Word Priming in Attractor Networks *

Suzanna Becker, Marlene Behrmann and Morris Moscovitch
The Rotman Research Institute, Baycrest Centre
3560 Bathurst St.
Toronto, Ontario, M6A 2E1

Abstract

We propose a new view of word priming in attractor networks, which involves deepening the basins of attraction for primed words. In a network that maps from orthographic to phonological word representations via semantics, this view of priming leads to novel predictions about the interactions between orthographically and/or semantically similar primes and targets, when compared on an orthographic versus a semantic retrieval task. We confirm these predictions in computer simulations of long-term priming in a word recognition network.

Connectionist models have strongly influenced current thinking about the nature of human memory storage and retrieval processes. One reason for their appeal is that they can account for a wide range of human performance on tasks such as word recognition (McClelland and Rumelhart, 1981), reading (Seidenberg and McClelland, 1989), and repetition priming (McClelland and Rumelhart, 1986). Further, because connectionist models make relatively specific assumptions about the mechanisms of cognitive processes, they can lead to novel predictions about human performance.

One of the most exciting developments in the last decade of human memory research is the characterization of implicit memory (Graf & Schacter, 1985; Schacter, 1985), a form of automatic, unconscious retrieval of previously encountered material. A widely used experimental method for testing implicit memory is repetition priming, in which the accuracy or speed of processing is measured on successive presentations of a target stimulus. Evidence of implicit memory is observed when subjects are more accurate or efficient in responding to previously studied targets than to new targets. The priming literature is highly relevant to connectionist models of learning and memory for two reasons: 1) Priming effects can be extremely long-lasting, ranging from minutes to many hours, or even months, and apparently reflect fundamental automatic ("unsupervised") learning processes employed by the brain. 2) When the prime and target are not identical, but have similar input and/or semantic features, the priming effects may range from facilitation to inhibition; these effects can shed light on the nature of human memory organization, and provide constraints on the representations employed in connectionist models.

In this paper, we first review the previous connectionist accounts of priming. We then propose a new view of word priming in attractor networks with orthographic and semantic levels of representation, which involves deepening the basins of attraction for primed words. This leads to some novel predictions about the interactions between primes and targets, which we explore in computer simulations.

Short- and long-term priming

Many forms of priming have been studied (for reviews, see Monsell (1985), Schacter (1987) and Richardson-Klavehn and Bjork (1988)), and several connectionist accounts of these phenomena have already been proposed. Short-term priming of a word by a semantically related word or context can be accounted for by Masson's model (1991) in which residual activation from the previous input influences the network's response to subsequent input. Bavelier and Jordan's model (1993) employs transient changes in baseline activity and threshold levels to account for a range of facilitatory and inhibitory repetition priming effects (repetition blindness, masked and unmasked priming). These models provide plausible accounts of many of the observed short-term priming effects.[1] However, longer-term priming (which persists across

*Financial support for this research is provided by a grant from the McDonnell-Pew Program in Cognitive Neuroscience (Grant number 92-40) to Becker and Moscovitch, and a Medical Research Council scholarship and operating grant to Behrmann.

[1]Bavelier and Jordan proposed that long-term repetition priming of a word is due to a long-term threshold reduction for the appropriate word-detector unit. However, this does not account for long-term form-priming.

the presentation of many intervening stimuli) clearly involves long-term changes in processing. It is this class of phenomena with which the current paper is concerned.

Long-term repetition priming

McClelland and Rumelhart (1986) proposed a mechanism for long-term repetition priming. In their distributed memory model, a recurrent network of units learns via an error-correcting procedure to store a set of training patterns and to perform pattern completion. Each learning step involves a large initial change in each weight (in proportion to the error for that weight), which rapidly decays down to a permanent or slowly decaying smaller change. Thus, each weight must store a history of the changes it has undergone as a function of each pattern it has recently seen. The model accounts for priming by the temporarily exaggerated weight changes on connections to units participating in recently-experienced patterns, making those patterns more likely to be completed in the near future. A similar mechanism for priming was proposed by Plaut (1991) to model perseveration effects in optic aphasia, using a combination of fast weights and slow weights (Hinton and Plaut, 1987). Short term correlations between units are stored by one-shot learning in the rapidly decaying fast weights, while the long term knowledge of the network is stored in the slow weights. [2]

Similarity-based Priming

In contrast to repetition priming, in which the prime and target stimuli are identical, still more interesting priming effects occur when the prime and target stimuli differ, but are related along some dimension. For example, in form-based priming (Forster, 1987), the prime and target are similar in the input space (sharing orthographic or phonological features) but are otherwise unrelated. The models described in the previous section provide fairly simple and plausible accounts of repetition priming, in which the same input pattern leads to more rapid or accurate responses upon repeated presentations. However, they make no specific predictions when the prime and target stimuli may be related at different levels of representation, as in form-based priming versus semantic priming.[3] In this case, stronger assumptions must be made about the nature of the representation and/or processing of stimuli.

Most of the experimental work on form-based priming has examined short-term effects. Such effects were first reported by Meyer et al., (1974) who found faster lexical decision for words primed by orthographically and phonologically similar words. However, subsequent attempts to replicate these findings have produced an array of apparently contradictory results. Forster and Davis (1991) report that facilitatory form-priming occurs only when the prime is heavily masked so that subjects cannot identify it. Forster et al. (1987) found orthographic priming in visual recognition of 8-letter but not 4-letter words, although for 4-letter words with unusual spelling patterns the effect was significant. The latter effect was accounted for by assuming an interactive activation model of word recognition. Residual activation from the prime presumably produces cross-activation of the target representation if it shares letters, but the prime also exerts inhibition on competing word responses. The net effect is assumed to be more strongly inhibitory for high frequency words, as they have more competitors. Slowiaczek et al. (1992) only found facilitatory auditory priming when the prime and target shared one phonological feature, and found *interference* if the overlap increased from one to two features; like Forster et al., they interpreted these findings in terms of an interactive activation model of word recognition. In the latter case, lexical interference between mutually inhibitory word representations was assumed to occur.

In contrast to the above experiments which studied short-term masked priming effects, long-term form-based priming has been much less studied. Rueckl (1990) has demonstrated long-term effects in tachistoscopic recognition of visually presented words, when each word was primed by many orthographically similar words. These results were interpreted using McClelland and Rumelhart's incremental learning account of long-term repetition priming, combined with the assumption that the amount of form-priming should increase with the degree of orthographic similarity between prime and target. Rueckl also reported better orthographic priming for words than non-words; using a connectionist interpretation,

[2] Plaut's model actually employed two different learning mechanisms for the two sets of weights, a Hebb-like correlational rule for the fast weights and a supervised error-correcting rule for the slow weights.

[3] But note that McClelland and Rumelahart did simulate a form of similarity-based priming, by testing with noisy versions of the stored patterns.

this effect was attributed to enhanced activity in the orthographic level of representation due to top-down support from the semantic level.

Priming in attractor networks

To the extent that connectionist interpretations have been invoked to account for priming effects, for the most part they have been unimplemented, and based on very simple network models with relatively constrained patterns of dynamic activity, such as McClelland and Rumelhart's (1981) interactive activation model of word recognition. These models provide reasonable accounts of basic repetition priming, and facilitatory form-based priming (which presumably occurs primarily at a pre-semantic level). However, to account for any potential interactions between semantic and lower level (perceptual or lexical) processing, a more sophisticated word representation is required which includes a semantic level.

One such model of word recognition was proposed by Hinton and Shallice (1991) to account for symptoms of reading impairment (dyslexia) under damage. The network was trained with the recurrent back-propagation learning procedure to map orthographic representations of words onto distributed semantic representations via a layer of hidden units, with the help of semantic "cleanup units". There were feed-forward connections from orthographic to hidden units, and reciprocal feedback connections between the hidden and semantic units, the cleanup and semantic units, as well as within the semantic layer. Using this architecture, Hinton and Shallice proposed a rather novel view of the mechanism underlying word recognition: when a word is presented to the network as an orthographic pattern, the network gradually settles into a stable semantic representation (an attractor state) once it has recognized a word. The current state vector of the input units can be thought of as a single point in a high-dimensional orthographic space, having dimensionality equal to the number of input units. Likewise, the current state vector of the semantic units can be thought of as a single point in semantic space. Because there is no systematic correspondence between orthographic and semantic features, the network must learn to map points which are nearby in the input state space onto much more distant points in the semantic state space. In contrast, semantically related words are typically represented by distant points in orthographic space, but nearby points in semantic space because they share many semantic features. When the network is damaged (by removing connections or adding noise to the weights), the reading pattern associated with deep dyslexia is observed: not only are visually similar words occasionally confused, but substitution errors involving semantically related words are also observed.

Adopting the Hinton-Shallice view of word recognition leads us to conceptualize the process of priming rather differently. When a word is presented to the network and its orthographic representation is activated, the network settles into a semantic attractor. Assuming Rumelhart and McClelland's view of long-term priming is correct, all the connections participating in the entire pathway between active pairs of neurons should be strengthened. This should increase the probability that the network will produce the same response when given the same input pattern in the future (even if the pattern is noisy or incomplete), by *deepening the attractor* for this pattern. Because the shape of the attractor basin has now been altered, the network's responses to other patterns should also be affected. Small perturbations in the input or semantic state of the network in the vicinity of this attractor should be pulled more strongly back to the attractor state. Attractors for semantically similar words should also be deepened, if they overlap with the prime along many dimensions. Likewise, at a pre-semantic level, a word primed by an orthographically related word should be recognized more quickly since its basin of attraction in orthographic space should also be deepened. On the other hand, priming with an orthographically similar word (with very different semantics) might *impede* the network's ability to retrieve the correct semantic representation for the target word; the mapping between orthography to semantics having been altered by priming, the network's state trajectory into the semantic attractor for the target word should be distorted so that there is now a stronger competing pull by the semantic attractor for the prime. We would therefore predict more errors in retrieving the correct semantics in the latter condition, and that correct retrieval should take longer for the target.

The present simulations were conducted to test two hypotheses. First, form-priming should produce both facilitation in (pre-semantic) word recognition and inhibitory priming on a task involving semantic retrieval. Although the latter effect has not been examined in human experiments, it is a novel prediction generated by our view of priming, and if it were to hold in computer simulations, would merit further investigation in the laboratory. Second, although long-

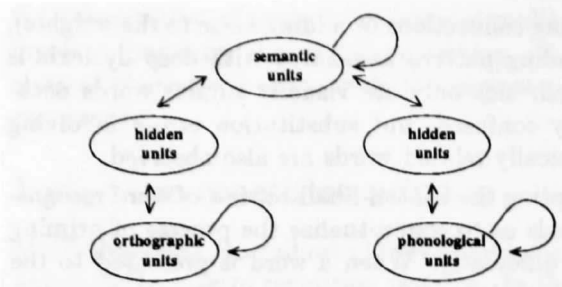

Figure 1: *The architecture of the network. Arrows indicate full connectivity amongst units within or between groups. There were 40 units in each hidden layer, 28 orthographic units, 33 phonological units, and 68 semantic units.*

term semantic priming has not been found on purported "pre-semantic tasks" such as lexical decision, our model predicts that when the prime and target words have highly overlapping semantics, on a *semantic retrieval task*, positive priming should be observed.

Priming experiments

The network

Our experiments were conducted using a slightly more elaborate version of Hinton and Shallice's network, pre-trained on the same set of forty words. Plaut (1991) extended the Hinton-Shallice model by adding a phonological output layer to the network. Plaut also showed that a network having a simpler architecture (omitting the cleanup units but adding more feedback connections) shown in Figure 1 could produce qualitatively similar behaviour when trained with the deterministic Boltzmann machine (DBM) learning procedure (Peterson and Anderson, 1987) rather than with back-propagation. For our experiments, we used essentially the same network as Plaut (Figure 1), except that each unit in the orthographic and phonological groups had a self-link (in addition to within-group links to each other unit). In order to be able to study the network's pre-semantic response time, we also used a slightly different mode of pattern presentation. Instead of presenting a pattern to the network by "hard-clamping" the states of the orthographic units and allowing the rest of the network to settle, we used a "soft-clamping procedure" in which the input value for each orthographic unit was treated as an activation on an extra incoming connection, multiplied by a fixed weight of 2.0. The input units were thereby permitted to settle to stable states along with the rest of the network.

The training set consisted of the same forty three- and four-letter words used by Hinton and Shallice (1991). There were 28 orthographic input units, one for each possible letter-position combination. Similarly, Plaut's phonological output representation consisted of 33 word-position-specific phonemic features. Words were grouped into five categories: indoor objects, animals, body parts, foods and outdoor objects. The Hinton-Shallice semantic features were chosen so that words in the same category had considerably greater overlap of semantic features than words in different categories. The network was first trained for 2000 iterations exactly as described by Plaut (1991) using the DBM learning procedure, with hard-clamping. The network was then trained an additional 1500 iterations using soft-clamping in the negative phase, with noise added to the inputs (with mean zero and standard deviation 0.05), and a smaller learning rate of 0.005 / fanin (the average number of incoming connections to each unit) until it learned to produce the correct semantic and phonological representations in response to each orthographic input pattern. Units used the hyperbolic tangent non-linearity, and had real-valued continuous activations ranging from -1 to 1. No priming effects were simulated during the training phase.

Simulations

Priming was simulated as proposed by Plaut (1991) using an extra set of correlational "fast weights" (Hinton and Plaut, 1987). Each fast weight was stored as an extra value on the same connection as the corresponding slow weight, and their values were added together to produce a single "net weight" on the connection for the purpose of state updates. Upon presentation of each prime, the network was permitted to settle to a stable state, after which each fast weight was set to the product of the pre- and post-synaptic activities, multiplied by a constant of 0.0003. The network was then presented with a test pattern, and again permitted to settle under the combined influence of both the fast and slow weights. To avoid cross-talk between trials, the fast weights were set to zero after each trial. Thus, the learning rule for setting the amount of priming on each connection is exactly the same rule employed in the positive phase of DBM learning, except that it is applied to the fast weights, and its effect decays to zero between priming trials. The fast and slow weights can be thought of, respectively, as relatively fast-decaying and more permanent components of the same memory trace (although for simplicity, we fixed the slow weights dur-

ing priming trials).

The priming trials were divided into four conditions: 1) Identity, 2) Form&Semantic, 3) Semantic, and 4) Form-based priming. In condition 1) the prime and target were identical (repetition priming), and in the latter three they were orthographically similar (sharing two or more letters) and/or semantically similar (i.e., they were drawn from the same semantic category). Each of the forty words was tested when primed with every other orthographically and/or semantically similar word, and with itself. Additionally, in order to compute word baseline priming scores, each word was primed with eight unrelated words (two from every other category). Each prime-target combination was repeated ten times, with different noise vectors (uncorrelated noise with mean zero and standard deviation of 0.01) added to the input. On each priming trial, two measures were taken: 1) a pre-semantic word recognition task: how quickly the orthographic input units reached equilibrium, and 2) a semantic retrieval task: how quickly the semantic units reached equilibrium. A group of units was considered to have reached equilibrium when no unit changed its state by more than 0.005. Reaction times to respond to targets were measured as percentages of average word baseline scores, according to the equation $RT = (C_b - C_p) * 100/C_b$, where C is the number of cycles to reach equilibrium, p subscripts denote post-priming scores, and bs denote baseline scores. Error rates were also recorded.[4]

Results

The error rates were 3% for Identical primes, 31% for Form&Semantic primes, 12% for Semantic primes, and 32% for Form primes. The mean percentage changes in reaction times and corresponding standard errors on correct trials for all conditions are shown in Figure 2. The main effects of prime type and test type were both highly significant ($p < 0.0001$). Repetition primes (identical primes and targets) had by far the largest effect: there was consistently strong positive priming on both the pre-semantic and semantic tests. In the former case priming produced on average a 7% reduction in response time, and in the latter case a 16% reduction.

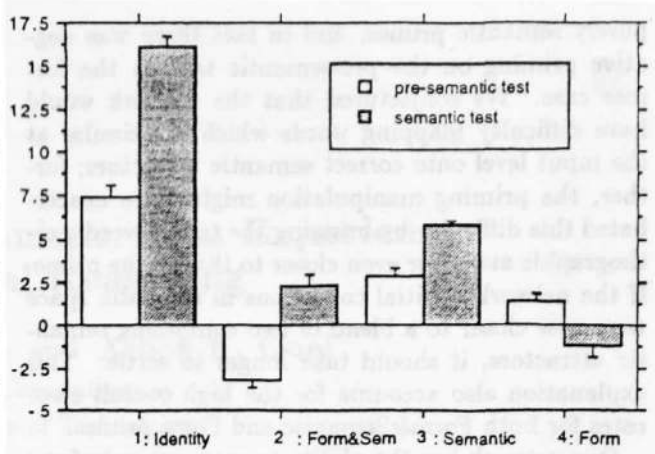

Figure 2: *Mean RT (with standard error bars) versus prime type. RT is averaged across trials, and is measured as a percentage change from baseline. A positive mean RT reflects a decrease in settling time.*

The interaction between prime and test type was also highly significant ($p < 0.0001$). Post-hoc Newman-Keuls comparisons of the means revealed that for the first three prime types, there was significantly greater priming on the semantic than the pre-semantic test ($p < 0.001$). For Form primes, the reverse pattern was seen: mean priming on the semantic test was negative, with significantly greater mean (positive) priming on the pre-semantic test ($p < 0.001$). Significantly less priming was found for Form&Semantic primes compared to purely Semantic primes on both tasks, and in fact in the former case, priming on the pre-semantic task was actually negative.

Discussion and conclusions

As predicted, priming with an orthographically similar word was found to produce an advantage on the pre-semantic test relative to the semantic test on correct trials. The reverse effect was seen for the other three prime types: when the prime was identical to or semantically related to the target, positive priming effects were greater on the semantic test. The much larger overall error rates for Form primes and Form&Semantic primes indicate that form-priming impedes the network's ability to map to the correct semantic attractor.

Somewhat counter-intuitively, when the prime was related to the target word in both form and semantics, there was much less semantic priming than for

[4]Trials on which either the network settled to the wrong attractor, or on which the raw reaction time scores were beyond two standard deviations of the mean unprimed pre-semantic or semantic reaction times (averaged over words, with ten noisy presentations of each unprimed word), were treated as errors and were not included in the priming measures.

purely semantic primes, and in fact there was negative priming on the pre-semantic task in the former case. We conjectured that the network would have difficulty mapping words which are similar at the input level onto correct semantic attractors; further, the priming manipulation might have exacerbated this difficulty by bringing the target word's orthographic attractor even closer to that of the prime. If the network's initial conditions in semantic space were now closer to a blend of two competing semantic attractors, it should take longer to settle. This explanation also accounts for the high overall error rates for both Form&Semantic and Form primes.

Our network has the ability to map not only from orthography to semantics to phonology, but also to do the reverse mapping; in future work, this will allow us to investigate whether there is cross-modal transfer in the various priming effects we have described.

Our simulations make some novel predictions about human performance under the various priming conditions. We are currently investigating the question of whether human subjects would exhibit qualitatively similar interaction effects of test type and prime type.

Previous connectionist accounts of priming have relied almost solely upon the interactive activation model of word perception (McClelland and Rumelhart, 1981) to explain inhibitory priming effects (assuming mutual inhibition between competing word detectors). We have proposed a different view of priming, based on the Hinton-Shallice model of word recognition as a process of settling into an attractor in semantic space. We have added to this model the ability to simultaneously settle to a stable interpretation in orthographic space. This combined model appears to have predictive power, and warrants further investigation.

Acknowledgements

We are grateful to Jeff Toth and Dave Plaut for helpful discussions.

References

Bavelier, D. & Jordan, M. I. (1993). A dynamic model of priming and repetition blindness. In *Advances in Neural Information Processing Systems 5* (To appear). Morgan Kaufmann.

Forster, K. I. (1987). Form-priming with masked primes: The best-match hypothesis. In M. Coltheart (Ed.), *Attention and Performance XII* (pp. 127–146). Hillsdale, NJ: Erlbaum.

Forster, K. I. & Davis, C. (1991). The density constraint on form-priming in the naming task: Interference effects from a masked prime. *Journal of memory and language, 30*:1–25.

Forster, K. I., Davis, C., Schoknecht, C., & Carter, R. (1987). Masked priming with graphemically related forms: Repetition or partial activation? *Quarterly journal of Experimental Psychology, 39*:211–251.

Goldinger, S. D., Luce, P. A., Pisoni, D. B., & Marcario, J. K. (1992). Form-based priming in spoken word recognition: the roles of competition and bias. *Journal of experimental psychology: Learning, Memory and Cognition, 18*(6):1211–1238.

Graf, P. & Schacter, D. L. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*:501–518.

Hinton, G. E. & Plaut, D. C. (1987). Using fast weights to deblur old memories. In *Proceedings of the Ninth Annual Cognitive Science Society Conference* (pp. 177–186). Erlbaum, Hillsdale, NJ.

Hinton, G. E. & Shallice, T. (1991). Lesioning a connectionist network: Investigations of acquired dyslexia. *Psychological Review, 98*:74–95.

Masson, M. E. J. (1991). A distributed memory model of context effects in word identification. In Besner, D. and G. Humphreys (eds), *Basic processes in reading, Visual Word Recognition* (pp. 233–263). Lawrence Erlbaum Associates.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception, part i: An account of basic findings. *Psychological Review, 88*:375–407.

McClelland, J. L. & Rumelhart, D. E. (1986). *J. L. McClelland, D. E. Rumelhart (eds.),* a distributed model of human learning and memory. In *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. II. Cambridge, MA: Bradford Books.

Meyer, D. M., Schvaneveldt, R. W., & Ruddy, M. G. (1974). Functions of graphemic and phonemic codes in visual word-recognition. *Memory and Cognition, 2*:309–321.

Monsell, S. (1985). Repetition and the lexicon. In *Progress in the Psychology of Language* (pp. 147–195). London: Erlbaum.

Peterson, C. & Anderson, J. R. (1987). A mean field theory learning algorithm for neural networks. *Complex Systems, 1*:995–1019.

Plaut, D. C. (1991). *Connectionist neuropsychology: The breakdown and recovery of behaviour in lesioned attractor networks.* PhD thesis, Carnegie Mellon University.

Richardson-Klavehn, A. & Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology, 39*:475–543.

Rueckl, J. G. (1990). Similarity effects in word and pseudoword repetition priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(3):374–391.

Schacter, D. L. (1985). Implicit memory: history and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*:501–518.

Seidenberg, M. S. & McClelland, J. L. (1989). A distributed developmental model of word recognition and naming. *Psychological Review, 96*:523–568.

Slowiaczek, L. M. & Hamburger, M. (1992). Prelexical facilitation and lexical interference in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(6):1239–1250.