Automatic and Controlled Processes in Semantic Priming: an Attractor Neural Network Model with Latching Dynamics

Itamar Lerner (itamar.lerner@gmail.com)
Interdisciplinary Center for Neural Computation, The Hebrew University of Jerusalem
Giv'at Ram, Jerusalem 91904 Israel

Shlomo Bentin (shlomo.bentin@huji.ac.il)
Department of Psychology and Interdisciplinary Center for Neural Computation, The Hebrew University of Jerusalem
Mount Scopus, Jerusalem 91905 Israel

Oren Shriki (oren70@gmail.com)
Department of Physiology and Neurobiology, Faculty of Medicine, Ben-Gurion University of the Negev
Be'er-Sheva, 84105 Israel

Abstract

Semantic priming involves a combination of automatic processes like spreading activation (SA) and controlled processes like expectancy and semantic matching. An alternative account for automatic priming has been suggested using attractor neural networks. Such networks offer a more biologically plausible model of real neuronal dynamics but fall short in explaining several important effects such as mediated and asymmetrical priming, as well as controlled effects. We describe a new attractor network which incorporates synaptic adaptation mechanisms and perform latching dynamics. We show that this model can implement spreading activation in a statistical manner and therefore exhibit all priming effects previously attributed to automatic priming. In addition, we show how controlled processes are implemented in the same network, explaining many other semantic priming results.

Keywords: Semantic priming; Attractor networks; Latching dynamics

Introduction

Semantic priming is one of the most important phenomena in the study of word perception and semantic memory. In a typical priming experiment (Neely, 1991), subjects are visually exposed to two words in succession, the prime and the target, and are required to silently read the prime and either name the target (pronunciation task), or decide whether it is a real word or not (lexical decision task). The target could either be semantically related or unrelated to the prime (or a nonword, in case of the lexical decision task). The priming effect is expressed as shorter average reaction times (RT) and reduced error rates in the related relative to unrelated condition. Sometimes, a neutral prime is used (e.g. a row of X's) to allow the differentiation between response facilitation (in the related condition) and inhibition (in the unrelated condition).

Computational accounts for semantic priming are divided between models based on automatic processes and those based on controlled processes. The most famous among the automatic accounts for priming is the spreading activation (SA) theory of Collins & Loftus (1975). This model suggests that concepts in semantic memory are represented by nodes that are connected to each other according to their semantic relatedness. When a concept is activated (by external or internal input) the activity spreads to related concepts (see figure 1). In priming experiments, activation of the prime concept (e.g. *table*) leads to activation of its related concepts (e.g. *chair*). This pre-activation facilitates the recognition of subsequent related targets. If an unrelated or a neutral target appears, no such head-start is available. Hence spreading activation may account for the facilitation but not the inhibitory component of semantic priming. Automatic priming can also be conceived in attractor networks with distributed representations of concepts (e.g. Mason, 1995). In such models concepts are represented by activity patterns of neurons' assemblies and semantic relationship is implemented as correlation between these representations. When the prime appears, the network converges on its corresponding activity pattern. When the target is then presented, the network changes its activity pattern from that of the prime to the one corresponding to the target. If the target is related to the prime, fewer changes need to take place due to the correlations; therefore, the convergence takes less time and a priming effect emerges.

While attractor networks are probably more true to the biological nature of real neuronal dynamics, they also fall short in explaining several important priming results. Mediated priming is one example (e.g. McNamara, 1992): It was found that word pairs which are indirectly related to each other (i.e., related only through a mediating word, like *lion* and *stripes*, related through *tiger)* can nevertheless prime each other. Allowing activation to spread to more than one step, SA theories could easily account for such effects. Attractor networks, on the other hand, cannot explain mediated priming since the activation patterns of indirectly related pairs are not correlated. Similarly, whereas SA models allow asymmetric connections between nodes and therefore allow asymmetric priming (in which the magnitude of priming varies according to which word in a given pair is designated prime and which is the target; e.g. *paycheck* vs. *check-pay*), such an effect cannot be obtained by

attractor network models because they rely on correlation, a symmetric trait by definition.

Here we present an attractor neural network which implements SA in a statistical manner. By doing so, we bridge between SA and attractor models and show how attractor networks can exhibit results like mediated and asymmetric priming. In addition, we discuss some controlled mechanisms like expectancy (Becker, 1980) and semantic matching (Neely, Keefe & Ross, 1989) and suggest how they may be interpreted within the same network.

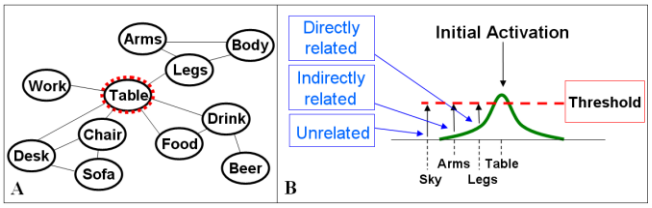

Figure 1: The spreading activation theory. (A) Related concepts connected in semantic memory. (B) Activation spreads through the network

Computational Model

Following the traditional separation between stages of processing (e.g. Smith et al., 2001), our model consists of 2 computational layers, lexical and semantic (Figure 2). We assume that after a string of letters is analyzed for orthographic composition, the result is fed to the lexical network where word identification occurs. If the letters form a real word, this word is 'recognized' by the lexical network and its activity is fed forward to the semantic network where the word's meaning is stored. However, the semantic network can influence lexical processing on line via feedback. Such a top-down effect contributes to semantic priming: when the semantic network is a priori 'tuned' to a concept with some relatedness to the newly arrived word, the lexical network recognizes this word quicker because both bottom-up and top-down pathways contribute to the recognition process (as opposed to the unrelated case, where the top-down pathway does not contribute). In the case of a neutral stimulus, none of the networks is activated and no transfer of information occurs.

The lexical and semantic networks are modeled as Hopfield-type attractor neural networks, with sparse representations and continuous-time dynamics (see Tsodyks, 1990). In our simulations, both the lexical and the semantic networks are fully connected recurrent networks, each composed of 500 neurons. Memory patterns (concepts) encoded by each network are binary vectors of size 500, with '1' indicating a maximally active neuron, and '0' an inactive one. The representations are sparse (i.e., a small number of neurons are active in each pattern) with p being the ratio of active neurons ($p<<1$). The connectivity between neurons assures stability of these patterns. External inputs to and from the network are always excitatory.

The neurons themselves are analog with activity in the range [0,1] and obey a logistic transfer function of their local input $h(t)$. The local input itself obeys a linear differential equation (following Herrmann, Ruppin & Usher, 1993) of the form:

$$\textbf{(1)}\ \tau_n \dot{h}_i(t) = -h_i(t) + \sum_{j=1}^{N} J_{ij} x_j(t) - \lambda \cdot (\bar{x}(t) - p) - \theta + \left[I_i^{ext}(t) - \theta^{ext} \right]_+ + \eta_i$$

In (1), τ_n is the time constant of the neuron, $x_j(t)$ is the activity at time t of the j-th neuron (with $\bar{x}$ indicating average over all neurons), J_{ij} are the connectivity weights, N is the number of neurons (500 in our case), p is the sparseness of the representations, λ a regulation parameter which maintains stability of mean activation, and θ is a constant neuronal threshold (See Herrmann et al. for details). The $[\ldots]_+$ symbol indicates a threshold linear function, such that $[x]_+=0$ for $x<0$, and $[x]_+=x$ otherwise. This leads the external input to the neuron, $I_i^{ext}(t)$, to be consequential only if it surpasses the constant external threshold θ^{ext}. Finally, η_i is a noise term drawn from a Gaussian distribution with some temporal correlations. Relatedness between concepts is implemented in the model as correlations between memory patterns (reflecting the degree of overlap between them). The stronger two concepts are related, the higher is their correlation. Unrelated patterns have a correlation near 0.

Two major differences distinguish the lexical from the semantic network. The first is that there are no correlations in the memory patterns of the lexical network. This is not to indicate there are no lexical relations (indeed, such relations obviously exist, at least at the phonological level), but merely to ensure that such relations would not influence the simulations. In fact, typical semantic priming experiments control for such confounds by selecting prime-target pairs that bare no lexical/phonological relations. The semantic network, however, includes correlated memory patterns representing semantic relations between concepts.

The second difference is, perhaps, the basic premise of our model: Unlike the lexical network (and the majority of previous attractor network models), our semantic network is associative in nature. Neuronal adaptation mechanisms at the synaptic level preclude the network from maintaining stability for long; therefore, the network, after converging to one attractor, leaves it quickly and jumps to another one. This process is stochastic in nature and can continue forever as long as no new input interferes. These jumps cannot be accurately predicted, but they tend to happen (although not necessarily) between correlated patterns. Such network behavior was termed 'latching dynamics' by Treves (2005). Specifically, short-term synaptic plasticity was modeled according to Loebel & Tsodyks (2002), with each synaptic weight of a neuron decreasing linearly with its activity:

$$\textbf{(2)}\ \dot{J}_{ij}(t) = \frac{J_{ij}^{\max} - J_{ij}(t)}{\tau_r} - U x_{\max} x_i(t) J_{ij}(t)$$

In (2), J_{ij}^{max} is the common Hopfield connectivity weight for sparse networks, τ_r is the time constant of recovery of the synaptic efficacy, and U is the rate of synaptic depletion.

The term x_{max} is a hypothetical maximum firing rate of a neuron (for example 100 pulses/sec) which adjusts the equation to fit a neural firing rate bounded by 1.

Links between the lexical and semantic networks are based on connections between active neurons in corresponding patterns (See figure 2): An active neuron in a certain word pattern in the lexical network sends excitatory connections to all active neurons in the corresponding concept-pattern of the semantic network, and vice-versa. Since correlations between patterns exist in the semantic network, one neuron in that layer could concurrently influence and receive input from different neurons activated in different patterns in the lexical layer. The lexical network also receives bottom-up input, representing the visual letter-string, which follows the same logic: Neurons belonging to the pattern presented to the lexical network receive excitatory inputs, while others receive no input.

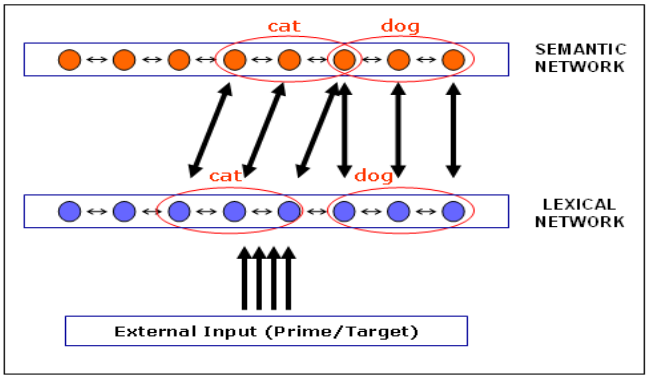

Figure 2: Architecture of the model. Two recurrent networks connected to each other with excitatory links. The semantic network contains correlated representations

Simulations

The simulations were run on an Intel Core 2 Quad CPU Q6600 with 2.4 Ghz and 2 GB of RAM. Simulations were written in Matlab 8a. In all the simulations, one numeric step represents 0.66ms.

Encoded Patterns

We encoded 17 memory patterns in each network. All patterns were binary vectors with equal mean activity and very sparse representations. In the semantic network, the following basic correlations between patterns were set: four groups, each containing four patterns, formed 'semantic neighborhoods' (patterns 1-4, 5-6, 9-12 and 13-16): Each pattern in a group was correlated with the other patterns in its group, but, with few exceptions (see below), no correlations existed between the groups. Correlations within a group had one of two values, representing two levels of direct relatedness. In addition, we also added some correlations between patterns of different neighborhoods to allow indirect priming investigations. For example, we added some correlation between pattern 2 and pattern 9, which resulted in patterns 1 and 9 being indirectly related. The 17th memory pattern was a 'baseline' pattern which the network was initialized to at the beginning of each trial, and was not correlated to any of the other patterns. In the lexical network, all 17 patterns were unrelated to each other. The 17th pattern was, again, the initial state for the network, and was not linked through top-down or bottom up lexical-semantic connections to the baseline pattern in the semantic network (thus forming a 'neutral' pattern).

Experimental Procedure and Data Analysis

Each trial began with the presentation of a binary vector to the lexical network, corresponding to one of its patterns (1's in the to-be activated neurons, 0's in the rest). This vector served as "prime". In neutral trials, pattern 17 (the neutral pattern) was presented. Two experiments were conducted. The first tested the general performance of the semantic network. The prime was presented for 100ms and it was always pattern no. 1. The network was allowed to advance according to the dynamic equations without further interference, for a total period of 3000ms. The procedure was repeated for 100 trials. Correlation of the momentary network state with each pattern, for each time point in the simulation, was averaged offline. The second experiment tested whether the performance of the model, when semantic priming occurs, corresponds with predictions based on human studies. The prime was presented for 100ms and followed by a target after 150 ms, hence creating a 250 ms SOA. The time interval from target onset until convergence of the lexical network indexed the reaction time, providing the network converged to the correct attractor. Primes and targets were either directly related (i.e., two patterns from the same neighborhood), indirectly related (two patterns from different neighborhoods but linked through a mediating pattern as explained earlier), unrelated (two patterns from different neighborhoods with no indirect connections), or neutral (in which the prime was the neutral pattern and the target any of the 'real' patterns). 100 trials were simulated for each relatedness condition, with prime-target pairs chosen randomly. Mean reaction times and standard errors were computed for each condition.

Results

Figure 3 presents the typical performance of the two networks (for presentation purposes, here we used a 1000ms SOA). Correlation of the state of each network with each of its stored patterns (including the memories and the neutral pattern) during a trial is presented in different colors, with convergence to a specific pattern indicated by its number appearing on top. The lexical network followed the external input, by converging to the corresponding memory pattern and keeping stability until a new input arrived. In contrast, the semantic network converged to the appropriate memory pattern, only to jump to other attractors in a serial manner, hence presenting latching dynamics. When a new external input arrived, the semantic network stopped its transitions and quickly converged to the corresponding memory pattern

shortly after the lexical network has done so. As evident in Figure 3, most jumps were within the neighborhood, while jumps to different neighborhoods occurred less frequently.

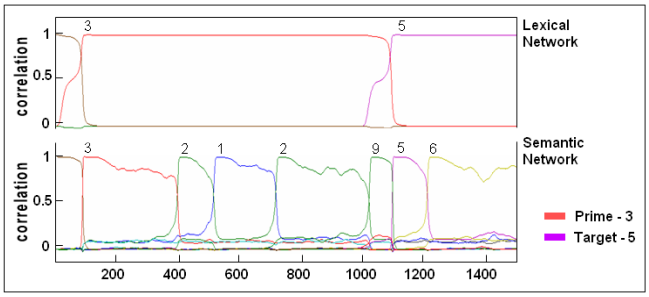

Figure 3: Typical behavior of the two networks

In the first experiment, trials always included pattern 1 as the prime. The mean correlation between the state of the semantic network and all its memory patterns was computed for each time point over trials. Figure 4A presents the correlations for five different time points after the prime onset. The x-axis represents different patterns according to their relatedness to pattern 1, with pattern 1 itself in the middle. Evidently, the mean correlations followed the principle of spreading activation: Initially, the concept represented by the external input has the strongest activation (correlation), its directly related concepts are activated to a smaller degree, and concepts not related to it are not activated at all. With time, as semantic transitions occur, the mean activation of the initial concept is decreasing, while activation in its related concepts increases. Indirectly related concepts also show some activation, with a delayed peak. Unrelated concepts receive no activation at all. After enough time, the mean correlation with each of the network's patterns is divided more or less equally, corresponding to a nearly deactivated state of the whole network (the mean activity would have reached near zero values in case more than 16 patterns were used).

In the second experiment, the mean RTs of the lexical network were computed and are presented in figure 4B. As can be seen, priming occurs for both directly and indirectly related pairs, although the effect is stronger in the direct case. In addition, weak relations produced smaller priming than strong relations. All these effects were significant at $p < 0.001$. There was no significant difference between the unrelated and neutral conditions, confirming that only facilitation occurred.

Discussion

The results of these simulations demonstrated that an attractor neural network with latching dynamics can implement spreading activation in a statistical manner. In a way, one could see the activity of nodes in the original spreading activation model as an average manifestation of the correlation in our attractor model. There is, however, an important distinction between our model and SA models: In our network, spreading is mixed with relaxation periods which correspond to the network reaching an attractor. In other words, activation does not spread in a monotonic manner like in the original SA model, but rather in jumps which fit the dynamical jumps from one attractor to another.

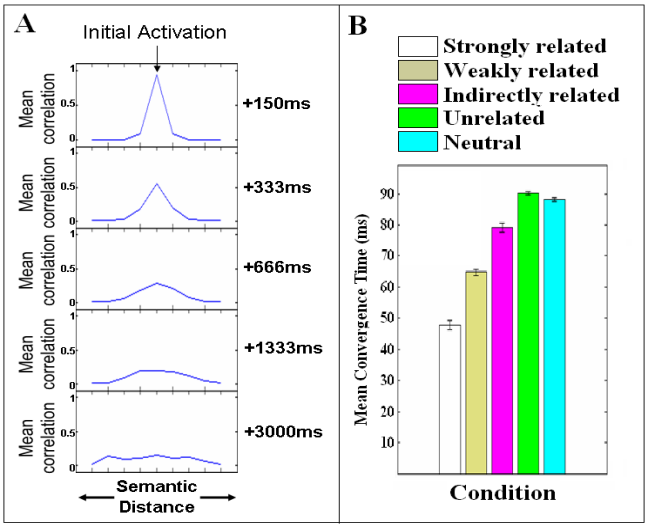

Figure 4: Simulations results. (A) Statistical spreading activation portrayed by the network as mean correlation over trials. (B) Mean convergence times of the lexical network for the different relatedness conditions

The results of the second simulation demonstrate how the dynamics in the semantic network affects the convergence time of the lexical network such that priming effects are produced. When the semantic network state is correlated with the target pattern at the moment the target word appears, its top-down influence shorten the lexical network's convergence times. Due to semantic transitions, such correlations may occasionally appear in indirectly related trials and produce the mediated priming effect. Although not explicitly simulated, these jumps can also produce asymmetry in priming: Transition probabilities from pattern A to pattern B are not necessarily equal to transitions from B to A. This asymmetry also allows making a distinction between semantic relatedness (as indicated by correlation) and associative relatedness (as indicated by the probability of one pattern leading to another pattern). Former attractor models relied solely on correlations between prime and target and therefore could not produce either mediated priming or asymmetry in priming.

Controlled Processes

When the SOA between prime and target is sufficiently long, subjects may decide to engage in specific strategies while responding. The general aim of such strategies is to shorten reaction times to the target. In contrast to the automatic nature of SA, strategies are considered to be under the subject's cognitive control.

A well known example of such strategies is expectancy (Becker, 1980). It is assumed that subjects may be able to realize that in part of the trials, the target is semantically

and/or associatively related to the prime and develop a set of expected targets from the prime's semantic "neighborhood". When the target appears, this "expected set" is searched first, while the general lexicon is searched only if the target is not included in the expected set. Obviously, when the target is found in the expected set its recognition time is accelerated. If it is not found there, however, its recognition is delayed by this initial screening procedure. Hence, the application of an expectancy strategy may account for both facilitation and inhibition of the priming effect. Two types of this strategy were identified (Becker, 1980): A 'prediction' strategy is used when the upcoming target is highly predictable. Only one item (or very few) is included in the expected set and, consequently, facilitation is robust while inhibition is negligible. A 'general expectation' strategy is used when more than a few items could potentially be targets and the expected set includes them all. Both facilitation and inhibition should result. However, subsequent studies have shown that not all conditions yield inhibition (for example, pronunciation tasks), which put this later strategy into question (Keefe & Neely, 1990). Either case, the requirements for this controlled process to be initiated are sufficiently long SOA, and a sufficiently salient proportion of related pairs in the stimulus set (called the 'relatedness proportion'), which makes such expectancies reasonable. Indeed, it was found that the relatedness proportion is positively correlated with priming, but only at long SOAs (Neely, 1991).

Controlled Processes in the Model

So far, we implicitly assumed that semantic transitions in the network happen automatically. We now turn to a different hypothesis: Semantic transitions may be controlled to some degree; therefore, while SA is the default behavior of the network when no interventions occur, other patterns emerge as soon as subjects attempt to control these transitions.

Controlling transitions can allow our model to implement expectancy (or at least the 'prediction strategy') if we consider the transition of the semantic network's state from a given prime pattern to another pattern as an 'expected' word for that prime. By default, such expected word is determined according to semantic relatedness principles. However, this conceptualization of expectancy makes it no different than our implementation of SA. What, then, makes expectancy a distinct mechanism? The answer is that expectancy can be modeled as the controlled operation of manipulating transition probabilities according to any information acquired by the subject up to that point, as to induce certain transitions and avoid others. For instance, expectancy can be realized by maintaining just one single transition in the semantic network (as opposed to many transitions in the default case). Another realization can be by controlling the variability of the semantic network's transitions, such that transitions will almost always occur from the prime to its most correlated pattern (as opposed to the more stochastic nature of transitions in the default state). The first suggestion can be implemented by allowing the network to make a single jump, as usual, but then stop any further transitions by lowering the background noise. This means, of course, that noise amplitude must be susceptible to cognitive control. We suggest that this is the equivalence of 'focusing attention' on the prediction. The second suggestion can be implemented by lowering the amplitude of the temporal correlations of the noise, which may be seen as focusing attention on the most probable prediction. Each of these two manipulations, as well as their combination, may have beneficial results: In case they succeed (i.e., the target indeed turns out to be the equivalent of the pattern the network has jumped to), an increase in priming is to be expected compared to the default case since all of the activated neurons of the semantic network would participate in accelerating the response. Without such intervention, the network is much less likely to be converged on the 'right' pattern when the target arrives, which implies that on average, only a minor set of the activated neurons will participate in the acceleration of response. Naturally, if the prediction is wrong, the response might be delayed compared to the default case. Hence this mechanism should be used only when there are good reasons to assume the target is predictable, that is, when the relatedness proportion is high. Moreover, the effect of these manipulations is expected to be most conspicuous on long SOAs, since on short SOAs there is usually not enough time for a transition to occur, let alone a series of transitions.

As an illustrative example, we have repeated simulation 2 for direct, indirect and neutral primes, for short/long SOAs, but this time we manipulated the amplitude of the noise. In one condition, the noise was reduced after the first transition in the semantic network (implementing the first mechanism we suggested for expectancy). In the other condition, no such manipulation was conducted. Figure 5 presents the results. As can be seen, the manipulation increased the facilitation effect, echoing the results in the literature (e.g. Neely, 1991). Naturally, this mechanism does not explain inhibition in priming, and following Keefe & Neely (1990), we speculate inhibition to be induced by different processes.

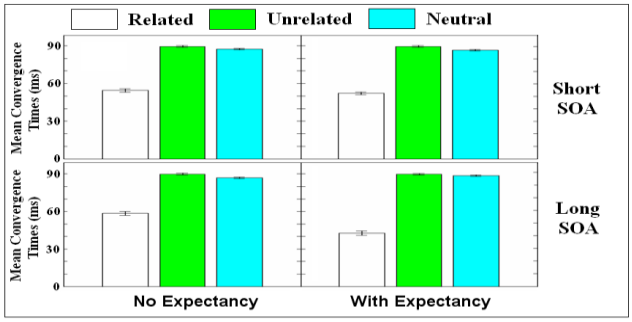

Figure 5: Average convergence times of the lexical network with and without an expectancy mechanism

Another controlled process presented in the literature is semantic matching (Neely et al., 1989). This process mainly involves decision making strategies which occur after lexi-

cal access to the target is achieved. In principle, it suggests that subjects engage in comparison between prime and target, which enables them to facilitate word and nonword responses in the lexical decision task.

While we did not attempt to fully model the semantic matching mechanism (which would necessitate incorporating a decision making mechanism), we would like to point out that any comparison between prime and target must depend on the prime being constantly activated in semantic memory throughout the whole trial, which in turn may indicate that no semantic transitions should occur in the semantic network. This, of course, can be achieved in our model by assuming a reduction in the noise amplitude immediately after lexical access of the prime (as opposed to the expectancy strategy case, where such a reduction is applied only after one semantic transition). We would therefore expect the usage of semantic matching to place severe limitations on spreading activation behavior, and specifically eliminate the indirect priming effect. Interestingly, this is exactly the result found in the literature (e.g. McNamara, 1992; see Neely, 1991, for a review).

General Discussion

Our main goal in the current study was to implement classical semantic processes related to semantic priming, with an emphasis on spreading activation, in a biologically-plausible framework of attractor neural networks. The results demonstrate that the basic characteristics of SA can be embedded in attractor dynamics while maintaining the same explanatory power of the original process. In addition, we show that controlled mechanisms involved in priming such as expectancy can be implemented within the same network, where the definition of 'controlled' is narrowed to the subject's influence on some specific parameters of the network.

Our network implies that real automaticity is the product of correlated representations. Direct semantic priming is a purely automatic process since, by definition, one pattern cannot be activated without partially activating its correlated patterns. On the other hand, processes which require a transformation from one representation to another can in principle be the object of cognitive control. Indirect priming can therefore be avoided by eliminating transitions in the semantic network. Spreading activation, by this view, is best seen as a default mechanism rather than a process which is completely automatic (cf. Smith et al., 2001).

Finally, a pure mathematical interpretation of the dynamics would suggest that the nature of the transitions between patterns in our model take the form of a Markov-chain, with the average correlation of the network with the various patterns forming a state vector and the transition probability matrix representing word association norms. Controlled strategies therefore represent a change in this matrix from the default values, which fit the subject's expectancies based on the accumulating data. Future inquiries may reveal the exact way by which these probabilities change as a function of the relatedness proportion, with Bayesian inference principles possibly governing this procedure.

Conclusion

Attractor neural networks have traditionally struggled with several important aspects of semantic priming compared to the more classical views. We have shown that an attractor network with latching dynamics can in fact implement some of these classical processes and serve as an equally competent model. Future work will need to specify in a more precise manner the exact ways by which strategies may influence our model's dynamics and how priming is affected by them.

References

Becker, C. A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory & Cognition*, *8*, 493–512.

Collins, A. M. & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

Herrmann, M., Ruppin, E. & Usher, M. (1993). A neural model of the dynamic activation of memory. *Biological Cybernetics*, *68*, 455–463.

Keefe, D. E., & Neely, J. H. (1990). Semantic priming in the pronunciation task: The role of prospective prime-generated expectancies. *Memory & Cognition*, *18*, 289–298.

Loebel A., & Tsodyks M. (2002). Computation by ensemble synchronization in recurrent networks with synaptic depression. *Journal of Computational Neuroscience*, *13*, 111–124.

Masson, M. E. J. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 3–23.

McNamara, T. P. (1992). Priming and constraints it places on theories of memory and retrieval. *Psychological Review*, *99*, 650–662.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. Humphreys (Eds.), *Basic processes in reading: Visual word recognition*. Hillsdale, NJ: Erlbaum.

Neely, J. H., Keefe, D. E. & Ross, K. L. (1989). Semantic priming in the lexical decision task: roles of prospective prime-generated expectancies and retrospective semantic matching. *Journal of Experimental psychology: Learning, Memory, and Cognition*, *15*, 1003–1019.

Smith, M. C., Bentin, S. & Spalek, T. M. (2001). Attention constraints of semantic activation during visual word recognition. *Journal of Experimental psychology: Learning, Memory, and Cognition*, *27*, 1289–1298.

Treves, A. (2005). Frontal latching networks: A possible neural basis for infinite recursion, *Cognitive Neuropsychology*, *22*, 276-291.

Tsodyks, M. V. (1990). Hierarchical associative memory in neural networks with low activity level. *Modern Physics Letters B*, *4*, 259-265.