COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 36 (2012) 1339–1382

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12007

# Spreading Activation in an Attractor Network With Latching Dynamics: Automatic Semantic Priming Revisited

Itamar Lerner,[a] Shlomo Bentin,[a,b,†] Oren Shriki[c]

[a]*Interdisciplinary Center for Neural Computation, The Hebrew University of Jerusalem*
[b]*Department of Psychology, The Hebrew University of Jerusalem*
[c]*Section on Critical Brain Dynamics, National Institute of Mental Health*

Received 12 December 2011; received in revised form 23 April 2012; accepted 2 May 2012

**Abstract**

Localist models of spreading activation (SA) and models assuming distributed representations offer very different takes on semantic priming, a widely investigated paradigm in word recognition and semantic memory research. In this study, we implemented SA in an attractor neural network model with distributed representations and created a unified framework for the two approaches. Our models assume a synaptic depression mechanism leading to autonomous transitions between encoded memory patterns (latching dynamics), which account for the major characteristics of automatic semantic priming in humans. Using computer simulations, we demonstrated how findings that challenged attractor-based networks in the past, such as mediated and asymmetric priming, are a natural consequence of our present model's dynamics. Puzzling results regarding backward priming were also given a straightforward explanation. In addition, the current model addresses some of the differences between semantic and associative relatedness and explains how these differences interact with stimulus onset asynchrony in priming experiments.

*Keywords:* Word recognition; Semantic priming; Neural networks; Distributed representations; Latching dynamics

## 1. Introduction

Related concepts tend to elicit one another in semantic memory. This simple and intuitive notion is firmly grounded in day-to-day experience, as well as in formal studies of human

Correspondence should be sent to Oren Shriki, Section on Critical Brain Dynamics, National Institutes of Mental Health, Bethesda, MD 20892. E-mail: oren70@gmail.com.
[†]Deceased July 13, 2012.

performance. For example, in free-association tasks, when subjects are instructed to respond with the first word that comes into their mind given a cue word, the response is often related in meaning to the cue (e.g., Deese, 1962); in sentence-verification tests, judgments regarding the semantic relationships between words are usually carried out faster for words sharing close semantic relations compared with words sharing distant relations (Collins & Quillian, 1969); in word-recognition studies implementing priming paradigms, subjects respond faster to a target word shortly after being exposed to a related word prime, compared to when the prime is semantically unrelated (Neely, 1991). Such findings have often been interpreted as supporting models of semantic memory in which the organization of knowledge is, at least partially, based on meaning-related neighborhoods.

One of the most prominent theories of semantic processing is the spreading activation (SA) model (e.g., Anderson, 1983; Collins & Loftus, 1975; Collins & Quillian, 1969). According to this model, concepts are represented by single units (or ''nodes'') and interconnected to each other in a network structure, which allows semantic activation to spread from one unit to another. The amount of activation that spreads is determined by the strength of the connection between two units, which represents their semantic/associative relatedness. The stronger two concepts are related to one another, the stronger is the connection between them (e.g., *table–chair* compared with *bed–chair* and *dog–chair*). When a concept is activated (e.g., by an external input), the activity level of its corresponding unit is set above a certain threshold, signaling its recognition by the system. The activity building at a particular node propagates to adjacent nodes automatically, thus elevating their activity. The concepts activated by proxy would, in turn, activate their own surroundings, and so activity spreads further and further in the semantic space (Fig. 1). If, during this SA process, the activity of a unit reaches its own threshold, its corresponding word becomes consciously perceived and available for evaluation. The SA stops when the original node is no longer externally activated, thus letting the remaining amount of activity in the network diminish with time and distance from origin (''dissipation of activation'').

A fundamentally different approach to modeling semantic memory is given by attractor neural network models (e.g., Masson, 1995; Moss, Hare, Day, & Tyler, 1994; Plaut, 1995; Plaut & Booth, 2000). The common assumption of such models is that concepts in semantic memory are represented by the distributed activity pattern (labeled ''memory pattern'') of an assembly of ''neurons.'' During a learning phase in which the concepts are introduced to the network, the connectivity among the neurons gradually changes until the memory patterns corresponding to the learned concepts become attractors in the network dynamics. Semantic relations are expressed in attractor models as correlations between memory patterns. In some models, this correlation is interpreted as feature overlap: If each neuron represents an explicit feature of a concept (e.g., *have four legs*), all concepts sharing this feature (*dogs*, *cats*, etc…) will have similar activity in the corresponding neurons. In other models, however, the correlation between patterns does not indicate distinguished shared features (see Jones, Kintsch, & Mewhort, 2006; Plaut, 1995). When the network is presented with an external cue corresponding to one of its stored memory patterns, the units' activity is gradually driven to this pattern until the network fully settles on its attractor state. This convergence represents the identification of the corresponding concept. As related concepts are

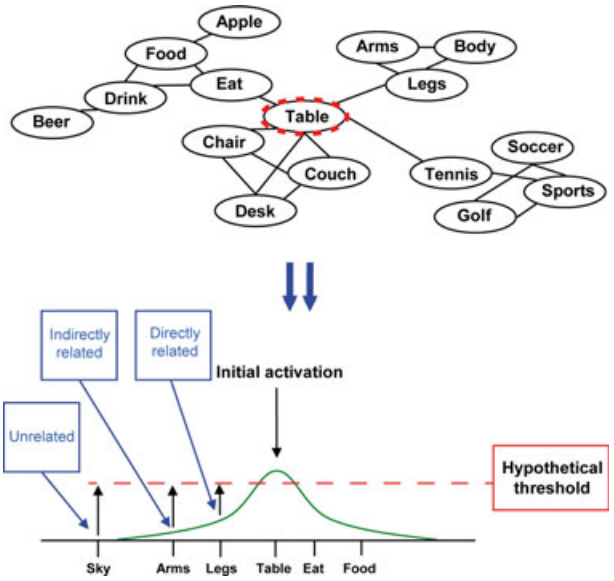


Fig. 1. Spreading activation operates on interconnected nodes within semantic memory. When one node is externally activated, activation spreads to related concepts, thus raising their baseline. Priming occurs when a pre-activated concept is presented as a target.

correlated, when one concept is ''activated'' (i.e., the network converges on it), its related concepts are partly activated in parallel. The dynamics in these networks is therefore short lived and terminates when the steady-state activity of an attractor is reached.

Spreading activation and attractor models have often been contrasted (e.g., McNamara, 2005; Thompson-Schill, Kurtz, & Gabrieli, 1998). SA models, being less constrained than attractor networks, are usually more flexible in accounting for various cognitive phenomena and, indeed, have been frequently used to simulate many well-known findings in the semantic memory literature (for a review, see McNamara & Holbrook, 2003). However, they are metaphorical models, which do not offer a mechanistic account of the dynamics in question. Hence, the traditional SA models may very well be seen as describing the dynamics in the semantic network (albeit in quantifiable terms) rather than pointing to its underlying (biologically inspired) sources. Attractor models, in contrast, may be somewhat more limited in their explanatory power due to their reliance on symmetric properties such as pattern similarity and their lack of emphasis on long-term dynamics beyond the typical convergence process. However, they provide mechanistic accounts of the processes involved and are more biologically oriented than SA, relying on principles such as distributed representations, attractor dynamics, and inter-neuronal connectivity driven by biologically inspired learning rules (e.g., Hebbian learning; Hopfield, 1982). While the biological plausibility of distributed representations was recently put into question (Bowers, 2009), they are, nevertheless, still a consensual view in neuroscience particularly when assuming sparse representations with low correlations between them (see, e.g., Plaut & McClelland, 2010; Quian Quiroga & Kreiman, 2010; Waydo, Kraskov, Quian Quiroga, Fried, & Koch, 2006). It is

therefore an interesting computational question whether attractor networks can be extended to reach explanatory power comparable to SA while retaining the principles that allow them a certain degree of biological feasibility.

In this study, we developed a model that attempts to take a step toward unifying the principles of SA and attractor networks into one coherent framework. We did this in the context of semantic priming, one of the frequently used paradigms in word recognition and semantic memory research that confronts many of the fundamental differences between the two approaches. We show how the introduction of biologically motivated adaptation mechanisms into attractor networks can lead to autonomous hopping between encoded memory patterns, which mimics SA dynamics and, consequently, how such a model can account for some of the basic findings in the priming literature.

An additional motivation for our work was that SA and attractor networks offer distinctively different takes on the general notion of semantic activation, which tap on several key issues in the semantic memory literature; these include the question of representation (local or distributed), degree of semantic activation (focused or spread), and the source of semantic and associative relatedness (static similarities between distributed representations or a product of dynamical changes in the system). Suggesting a model that combines the classical principles of SA and the more biologically plausible principles of attractor networks, we hope to present some of these issues in a new perspective.

This article has the following structure: We begin by discussing how SA and attractor models explain semantic priming and what the discrepancies between these models are. Then, we introduce our model, followed by simulations that demonstrate its basic traits and capability to explain previously reported semantic priming and free-association results in human performance studies. Finally, we discuss some implications of the model and how it may relate to several other theories in the field.

## 2. Semantic priming

### *2.1. Basic experimental findings*

Since its introduction in the early 1970s (Meyer & Schvaneveldt, 1971), semantic priming has been among the most widely investigated phenomena in the research of semantic memory. In a typical priming experiment (Neely, 1977; See Neely, 1991; McNamara, 2005, for reviews), the participant is presented with two words in succession, the prime and the target, with either a short or a long stimulus onset asynchrony (SOA). Frequently used procedures involve reading the prime silently and either naming the target (pronunciation task) or deciding whether it is a real word (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 semantic priming effect refers to the finding that the average reaction time (RT; pronouncing the second word or deciding it is a real word) is shorter, and error rates are lower when the two words are semantically related to each other, compared with when they are unrelated. Experiments have shown that priming may reflect both facilitation (i.e., a

prime accelerates the RT to a related target) and inhibition (the prime delays RT to an unrelated target) compared with a condition in which a neutral stimulus (such as a row of X's) takes the role of the prime.

The nature of the relations between primes and targets and how it influences the priming effect has often been the focus of attention in the priming literature. Words can be only semantically related (e.g., *trout–salmon*), could be episodically associated even without a semantic relationship (e.g., *pillar–society*), or could be related both semantically and associatively (e.g., *dog–cat*). Across all three types of relationships, priming is augmented for pairs which are closely related to each other (a ''strong'' connection) compared to pairs in which the relationship between the words is weaker (Lorch, 1982; Neely, 1991). Priming seems to be stronger for pairs that are both semantically and associatively related (as determined by free-association norms; see Nelson, McEvoy, & Schreiber, 2004) compared with pairs that are purely semantically related (the so-called associative boost effect; Hutchison, 2003; Lucas, 2000; Moss et al., 1994).When both semantic and associative relationships between prime and target exist, the magnitude of priming tends to increase with SOA (e.g., de Groot, 1984, 1985), whereas when the prime and target are only semantically related, priming is less affected by SOA (Lucas, 2000). Priming can also be asymmetric, so that the size of the priming effect changes pending on which of the two words in a pair is the prime and which is the target. This asymmetry is best demonstrated for pairs in which the words are associatively related in one direction (e.g., *stork–baby*), but not in the opposite direction (e.g., *baby–stork*). For such pairs, priming could be very effective when the prime and the target preserve the association, while if presented in the backward direction (termed ''backward priming''), the effect is smaller and may even disappear with sufficiently long SOAs (e.g., Kahan, Neely, & Forsythe, 1999). Finally, word pairs in which the prime is related to the target only indirectly through a mediating word (termed ''mediated priming''; e.g., *lion–stripes*, mediated by *tiger*) were shown to yield priming effects when strategic processes related to decision making are prevented (Balota & Lorch, 1986; Neely, 1991). These effects, however, are smaller compared to the priming of directly related items.

In general, models of semantic priming have focused on either automatic or controlled mechanisms contributing to the effect. Controlled mechanisms refer to specific strategies which subjects can intentionally use in an attempt to maximize the efficiency of their response to the target, producing either facilitation or inhibition at long prime-target SOAs. Automatic priming, on the other hand, results from the structure, dynamics, and connectivity of the semantic storage itself and is allegedly independent of subjects' strategies. These mechanisms typically contribute to the facilitation of target processing, primarily at short SOAs, without evident inhibitory effects. SA models and attractor networks account mostly for automatic priming.

### *2.2. The SA account*

Spreading activation theories explain semantic priming, assuming that when the semantic node which represents the prime is activated, the activation spreads automatically to related nodes (for review, see Neely, 1991). This wave of activation raises the baseline activity of

such nodes and, therefore, reduces the amount of additional activation needed to bring them above threshold when addressed bottom-up (Fig. 1). Hence, SA can facilitate the recognition of targets that follow semantically related primes, as reflected by faster RTs to such targets. The magnitude of this facilitation is proportional to the strength of the connection between the prime and the target nodes; therefore, SA can naturally account for the positive relation between the magnitude of the priming effect and the strength of the semantic connection. As activation propagates beyond the immediate neighbors of the prime, SA accounts not only for priming based on direct semantic relatedness but also for mediated priming. Moreover, as the level of activation is reduced by distance from origin, SA correctly predicts that the magnitude of the mediated priming effect should be smaller than that of direct priming.

Spreading activation theories often do not make a clear distinction between semantic and associative connections and exploit the same mechanism to account for both semantic and associative priming (Lucas, 2000). As the reciprocal connections between two nodes in the SA network need not be equal, asymmetric priming can readily be produced. However, differences between semantic and associative priming cannot be accounted for by SA theories due to the lack of distinction between these two types of connections. In addition, backward priming, which is evident for items with no prime-to-target relations, could not be easily explained by SA mechanisms because only a unidirectional connection between the corresponding nodes should be present in such cases. Finally, SOA effects on priming are predicted only when assuming that the typical time of activation spread is in the order of hundreds of milliseconds. If, however, activation spreads very quickly (e.g., Lorch, 1982; Ratcliff & McKoon, 1981), no such effects should exist.

### *2.3. The attractor-network account*

Attractor networks account for semantic priming by relying on the fundamental assumption that semantically related concepts have correlated representations (e.g., Masson, 1995; Moss et al., 1994; Plaut, 1995; Plaut & Booth, 2000). When the prime is presented, the activity pattern of the network begins to converge on its corresponding attractor. As distributed representations imply, by definition, that neurons are shared among different memory patterns, convergence to the attractor representing the prime necessarily activates some of the neurons included in the memory patterns that represent its related concepts. As a result, when the target is presented and the network begins traveling toward its attractor, fewer neurons will have to change their activation status when the transition is from a prime to a related (correlated) target compared to when it is to an unrelated target. As the prime pattern constitutes an attractor of the network dynamics, it tends to resist changes in neuronal activation applied by the presentation of the target; therefore, the fewer the neurons which need to change their status during the transition, the less resistance would the transition face and, consequently, the faster the network would converge to the target's attractor, reflecting its recognition. Hence, neural networks with distributed representations and attractor dynamics can easily account for the acceleration of word recognition in semantic priming experiments.[1] In addition, if stronger semantic relatedness is interpreted as stronger correlations,

the above account can easily explain the dependence of priming on the strength of the prime-target semantic relations.

Attractor networks, however, struggle to explain several specific features of priming. First, as indirectly related prime-target pairs should have uncorrelated representations (as there are no direct relations between them), mediated priming is not easily accounted for by such models. Second, as the dynamics ends with the convergence on the prime, it places strict limitations on the time window during which SOA can influence priming unless the convergence process is assumed to last for seconds. Third, as priming in such networks depends on correlation, which is a symmetric trait, they cannot easily account for asymmetric priming. Finally, relying on pattern correlations as the main explanatory tool, attractor models cannot distinguish between semantic and associative connections.

Some of the above deficiencies have been addressed in various ways. For example, several models (Moss et al., 1994; Plaut, 1995) showed that associative relations can be formed in the network independently of semantic relations. If, while the network is trained, certain concepts are coupled so that they frequently appear in succession, the network may learn this temporal consistency. Later, in a priming simulation, when the same concepts appear as prime and target in the order they were learned, the network would tend to converge on the target faster than when an unassociated pair is used, thus demonstrating associative priming independent of semantic relations. It also allows associative priming to be asymmetric and increase with SOA (Plaut, 1995). However, asymmetry in pure semantic priming, which is still based on correlations in these models, cannot be explained by this mechanism.

An additional attempt to address a weakness of attractor network models of semantic priming regarded the effect of mediated relationships. Some authors postulated that indirectly related word pairs actually have a weak direct relatedness between them, allowing mediated priming to occur in attractor networks much in the same way as direct priming (e.g., Jones et al., 2006; McKoon & Ratcliff, 1992; Plaut, 1995). This suggestion, however, is challenged by several empirical findings (e.g., Jones, 2012) and its validity is strongly debated in the literature (see, e.g., McNamara, 2005). All in all, attractor dynamics seems to lack some of the flexibility that SA dynamics offers and, consequently, falls short in accounting for various priming results which SA models can comfortably address.

## 3. The current model

Following the traditional separation between levels of processing (e.g., Borowsky & Besner, 1993; Smith, Bentin, & Spalek, 2001), we speculate the existence of three different computational levels, represented by three networks: orthographic, lexical/phonological, and semantic. In line with other connectionist models (e.g., Huber & O'Reilly, 2003; McClelland & Rumelhart, 1981; Seidenberg & McClelland, 1989; see also Plaut, 1997), we assume that visual input containing words activates the orthographic layer, where letters are identified. The output of this process is fed into the lexical/phonologic network where real words are recognized and fed forward to the semantic network where the word's meaning is

represented. Importantly, these processes are interactive all the way down: The semantic network can influence the lexical network by feedback, and so is the case between the lexical and the orthographic networks.[2] Top-down effects contribute to semantic priming: When a newly arrived word (the target) is related in some way to a word which the semantic network is ''tuned'' on (the prime), the lexical network can recognize this target faster than if the prime and the target are not related because both the bottom-up and the top-down streaming contribute to the recognition process. When a neutral stimulus (i.e., a stimulus which does not represent a word) is presented to the lexical network, neither the lexical nor the semantic network is activated and no information transfer occurs (see McNamara, 2005; for a similar conceptualization in an interactive-activation model)[3].

Being concerned here primarily with semantic effects, we fully modeled and simulated only the lexical and semantic networks, as they are directly involved in the manifestation of this phenomenon. All other processes, including the visual input, the activity in the orthographic network, and its output, were unified to a simple external bottom-up input arriving to the lexical network. The lexical and semantic networks were modeled as attractor neural networks with sparse, binary representations and continuous-time dynamics (see Hopfield, 1982, 1984; Tsodyks, 1990). Sparse representations have lately been supported by neurophysiologic evidence and are considered the rule in cortical codes (e.g., Waydo et al., 2006).

### *3.1. The semantic network*

#### *3.1.1. Basic properties*

The semantic network in our model is a fully connected recurrent network composed of 500 units (''neurons''). Memory patterns encoded to the network, representing concepts (here-to-end labeled ''concept patterns''), are binary vectors of size 500, with ''1'' indicating a maximally active unit and ''0'' an inactive one. The representations are sparse (i.e., a small number of units are active in each pattern) with $p$ being the ratio of active units ($p << 1$, equal for all patterns). When an external input attempts to activate units that are part of a specific memory pattern in the network, the activity of the entire network is driven by the internal connectivity to gradually converge on this pattern. The connectivity matrix between the units assures the patterns' stability. External inputs are always excitatory.

The units themselves are analog with activity $x_i$ in the range [0,1] and reach binary values when converged on one of the memory patterns. The activity of the $i$th unit obeys a logistic transfer function of the form:

$$x_i(h_i) = \frac{1}{1 + e^{-\frac{h_i}{T}}} \tag{1}$$

With $T$ being a gain parameter,[4] $h_i$ represents a low-pass filtered version of the instantaneous local input to the unit. Following Herrmann, Ruppin, and Usher (1993), the local input obeys the following linear differential equation:

$$\tau_n \frac{\mathrm{d}h_i}{\mathrm{d}t}(t) = -h_i(t) + \sum_{j=1}^{N} J_{ij}(t)x_j(t) - \lambda(\bar{x}(t) - p) - \theta + \left[I_i^{\mathrm{ext}}(t) - \theta^{\mathrm{ext}}\right]_+ + \eta_i \quad (2)$$

Here, $\tau_n$ is the time constant of the unit, $x_j$ is the activity of the $j$th unit (with $\bar{x}$ indicating average over all units), $N$ is the number of units (500 in our case), $p$ is the sparseness of the representations mentioned earlier, $\lambda$ is a regulation parameter which maintains stability of mean activation, and $\theta$ is a constant unit-activation threshold, which can also be seen as global inhibition (see Herrmann et al., 1993, for details). The $[...]_+$ symbol indicates a threshold linear function, such that $[x]_+ = 0$ for $x < 0$, and $[x]_+ = x$ otherwise. The use of this function allows the external input to the unit, $I_i^{\mathrm{ext}}(t)$, to influence the network activity only if it surpasses some constant external threshold $\theta^{\mathrm{ext}}$. Finally, $\eta_i$ is a noise term drawn from a Gaussian distribution with standard deviation $\eta_{\mathrm{amp}}$ and temporal correlations $\tau_{\mathrm{corr}}$ (see details later). The (maximal) connectivity matrix of the network is determined according to a Hebbian-inspired rule (Tsodyks, 1990):

$$J_{ij}^{\max} = \sum_{\mu=1}^{P} \frac{(\xi_i^{\mu} - p)(\xi_j^{\mu} - p)}{Np(1-p)} \quad (3)$$

In(3), $P$ is the total number of memories encoded into the network, and $\vec{\xi}^{\mu}$ is the $\mu$th memory pattern.

Relatedness between concepts is implemented in the model as correlations between memory patterns (reflecting the degree of overlap between them), defined for two patterns, $\vec{\xi}^{\mu}$ and, as:

$$m(\vec{\xi}^{\mu}, \vec{\xi}^{\upsilon}) = \sum_{i=1}^{N} \frac{(\xi_i^{\mu} - p)(\xi_j^{\upsilon} - p)}{Np(1-p)} \quad (4)$$

The higher two concepts are related, the stronger their correlation is; unrelated patterns have a correlation near 0.

### *3.1.2. Latching dynamics*

In most attractor networks which were used to simulate semantic priming, the dynamics lead the network to converge to a certain pattern from which only a new external input could drive it away. A few studies, however, have suggested the possibility of an additional, long-term dynamical process (compared with the relative short one which governs the convergence phase) based on neuronal adaptation mechanisms. Experimentally, adaptation mechanisms have been assumed to take part in several cognitive functions operating on different levels of processing and a variety of timescales, ranging from visual mechanisms (e.g., perceptual priming; Huber & O'Reilly, 2003) to lexical mechanisms such as phonetic-to-lexical processing (e.g., the verbal transformation effect; Warren, 1968) and lexical-to-semantic processing (as in the semantic satiation effect; Amster, 1964; Lambert & Jakobovits, 1960; Tian & Huber, 2010). Many of these effects were shown to be captured by network

models implementing neural adaptation between different computational layers (e.g., Huber & O'Reilly, 2003; see General Discussion for more details). In attractor networks with multiple steady states, adaptation mechanisms can prevent units from maintaining a constant firing rate and make the network unable to hold its stability for long. Therefore, the network state autonomously leaves the initial attractor and converges to a different one. The process may repeat again and again with the network ''jumping'' from one attractor to another, simulating what may be seen as free associations. This type of jumping was termed ''latching dynamics'' by Treves (2005) and was investigated by his group (e.g., Kropff & Treves, 2007; Russo, Namboodiri, Treves, & Kropff, 2008), as well as by others (Herrmann et al., 1993; Horn & Usher, 1989; Kawamoto & Anderson, 1985). Mechanisms that cause adaptation can range from dynamic thresholds (e.g., Herrmann et al., 1993) to dynamic synapses (e.g., Bibitchkov, Herrmann, & Geisel, 2002). It was also found that there is a greater tendency for network transitions between correlated patterns than between uncorrelated ones. This bias occurs because neurons in attractor networks are typically noisy and hence do not adapt at exactly the same rate. Consequently, when some neurons are already incapable of maintaining their activity due to fast adaptation, other neurons belonging to the same memory pattern may still maintain their activity. As a result, the network leaves the original attractor and settles to a correlated attractor in which the slowly adapting neurons are still active (see Herrmann et al., 1993).

We implemented adaptation in the semantic network using short-term synaptic depression. This process has been shown to exist in cortical synapses (e.g., Tsodyks & Markram, 1997) and is thought to have several computational advantages (e.g., Pfister, Dayan, & Lengyel, 2010). Yet other adaptation mechanisms would have led to similar results.

Short-term synaptic depression was modeled according to Tsodyks, Pawelzik, and Markram (1998). In line with this model, the synaptic efficacy of each unit (i.e., the efficiency of its synaptic transmission to other units) decreases linearly with its activity:

$$\frac{\mathrm{d}s_i(t)}{\mathrm{d}t} = \frac{1 - s_i(t)}{\tau_r} - U x_{\max} x_i(t) s_i(t) \tag{5}$$

Here, $s_i$ is the synaptic efficacy of the $i$th unit, $\tau_r$ is the time constant of recovery of the synaptic efficacy, and $U$ is the utilization of the available synaptic resources. The term $x_{\max}$refers to a hypothetical maximum firing rate of a unit (e.g., 100 spikes/sec), and it was needed because in the original equations (Tsodyks et al., 1998), the firing rate of the units was not bounded by the range [0,1] as it was in our case. The synaptic strength for a given efficacy at a given time is determined as the maximal weight multiplied by the efficacy:

$$J_{ij}(t) = J_{ij}^{\max} s_j(t) \tag{6}$$

The result of adding short-term synaptic plasticity to the units is that the stability of a pattern cannot be maintained by the network for long. This is because the efficacy of both the

excitatory synaptic connections among the active units in the pattern and the inhibitory connections from the active units to silent units decreases with time. Consequently, after a given time, the network will leave the attractor and converge to a different one. During this time, depleted synapses have the opportunity to recover.

#### *3.1.3. Noise*

The noise term in our network, $\eta_i$, is drawn from a normal distribution with temporal correlations, independently for each unit. Temporal correlations on the order of tens of milliseconds are evident in physiologic data and may reflect filtering processes associated with synaptic integration (Zador, 1998). In addition, synchronous activity of external networks may also lead to temporal correlations in the noise, which may have important computational consequences (Mato, 1999). In our model, the correlations cause occasional ''drifts'' in the units' activity consisting of noise-driven sporadic rises or decreases which last for more than a few milliseconds (in contrast to the white noise case, where the lack of temporal correlations allows only instantaneous sporadic changes). These drifts are important as they allow a wide variety of transitions between patterns induced by the latching dynamics. Although typically the network jumps from one memory pattern to a strongly correlated one, it could nevertheless perform occasional transitions to less strongly correlated patterns. If the temporal correlations were set to zero, transitions were almost always from one pattern to the one most correlated to it. In addition, the noise amplitude itself influences the stability of the attractors and, consequently, affects the rate of transitions in the semantic network. Its value was set to allow a transition rate which fits previously published experimental results. Different values can slow or even halt transitions.

### *3.2. The lexical network*

Like the semantic network, the lexical network in our model is fully recurrent and comprised 500 units. We labeled the memory patterns in the lexical network as ''word patterns.'' The equations governing its dynamics are similar to those of the semantic network, with two important changes:

1. There are no correlations between the word patterns in the lexical network. This is not meant to indicate that there are no lexical relations in natural languages (indeed, such relations obviously exist, at least at the phonological level, e.g., ''rat''–''bat,'' ''cable''–''table''), but merely to ensure that such relations would not add unnecessary noise to our simulations. In fact, typical semantic priming experiments control for such possible confounds by selecting prime-target pairs that bare no lexical/phonological relations within a pair. Anyhow, this is a simplification which should not influence the average pattern of results.
2. The lexical network does not implement latching dynamics. This is another simplification, which further reduces the variability in the lexical network to allow emphasizing the effect of the semantic network on lexical convergence. Moreover, from a conceptual point of view, it stands to reason that semantic networks are more associative in

nature than lexical networks, as indicated by association norms (e.g., Nelson et al., 2004); free associations are based more on the meaning of words rather than their lexical/phonologic properties. In practice, latching dynamics was eliminated in the lexical network by decreasing the rate of the synaptic depression, $U$, to a very small level. Indeed, although there is still no direct evidence of systematic differences in synaptic depression between brain regions, Tsodyks and Markram (1997) have found that there is a wide variety of synaptic depression rates among neocortical neurons which strongly affect their computational properties.

### *3.3. Connectivity between the networks*

The links between the lexical and semantic networks are based on connections between active units in corresponding patterns (Fig. 2). An activated unit belonging to a certain word pattern in the lexical network sends excitatory connections to all active units in the corresponding concept pattern of the semantic network and vice versa. Given the distributed nature of the semantic representations and the correlations in the semantic network, the activation of one word pattern in the lexical network activates to different extents all semantically related concept patterns in the semantic network. This partial activation is fed back to the corresponding word patterns in the lexical network and adds to its activation by the bottom-up input from the orthographic network. The bottom-up input is also excitatory and determines to which word pattern the lexical network will converge by influencing only the corresponding active units in this pattern.

To allow some separation in the computational processes within each layer, the semantic and lexical networks respond to external inputs if, and only if, they surpass a certain threshold (see Eq. 2). Lexical-to-semantic connections are set to be stronger than the semantic-to-lexical connections. The logic for this asymmetry is that in word-recognition experiments, the required behavior is governed by stimulus-dependant processes, which encourage

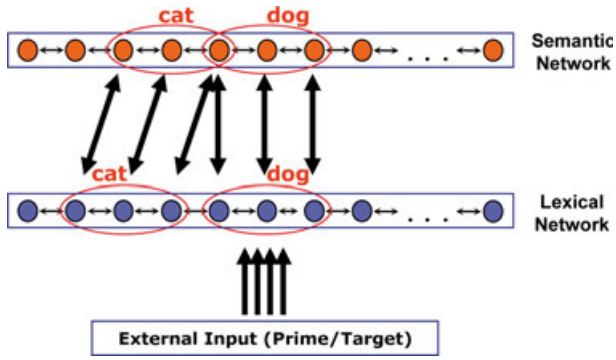


Fig. 2. The architecture of the model. Patterns representing related concepts are correlated in the semantic network but uncorrelated in the lexical network. Active units of two toy example patterns representing ''dog'' and ''cat'' are marked. Connections between networks are from active units of a pattern in one network to all the corresponding active units in the other network. For simplicity, only some of these connections are drawn.

bottom-up information transmission. Top-down processes (like the influence of semantics on lexical access) are not essential and can be readily reduced by scaling down the appropriate connection strength.[5] This reduction, however, is not absolute and still allows for some top-down semantic influences as the ones causing semantic priming. As a consequence of this asymmetry, the lexical network affects the activity in the semantic network more quickly than the other way around (the threshold could be surpassed more easily due to the stronger connections) and allows it to be influenced by the semantic network only while it is also activated by the bottom-up external input.

To further increase the independence of each layer, the lexical-to-semantic connections are also subject to synaptic depression with a slow recovery time (for a similar approach, see Huber & O'Reilly, 2003). This causes the bottom-up influence of the lexical network to diminish after a typical time interval, letting the semantic network engage in latching without further disturbance (until a new bottom-up external input arrives and the lexical network converges to a new pattern). Nevertheless, we assume only minimal suppression of semantic-to-lexical connections, as these links are, as described above, weak in the first place. In addition, the bottom-up external input to the lexical network is modeled as constant for as long as a word is assumed to be visible, and it diminishes abruptly when the visual word disappears.

### *3.4. Basic behavior of the lexical and semantic networks*

Fig. 3 demonstrates typical examples of one-trial activations of the network in a typical semantic priming simulation. Correlation of the activation pattern along time for each network with each of its stored patterns (including the real memory patterns and the neutral one) during a trial is presented in that figure in different colors, and convergence to a specific pattern is indicated by its number appearing on top. The lexical network follows the external input by converging to the corresponding memory pattern and keeping stability until a new input arrives. In contrast, the semantic network converges to the appropriate memory pattern, only to jump to other attractors in a serial manner, hence presenting latching dynamics. When a new external input arrives, the semantic network stops its transitions and quickly converges to the corresponding new memory pattern a little after the lexical network has done so (its reaction is much quicker than the lexical network's due to the strong lexical-to-semantic connections).

## 4. Simulation 1

This simulation examined the free dynamics of the semantic network and its relation to SA.

### *4.1. Method*

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

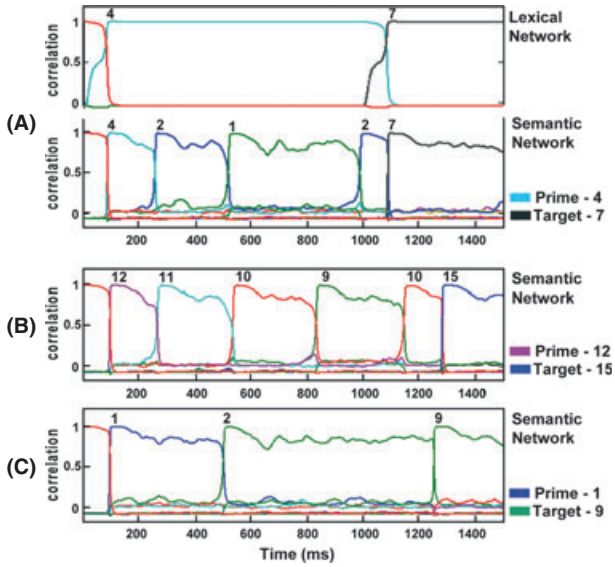


Fig. 3. Correlation of the network state with its different memory patterns as a function of time. Each pattern is indicated by a line with a different color (not all correlation lines are visible at all times, as often they coincide). Moment of convergence to a specific pattern is indicated by the corresponding pattern number above the appropriate line. (A) Typical dynamics of the semantic and lexical networks. The semantic network presents latching dynamics, while the lexical network is stable. (B and C): More examples of the dynamics of the semantic network, showing the stochasticity of transitions.

### *4.1.1. Encoded patterns*

Seventeen different memory patterns were encoded in the semantic and lexical networks. Within each network, these patterns comprised binary vectors with equal mean activity and a very sparse representation.[6] In the semantic network, the basic correlations between patterns were a priori set as following (Fig. 4A): Four groups, each containing four patterns, formed ''semantic neighborhoods'' (patterns 1–4, 5–8, 9–12, and 13–16), so that each pattern in a neighborhood was correlated with the other patterns in its neighborhood, and with few exceptions (see below), no correlations existed between the neighborhoods.[7] All correlations within a semantic neighborhood were equally strong. The 17th memory pattern was a ''baseline'' pattern to which the network was initialized at the beginning of each trial, and it was not correlated to any of the other patterns. This baseline ensured that the network would not readily converge to one of the ''real'' patterns when the trial begins and its stability allowed the network to maintain activity until the prime began influencing it (see Rolls, Loh, Deco, & Winterer, 2008 for another example of modeling baseline activity as a stable state of the system).

To produce indirect relatedness (in addition to direct relatedness) in the semantic network, we modified the above basic encoding structure so that a correlation was introduced between a pattern in one neighborhood and a pattern in a different neighborhood. This correlation was based on other units than the ones forming the correlations within each neighborhood. For example, whereas patterns 1–4 formed a semantic neighborhood and patterns 9–12 formed a different semantic neighborhood, we slightly changed the encoding of

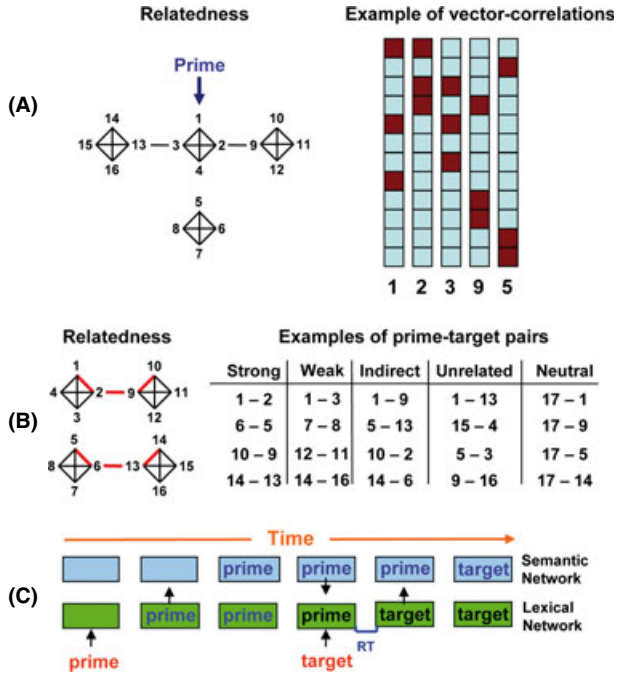


Fig. 4. (A) Patterns used in Simulation 1. The left column shows the semantic organization of patterns by neighborhood. The right column presents a simplified illustration of the relatedness as represented by vector correlations in the network for several representative concepts (brown/light blue colors representing values of 1/0). (B) Patterns used in Simulation 2. The left column shows the semantic organization of patterns by neighborhood, with weak relatedness indicated in black lines and strong relatedness by red lines. The right column presents examples of prime-target pairs used in the simulation trials, organized by relatedness condition. (C) Example of expected chain of events in a semantic priming simulation. Lexical network converges to the prime pattern, followed by convergence of the semantic network. When target appears, the lexical network converges to the appropriate target pattern under the influence of the semantic network. No latching dynamics is assumed in the example.

patterns 2 and 9 to introduce a correlation between them (see the vector examples in Fig. 4A). Consequently, patterns 1 and 9 became indirectly related (mediated by pattern 2). Similarly, we correlated pattern 3 with pattern 13, which resulted in an indirect relatedness between patterns 1 and 13.

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 any of the 17 patterns in the semantic network (thus forming a ‘‘neutral’’ pattern; see Simulation 2 for a more extensive discussion of ‘‘neutral’’ patterns).

*4.1.2. Experimental procedure*

The simulation comprised 100 trials. Each trial began with the lexical and semantic networks converged on their respective neutral patterns. An external input (always pattern 1) was presented to the lexical network immediately after the trial began. This input was a

binary vector corresponding to the appropriate memory pattern of the lexical network (1's in the to-be activated units, 0's in the rest). One hundred milliseconds after trial onset the external input was removed and the network's activity followed the dynamic equations without further interference, for a total period of 3,000 simulated milliseconds (4,500 numeric steps). No additional input was presented in the current simulation. Correlation of the momentary state of each network with each pattern, for each time point along each trial, was stored and averaged offline.

### *4.2. Results*

The mean correlation between the state of the semantic network and each of its encoded memory patterns was computed for each time point over trials. Fig. 5 presents these

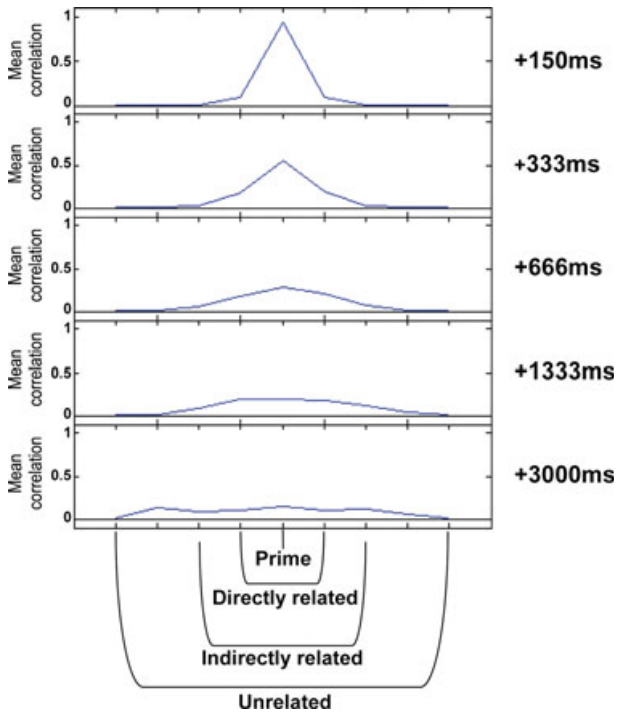


Fig. 5. Spreading activation behavior of the semantic network in Simulation 1. Mean correlation of various memory patterns with the state of the network is displayed at five points in time after prime onset. Middle of the *x*-axis corresponds to the prime pattern (pattern 1), to its right and left are two of its related patterns (pattern 2 and 3), and next are the indirectly related patterns (patterns 9 and 13). Further to the right and left are double-indirect patterns (patterns 12 and 16), and on the edges are two unrelated patterns (patterns 5 and 6). Because of finite size effects of the network, unrelated patterns do not yield an exact 0 correlation, and the displayed results are corrected for such bias. The mean correlation presents a ''spreading out'' behavior, initially concentrated on the prime, then spreads to the related and indirectly related patterns, and finally distributed evenly between many patterns. In very large networks, the final stage is expected to be distributed between many more patterns, yielding an insignificant correlation to each of them, a situation resembling the activation dying out.

correlations for five different time points after the prime onset. On the *x*-axis, the middle point shows the mean correlation with the actually presented prime concept, the two adjacent points to left and right of the center show mean correlations with two of the neighbor patterns (patterns 2 and 3), the two points further to the right and left show the correlation with two indirectly related concepts (patterns 9 and 13), and the two most extreme points show the correlation of the network state with two unrelated concepts (patterns 5 and 6). As can be seen in this figure, the mean correlations followed the principles of SA. 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 (due to the latching dynamics), the mean activation of the initial concept decreases, while the related concepts are activated more and more. Indirectly related concepts show some activation, with a peak rising later. Unrelated concepts receive no activation at all throughout this period. After 3,000 simulated milliseconds, the mean correlation with each of the network's patterns is distributed more or less equally, corresponding to a nearly deactivated state of the whole network (which can be seen as a statistical implementation over aggregated trials of the dissipation of activation with time which characterizes SA theories).

## 5. Simulation 2

As elaborated in the Introduction, SA and attractor networks have often been contrasted using the semantic priming paradigm. Whereas semantic priming in SA stems from the existence of connections between related concept nodes, attractor networks mostly attribute priming effects to correlations between patterns. These two explanations differ, among other things, in their prediction about mediated priming: As mediated items are not related directly, they should not be correlated; therefore, simple attractor networks (without latching dynamics) are unable to account for mediated priming results. SA, on the other hand, allows activation to spread for long distances in semantic space and, therefore, predicts mediated priming. In the present simulation, we tested whether our model yields basic semantic priming effects and how different prime-target relations, including mediated relations, modulate priming. In particular, we explored whether the pattern of priming effects in the simulation corresponds with findings previously reported in human studies.

### *5.1. Method*

The general methods were similar to those used in Simulation 1.

#### *5.1.1. Strong versus moderate and direct versus indirect relations*

To produce varying degrees of direct relatedness, we changed the encoding of two specific patterns in each neighborhood (e.g., patterns 1 and 2), so that their correlation was higher than all the others within the neighborhood (e.g., the correlations between patterns 1 and 3, 1 and 4, 2 and 3, 2 and 4). Independent units were used to produce this additional

correlation so that it would not interact with any already-encoded correlation. This procedure resulted in two levels of relatedness within a neighborhood.

Mediated priming was produced like in Simulation 1. Specifically, we introduced a correlation between pattern 2 and pattern 9, and between patterns 6 and 13. Consequently, patterns 1 and 9 and patterns 5 and 13 became indirectly related (mediated by patterns 2 and 6, respectively; see Fig. 4B).

#### *5.1.2. Experimental procedure*

Each trial consisted of the presentation of two inputs, a prime followed by a target, each being one of the pre-encoded lexical patterns. The relatedness between the prime and the target could be strong, moderate, indirect, or unrelated. For example, as patterns 1 and 2 were a strongly correlated pair within the semantic neighborhood of patterns 1–4, and pattern 2 was also correlated outside its neighborhood to pattern 9, then presenting the patterns 1 and 2 as prime and target, respectively, formed a strong and directly related condition, 1 and 3 a moderate and directly related condition, 1 and 9 an indirectly related condition, and 1 and 16 an unrelated condition (see examples for all experimental conditions in Fig. 4B). In addition, a neutral condition was presented with the prime being pattern 17 and the target being any of the ''real'' word patterns (1–16). Since no connections exist between the neutral patterns of either network, this condition was, in fact, equivalent to not presenting the prime at all.[8] Primes and targets were randomly chosen from within the possible combinations for each condition, with 100 trials in each condition.

Each trial started with the presentation of an external input to the lexical network which served as ''prime.'' After 100 simulated milliseconds, this external input was removed, and a new external input corresponding to the target was presented to the lexical network with 250-ms SOA (cf., Balota & Lorch, 1986). The RT to a target was measured from its onset until the convergence of the lexical network (''correct'' convergences to the target attractor were always achieved). Convergence was defined as the network's state reaching a 0.95 correlation with the relevant memory pattern. Fig. 4C presents an example of this chain of events in a non-neutral trial (for simplicity, no semantic transitions were assumed in this example).

### *5.2. Results*

The lexical network's RT was computed separately for each prime-target relatedness condition (Fig. 6A).[9] RTs were shortest for the strongly related pairs ($M$ = 47.81 simulated milliseconds), followed by the moderately related pairs ($M$ = 64.81 ms), the indirectly related pairs ($M$ = 79.1 ms), and the unrelated and neutral pairs ($M$ = 90.27 and $M$ = 88.13 ms, respectively). Fig. 6B presents the main facilitation effects (relative to the neutral condition) compared with data from human experiments (taken from Lorch, 1982; Balota & Lorch, 1986). Mirroring the empirical findings in humans, there was a gradient of the priming magnitude which decreased both with semantic relatedness and with directedness of this relation.

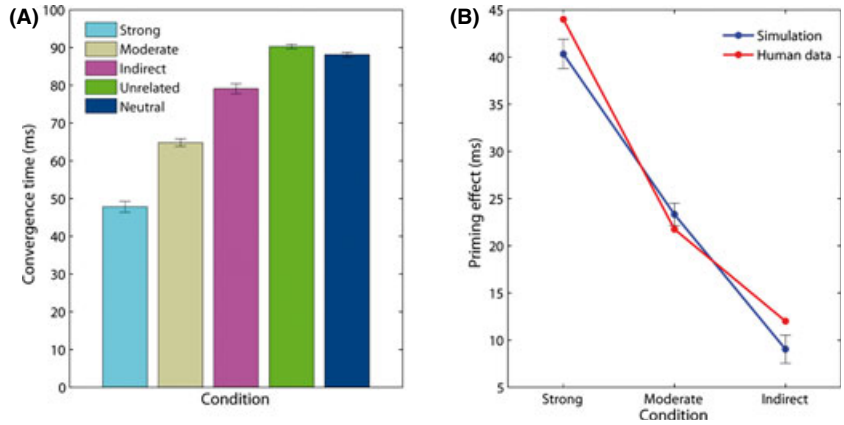


Fig. 6. (A) Mean convergence time of the lexical network for the various conditions of Simulation 2. (B) Facilitation effects in Simulation 2 compared with human experiments. Human results taken from Lorch, 1982 (table 2, 150 and 300 ms SOAs; table 3, 200 and 400 ms SOAs; high and low dominance exemplars representing strong and moderate connections); and Balota & Lorch, 1986 (table 2, 250-ms SOA only). Error bars represent ±1 standard error of the mean.

### *5.3. Discussion of simulations 1 and 2*

The results of the first simulation showed how an attractor neural network with latching dynamics can implement SA on average over trials. This pattern is manifested in the semantic network and it is based solely on its characteristics, with no dependence on the characteristics of the lexical network (or even its mere existence). The only condition for activation to ''spread'' from a particular network pattern to its ''neighbors'' is that the input to the semantic network would diminish and let the latching dynamics run freely (as was implemented in the present model by postulating short-term synaptic depression in lexical-to-semantic connections). The pace at which the activation spreads depends on the rate of transitions from one attractor to another during the latching dynamics. The faster the transitions are the faster activation spreads.

As mentioned above, latching may be achieved by various mechanisms. Yet the general characteristics of the SA as revealed by the present simulation did not depend on the specific adaptation mechanism which has been implemented. Rather, the most important factor which determines where and how quickly the activation spreads is the correlation between the patterns. Transitions occur more frequently between correlated patterns and within a semantic neighborhood than between uncorrelated patterns contained in different neighborhoods (Fig. 3). In other words, the correlations between concept patterns play the role played by the connections' strength between various word nodes in the original SA theory. However, the probability of the network to jump from one pattern to another is not simply determined by the correlation strength between two concepts. Rather, this probability is determined by the relative strengths of the correlations that a particular concept pattern has with *all* the other concept patterns. For example, if one pattern (''pattern A'') has a 0.1 correlation with another pattern (''pattern B'') and no correlations with any other pattern,

there is a high probability that the network would jump from pattern A to pattern B. On the other hand, if pattern A is also correlated with another pattern, C, with a 0.2 correlation, the most probable jump would be from A to C rather than from A to B; the probability of the jump from A to B is reduced, even though the strength of the corresponding correlation is the same. This characteristic of our model resonates with Anderson's model of SA (1983), in which the connection strengths between nodes are scaled by the total amount of connections (a hypothesis which is essentially required to account for other linguistic phenomena, such as the fan effect; see Anderson, 1974). It is also in line with some spreading activation models in which the connection strengths are determined by free-association norms (e.g., Spitzer, 1997). In such models the connection strength from *dog* to *cat*, for example, is determined by the probability that cat will be the first association of dog in a free-association task. The total strength of connections from dog to all of its associates must sum up to 1 (as it represents probability), and therefore, each of the connections is influenced by all the others. This is exactly what should be expected from the way SA is implemented in our model, which, being expressed on average over trials, provides, in fact, a natural scaling for the connections' strength. Indeed, one could see the activity of nodes in the original SA model as an average manifestation of both the correlations between the patterns in our semantic network and the probability of associative occurrences achieved from association norms. As will be demonstrated in Simulation 3, the probability of a transition from one pattern to another in our model resembles, in principle, the probability of a corresponding association between two concepts in free-association norms.

There is, however, an important distinction between the average performance of the semantic network in our model and that of the original SA model. In our network, spreading is temporarily interrupted by 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 reflect the dynamical transitions from one attractor to another. This implies that at very short SOAs, the ''spreading'' is actually entirely dependent on the network's correlations as it relaxes on the prime's attractor and may, therefore, seem instantaneous with respect to the prime's immediate neighbors (cf., Plaut, 1995). Only at longer SOAs can transitions participate in the dynamics and allow spreading to carry on. Thus, immediate and distant (i.e., indirect) neighbors in our model have different status in terms of the activation spread, in contrast to the classical SA in which no such distinction exists.

The results of the second simulation demonstrated how the dynamics in the semantic network affects the convergence time of the lexical network. As can be seen in Fig. 6, mimicking semantic priming effects in humans, the time needed for the convergence of the lexical network on the target's word pattern is shorter if prior to its appearance the semantic network converged on a concept pattern that is correlated (i.e., related) to that target's concept. This result is achieved because a number of units that are activated in the semantic network are connected to the units that would be activated by the target pattern in the lexical network, and this partial top-down pre-activation facilitates the convergence of the lexical network on the target. As the magnitude of facilitation is proportional to the amount of shared units, the stronger the prime and the target concept patterns are correlated in the semantic

network, the faster would the target ''recognition'' by the lexical network be (compare the turquoise and the beige bars in Fig. 6A). Unrelated prime and target patterns do not share any active units and, indeed, did not facilitate each other.

In addition to direct priming, mediated priming effects were also apparent in Simulation 2. These effects stemmed from trials in which the semantic network committed a transition from the prime's pattern to another pattern before target onset. Often, this new pattern is correlated both to the prime and to the upcoming target (as in the example of *lion* being related to *tiger,* which is related to *stripes*). Consequently, when the target appears, the semantic network will already be converged to a pattern correlated to it (even if the original prime is not). This latter correlation yields the observed mediated priming effect.

As the semantic-to-lexical connections in our model are purely excitatory, we expected only facilitatory effects and, therefore, the unrelated and neutral targets, both representing patterns uncorrelated to the prime, should have yielded equivalent convergence times. This was, indeed, the general observed result of the simulation: Although the unrelated and the neutral conditions were not identical,[10] the difference between these two conditions was very small, an order of a magnitude smaller than the facilitation effect (Fig. 6A). Therefore, our results, based on our implementation of neutral trials, are in agreement with the view that automatic priming mechanisms primarily contribute to facilitation of related targets.

## 6. The emergence of asymmetric priming effects and their modulation by SOA

In the previous simulations, we showed how the combination of correlation between concept patterns and synaptic adaptation in the semantic network yields a spreading-activation-like dynamics, allowing for both direct and mediated priming to emerge. Next, we will show how these two mechanisms combine in sophisticated ways to account for previously reported asymmetry in associative relations. We will demonstrate how this asymmetry is modulated by SOA and how it can be related to the difference between semantic and associative priming, as well as to backward priming.

There is an important distinction between the correlation of two patterns and the probability of transitions between them. This can be shown by a simple example: Imagine that pattern B is correlated to the same degree to both patterns A and C, whereas neither A nor C is correlated to other patterns. If the network rests on pattern A and then jumps, it will almost definitely jump to pattern B, its only correlated pattern. On the other hand, if it rests on pattern B, jumps to A and C would be equally probable. Therefore, although A and B are symmetrically correlated, the transition probabilities from one of them to the other are asymmetric. This example demonstrates the fundamental characteristic of our network's dynamics: The transition probabilities from one pattern to another are influenced by the entire structure of pattern correlations encoded in the network.[11]

In priming experiments, the transition probabilities can influence RT. When the semantic network frequently jumps to the pattern representing the upcoming target (before its actual appearance), the mean RT to the target should decrease considerably (and priming effects should consequently increase) as many more units representing the target concept contribute

to facilitating the convergence of the lexical network. If, however, transitions are mostly to a pattern other than the target, RTs may not change or may even increase, depending on where the network jumped to. Asymmetry in priming may therefore arise as a function of the asymmetry in jumps. This asymmetry should interact with SOA: With short SOAs, there are fewer semantic transitions, and therefore the asymmetry in probabilities has less impact. With longer SOAs, semantic transitions are probable and, consequently, asymmetry emerges.

A common finding in the priming literature is that the size of the priming effect is modulated by SOA to different degrees, depending on the prime-target relatedness type. Whereas the priming effect for associated pairs increases with SOA (de Groot, 1984, 1985; de Groot, Thomassen, & Hudson, 1986; Lorch, 1982; Neely, 1991), semantic priming for related but unassociated pairs is less influenced by SOA and may even decrease for backward-related pairs (Kahan et al., 1999; Lucas, 2000). Several authors suggested that this difference between the SOA effects on associated and unassociated pairs may be attributed to episodically learned connections between linguistic items based on co-occurrence. Episodically learned associations could be formed either between concepts in the semantic system (e.g., Herrmann et al., 1993; Silberman, Miikkulainen, & Bentin, 2005) or between words in the lexical system (Fodor, 1983; Lupker, 1984). These connections were suggested to affect priming mostly at long SOAs, assuming that time allows more efficient processing of the prime, leading to a greater impact of the learned prime-target associations (Plaut, 1995). Alternatively, other authors explained the SOA-dependent increase in priming between associated pairs relying on controlled mechanisms, which supposedly take time to initiate and therefore contribute to priming only at long SOAs (e.g., Neely, 1977; Neely, 1991).

In our model, SOA is expected to influence the magnitude of priming as transitions in the network cause the focus of semantic activation to change, in a given trial, from the prime to its surrounding neighborhood. This SOA dependency occurs without assuming any particular connectivity beyond the correlation structure of the encoded patterns and, most important, is expected to differ for symmetrically and asymmetrically associated pairs because of their different transition probabilities. The difference between associative and semantic priming may therefore be a product of such asymmetry.

## 7. Simulation 3

The goal of this simulation was to explore how a more complex structure of network correlations affects the transitional probabilities of the network and the priming effect, as a function of SOA. The correlation structure was determined by the association norms of four specific concepts within one semantic neighborhood (Nelson, McEvoy, & Schreiber, 1998) and was designed to mimic their mutual free-association probabilities. First, we showed how such probabilities can be roughly implemented in the semantic network despite its small size and small variety of correlation strengths. Second, we verified that several known characteristics of association response times (e.g., Goldstein, 1961; Schlosberg & Heineman, 1950) are roughly reproduced by the transition latencies in our model. Third, we

examined the priming effect and its modulation by SOA for prime-target pairs that differ in their forward and backward transition probability.

### 7.1. Method

#### 7.1.1. Encoded patterns

The present simulation focused on a semantic neighborhood consisting of patterns analogous to four animal concepts—*dog*, *cat*, *mouse,* and *kitten*. Based on the human-derived association norms, the *cat* pattern was strongly correlated to both *dog* and *mouse*, and moderately correlated to *kitten*. All other correlations within the neighborhood were weak. Each animal concept, with the exception of *kitten*, also had idiosyncratic moderate correlations with concepts outside the neighborhood (*kitten* was the exception because, according to the association norms, it hardly has any significant forward or backward connections outside the neighborhood). *Dog*, in particular, had an idiosyncratic correlation with the concept *beware*, which belonged to another semantic neighborhood consisting of *beware*, *danger*, *caution*, and *careful*, all weakly correlated among themselves. The total number of memory patterns encoded in the network (including the baseline pattern) was 17, as in previous simulations. A summation of the structure is depicted in Fig. 7.

#### 7.1.2. Experimental procedure

First, we assessed the probability of associations between the animal concepts as reflected by the first transition from each stimulus. This assessment was based on 4,000 trials in

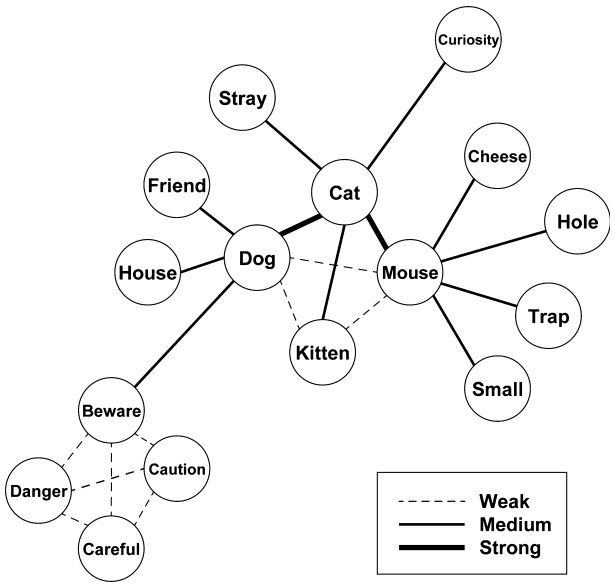


Fig. 7. Structure of the semantic memory used in Simulation 3. Sixteen memory patterns representing 16 different concepts were used, with three degrees of correlation strengths. Width of the connecting lines represents the strength of the correlations.

which each of the four animal concepts were presented 1,000 times to the lexical network as single stimulus. Trials duration was 1 second, to elevate the chances of at least one transition in the semantic network. Identically, 1,000 trials were run with the concept *beware*, to examine its tendency to jump to the *dog* concept. The observed probabilities of these transitions are presented in a color-bar graph next to their real values based on the human association norms (Fig. 8A). As can be seen, albeit not identical, the transition probabilities of the network closely resembled the trends observed in the human data. Specifically, the concept patterns corresponding to *dog* and *cat* were associates of each other, while *dog* was also a moderate associate of *beware*, but not vice versa. The concept *cat* had, in addition, some moderate associations with *kitten* and *mouse*. Of particular interest, the association to *kitten* was dramatically asymmetric, with *kitten* leading to *cat* almost nine times more often than the other way around.

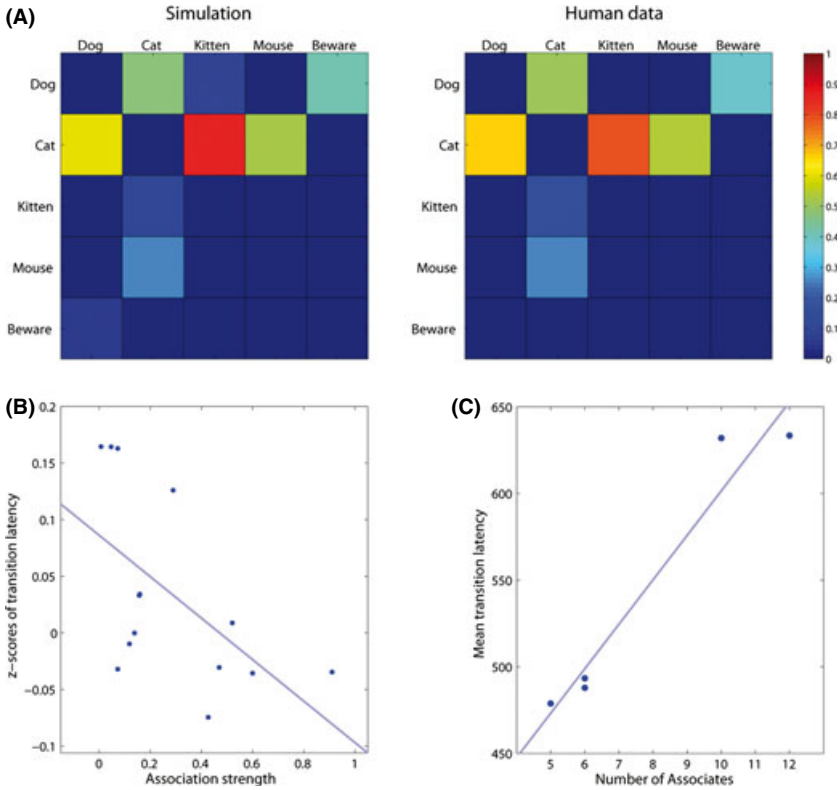


Fig. 8. Behavior of the model using associative and non-associative pairs in Simulation 3. (A) First-transition probabilities of the five main concepts in the network. Probabilities are indicated by colors ranging from 0 (dark blue) to 1 (red). Columns represent the presented words and rows represent their associations. The simulation results are compared with data taken from human association norms (Nelson et al., 1998). (B) Associative strength (measured as frequency of occurrence) versus standardized transition latency of first transitions in the network. (C) Number of associates of the five main concepts in the network versus their mean transition latency.

Next, we calculated the average transition latency of each of the associations produced by the network by computing the time from the appearance of the stimulus until the first transition has concluded, separately for each association. Following this stage, priming was simulated as in Simulations 1 and 2, with a target following a prime at a pre-designated SOA. Several representative prime-target pairs within the animal neighborhood were used, consisting of both high and low mutual transition probabilities. In addition, a couple of pairs consisting of one concept within the neighborhood and one concept outside it were used. For some of the strongly associated prime-target pairs, priming simulations of the corresponding backward direction (where the prime and target switched roles) were also conducted. Finally, neutral trials (with targets chosen randomly from the memory patterns) were conducted for comparison with the related trials. Each priming pair was repeated for 100 trials, and for 7 SOAs equally distributed from 150 to 450 ms.

### *7.2. Results*

#### *7.2.1. Association latencies*

Experimental research on association norms often finds a negative correlation between association strength (defined in terms of frequency of occurrence for a given cue) and association response time (e.g., Schlosberg & Heineman, 1950), as well as a positive correlation between the number of associates of a cue and the average latency of all of its associates (Flekkoy, 1973; Goldstein, 1961). To examine how well the network associations resemble human data, we examined whether these trends exist in our results.[12]

Association strength (defined here as the observed probability of a particular transition) and association latency were not correlated when the entire spectrum of associations from the five stimuli were considered. However, it was evident that the latencies for a particular stimulus were strongly affected by the total number of active units that stimulus shared with all the other concept patterns. Stimuli that shared a large number of units with other patterns (such as *dog, cat,* and *mouse*) were less stable and tended to yield considerably faster transitions than stimuli that shared few units (*kitten* and *beware*). This difference, which blurs the correlation of interest, might not be as robust in the representations of real concepts in humans, and, therefore, it is most probably a confound created by the specific encoding structure used in our simulations (see Discussion). We therefore controlled for the difference in the total number of shared units by computing the *z*-scores of each of the latencies, that is, comparing each raw latency value with the association latencies of its deriving stimulus (e.g., the *z*-score of *cat–kitten* was computed in comparison with the average transition latencies stemming from *cat*, while the *z*-score of *kitten–cat* was computed compared with the average latencies stemming from *kitten*). These latency *z*-scores are plotted against association strength in Fig. 8B. Corresponding to published findings in humans, there was a significant negative correlation ($r = -.57$; $p < .04$) between the measures, indicating that strong associations tended to occur faster than weaker associations.

We also looked at the total number of associates of each of the five patterns as a function of its average association latency (over all the associations stemming from it; naturally, this computation used raw RT values and not *z*-scores). These data are plotted in Fig. 8C. Again,

mirroring findings from human associations, we found a positive correlation between these two measures ($r = .97$). Although only five data points were available in the present data, this correlation was highly significant ($p < .007$).

*7.2.2. Priming effects*

The priming effects of the pre-specified prime-target pairs, as a function of SOA, are presented in Fig. 9. Three notable results are apparent: First, the priming of a target by a strongly associated prime was higher, across SOAs, compared with priming between unassociated pairs (compare the three upper curves to the rest in Fig. 9). Second, the priming effects for strongly associated pairs increased with SOA (Fig. 9, red, blue, and purple curves), while the priming for non-associative pairs was unaffected by SOA or even decreased (cyan, yellow, pink, and green curves). Third, pairs that were asymmetrically associated to each other, as reflected by their transitional probabilities, yielded asymmetric priming, with the asymmetry growing with SOA (compare purple vs. cyan and brown vs. green curves).

*7.3. Discussion*

The results of Simulation 3 demonstrate the importance of the correlation structure between encoded patterns in determining the transition probabilities between concepts and the consequence of these probabilities on priming. Although the structure of the semantic network in our model is determined by the symmetric correlations between concepts, the transition probabilities resulting from this structure are not symmetric, reflecting directional associative relations as well. This characteristic allows the semantic network in the model to

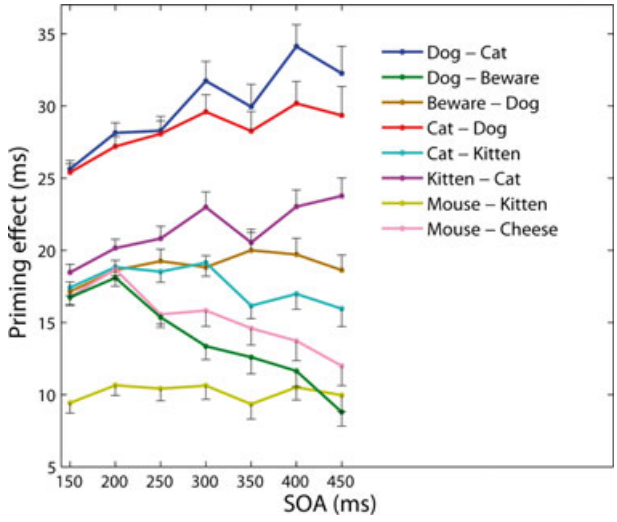


Fig. 9. Mean facilitation effects as a function of SOA for the prime-target pairs presented to the network in Simulation 3. Error bars represent ±1 standard error of the mean.

exhibit complex dynamics yielding asymmetric associative priming as well as an interaction between the effect of associative strength and SOA. The most important conclusion from this simulation is that differences in the association strength between concepts and the resulting differences in their priming patterns can be based solely on the correlation structure of the encoded memories without requiring the formation of explicit associative connections between specific words through designated training (cf., Plaut, 1995).

The observed transition latencies between memory patterns matched several reported characteristics of experimental RTs in free-association tasks (Flekkoy, 1973; Schlosberg & Heineman, 1950). The association strength between concepts was negatively correlated with the *z*-scores of the transition latencies, although not with the absolute latency values. The lack of correlation with the absolute latency values could be explained by the fact that the concept patterns in our simulation were apparently divided into two distinct groups, one which contained patterns sharing a large number of active units with other patterns (*dog*, *cat*, *mouse*) and another which contained patterns with few shared units (*kitten*, *beware*). The difference in average latencies between the two groups was relatively large (∼150 ms) and probably blurred the effect of association strength on latency. Indeed, separate post hoc examinations of the association strength within each of the two groups in isolation yielded negative correlations with the absolute scores just as for the *z*-scores of the combined group (although these correlations did not reach statistical significance due to the small number of data points in each group separately). Whether the discrepancy between the two groups is of theoretical interest remains to be explored. It may be that the difference does not have an analog in the representations of real concepts in the brain and, therefore, should be considered an artifact stemming from the relative small size of our network (that limits the maximum amount of encoded concepts and prevents equating the number of shared units across all patterns). Another possibility, however, is that this difference in the number of shared units does reflect, to some extent, real diversity in the way concepts are represented in the semantic system; in that case, a more comprehensive comparison of the associative latencies may require identifying how the total number of shared units of a pattern in the network maps to known characteristics of real concepts (examples of which may include, perhaps, semantic set size and familiarity). However, as the patterns in our network represent real concepts only crudely, we leave further investigation of this issue to future studies in which a more transparent analogy between concepts and their computational representations will be determined.

Another feature of the transition latencies that resembled human associative RT data was the positive correlation between the average association latency from a stimulus and the total number of associates of this stimulus. That is, the more associates a concept has, the longer the associates' latency is on average. As shown by Flekkoy (1973), this result cannot be only a by-product of the negative association between association strength and association latency. In our network, this correlation stemmed from the fact that the stimuli tended to produce occasional idiosyncratic transitions (i.e., transitions from a stimulus to a concept pattern not correlated to it directly, such as *kitten–beware*), and this tendency differed systematically between the five stimuli. The fewer units a stimulus pattern shared with other concept patterns, the more idiosyncratic transitions it produced. As, as discussed above, concepts that shared few units also tended to yield slower transitions, a correlation emerged.

The results of Simulation 3 did not portray precisely the difference between semantic and associative connections. Rather, they suggest that these two types of connections can be based on a similar underlying mechanism and that the common differentiation between them may not necessarily represent a substantial connectivity difference in terms of the neural network involved. An associative connection between two concepts exists when there is a high probability for a transition from one to the other. Semantic connections, on the other hand, cannot be defined as simply. Specifically, the fact that two patterns are correlated does not necessarily mean that they are semantically related. It might be argued that semantic relations can usually be attributed to a group of concepts with dense interconnectivity (as in the case of the animal concepts which all share some connections), thus forming semantic neighborhoods; but the emergence of a more elaborate semantic structure might be the product of reading out information from the semantic network at higher cognitive stages. Be that as it may, the lack of an absolute definition for semantic relations in our network did not prevent it from portraying the priming results observed in human studies, differentiating between associated- and unassociated-related pairs. Priming of associatively related prime-target pairs was stronger compared with unassociated pairs and also increased with SOA (cf. de Groot, 1985; de Groot et al., 1986; Hutchison, 2003; Plaut, 1995). In contrast, priming of unassociated pairs did not rise with SOA, and for backward-related pairs such as *dog–beware*, it even decreased. This general ''advantage'' of associated over unassociated pairs is caused in our model by two factors: First, with sufficiently long SOAs, the semantic network tends to jump from the prime to an associated target, increasing priming significantly. Second, associatively connected pairs generally (although not always) tend to be more strongly correlated than unassociated pairs, and this allows them to have a higher influence on the lexical network. These findings are reminiscent of the ''associative boost'' effect (e.g., Lucas, 2000; Moss et al., 1994), in which associative + semantically related pairs tend to elicit stronger priming compared with pairs which are only semantically related. As our model asserts that the same correlation mechanism can be responsible for both types of relatedness (associative and semantic), the very fact that there is an association between words in a pair implies that, first, the words in that pair are typically more strongly correlated than the words in most unassociated (e.g., purely semantically related) pairs to begin with, and second, that there are frequent transitions between the words comprising this pair which further increase their priming effect. Therefore, the ''boost'' stemming from the additional associative connection becomes a straightforward outcome.

A related view, attributing both semantic and associative relations to correlations between distributed representations, is proposed by semantic-space models of priming such as HAL, LSA, or BEAGLE (for a review, see Jones et al., 2006). In these models, an extensive vocabulary of words is represented in vectorial space according to an elaborate co-occurrence measure of these words in large text corpora (the models differ in the specific way in which they instantiate this mapping). Priming is attributed by such models to the differences in the vector correlations of unrelated versus related word pairs. Similarly, differences between associative and semantic priming are attributed to correlation differences between the representations of associatively and semantically related pairs. These models fit well with distributed attractor networks (and ours in particular) because in both types of models,

pattern correlations are the source of priming. Attractor models, however, add dynamics to these static representations and therefore can better illustrate how priming is modulated by SOA and task constraints.

Another important phenomenon that our results address is backward priming. To reiterate, backward priming refers to pairs that are related in only one direction (*Stork–Baby*), but nevertheless yield priming even when presented as prime and target in the reverse direction. Previous explanations of backward priming have usually attributed this effect to controlled processes that operate solely for lexical decisions; this is because, unlike automatic priming, these processes are thought to allow bidirectional evaluation of the stimulus pairs. This account, however, was found to be inconsistent with later demonstrations of backward priming in pronunciation tasks at short SOAs (Kahan et al., 1999; Thompson-Schill et al., 1998). A different attempt to account for backward priming was made within the framework of distributed network models (e.g., Plaut & Booth, 2000). This approach presumes that a backward association between a target and a prime is sufficient for their semantic representations to share some semantic features. In that case, backward priming should not be different from the usual forward priming because the correlation between the representations of the two words is symmetric. This hypothesis was supported by findings showing that at short SOA asymmetrically related pairs elicit similar priming effects whether presented in the forward or backward direction (Thompson-Schill et al., 1998). Nevertheless, although successfully accounting for backward priming effects at short SOAs (regardless of task), the ''shared features'' account cannot easily explain why backward priming does not appear in pronunciation under longer SOA conditions (Kahan et al., 1999; Peterson & Simpson, 1989). Acknowledging these problems, Kahan et al. (1999) suggested that other mechanisms might contribute to backward priming at short SOAs, but no such mechanisms have been specified to date (see also Franklin, Dien, Neely, Waterson, & Huber, 2007).

Our model can easily account for the SOA-dependent backward priming. Consistent with previous models based on distributed representations, backward priming in our model, like forward priming, stems from the existence of correlations between primes and targets (as in the pair *Dog–Beware*). However, due to the directionality of the associations of such pairs (Figs. 8A and 9), the initial correlation decreases with SOA as transitions become more probable, thus diminishing the impact of the prime and eliminating the backward priming effect. Therefore, our model naturally accounts for a nonstrategic, automatic backward priming effect which strictly depends on short SOA conditions (for a related view, see the discussion about backward priming in ACT$^*$ in McNamara, 2005).

To conclude this discussion, it is important to acknowledge that although the present simulation exhibited significant asymmetric properties, we do not suggest that direct one-to-one connections between semantically unrelated concepts are completely absent. It is reasonable to assume that certain concepts do, indeed, tend to activate each other while not having any overlapping representations, and that some of the particular asymmetries in association norms could be attributed to these cases. Nevertheless, we emphasize that at least some of the differences between semantic and associative priming do not require additional elaborations and can be attributed to the correlation structure alone. It is also worth noting that the tendency of the priming effect to increase or decrease, as seen in Fig. 9, depends on the

existence of one single transition; more transitions might, in fact, reduce the effect, as is already seen to some degree at the longest SOA (see two upper curves in Fig. 9). This reduction can be prevented if one assumes conditions in which transitions beyond the first are banned. Indeed, as we show in a more recent study (I. Lerner, S. Bentin, & O. Shriki, unpublished data), our network can consistently produce strong priming effects when various parameters in the network, including parameters that affect the rate of latching dynamics, are strategically modulated. As we show there, these modulations can reflect the involvement of known controlled processes operating in semantic priming such as expectancy and semantic matching.[13]

## 8. General discussion

The main goal of this study was to develop a unifying model of semantic activity in which principles of distributed attractor networks and SA mechanisms are combined to explain semantic priming phenomena. Adding biologically motivated neural adaptation mechanisms to an attractor neural network, we first showed how semantic transitions yield a dynamic development of the mean correlations of the network, implementing SA when averaged over trials. Next, supporting the mechanistic account of our model, we replicated known semantic priming patterns observed in human studies and suggested how transitions in the network can be linked to free associations in humans. Furthermore, based on its characteristics, the model accounted naturally for findings that were insufficiently explained by previous attractor neural network models, such as mediated and asymmetric priming. In addition, the model provided new insights about debated issues in the semantic priming domain such as the mechanisms involved in backward priming as well as the difference between semantic and associative relatedness and its relation to feature overlap in attractor networks.

### *8.1. Implications of the present model*

The dynamics of our model predict several effects that were insufficiently explored by previous human semantic priming studies. Among those, perhaps, the most central prediction, although the most difficult one to test, is that RTs should have a higher variance when transitions have the opportunity to facilitate the response (e.g., in related trials) than when transitions cannot facilitate the response (e.g., in unrelated trials) or do not occur. This pattern should exist because latching dynamics is stochastic (i.e., different semantic transitions may occur in each trial at varying probability) and, therefore, the spectrum of possible states which the semantic network can take due to transitions increases with time. For example, after initially converging on *cat*, the network can later converge on *milk* or *dog*, then on *white*, *cheese*, or *friend* and *leash*, and so on. As the convergence time of the lexical network is influenced by the momentary state of the semantic network, word-recognition times should vary accordingly, depending on the specific mix of shorter and longer RTs over trials. This predicted outcome, which stems from the dynamics of our model, is actually a

version of the well-known increase in variance with time for ''random walk'' processes. Testing it in humans, however, is not trivial; RT variability is determined by many factors, including nonsemantic ones like motor control and attentional lapses. Therefore, subtle variations in RTs at the order of 20–30 ms, like the ones caused by semantic facilitation effects, result in very small increases in the RT variance which are likely to be concealed by much stronger sources of variability. In addition, the degree of variability is also known to be influenced by the absolute lengths of the RTs with longer RTs yielding larger variability (as a result, e.g., of the noise having more opportunity to affect the dynamics with longer times; e.g., Ratcliff, Gómez, & McKoon, 2004). Therefore, to examine this prediction with sufficient reliability, it is required to (a) test a very large number of participants and (b) balance the expected mean RTs of the examined conditions (e.g., related vs. unrelated) such that the danger of confounding differences in variance with differences in means is reduced. A possible way to avoid the need to equalize the absolute RTs across the two conditions is to examine a slightly different form of the prediction, namely that the RT variance should increase with SOA in the related condition but remain roughly the same (or increase less) across SOAs in the unrelated condition. Using this procedure, variances are compared within conditions instead of between conditions and the influence of the mean RTs should play a smaller role. Nevertheless, a large number of participants would still be required.

A second central prediction stemming from our model is concerned with the time course of mediated versus direct priming. As the spreading of activation in our model occurs in jumps rather than continuously, these jumps might induce nonlinear changes in the pattern of semantic priming effects along time, which should be expressed differently in different types of related pairs. For example, the effect of SOA on priming in a pronunciation task should reveal different time courses for indirectly related and directly related pairs. Mediated priming is expected to have a late onset compared with direct priming, as it requires a semantic transition to take place, whereas direct priming does not. Examining direct and mediated priming, while changing the SOA in small steps, is expected to reflect this difference. Whereas the time course of priming for different relatedness conditions have been examined in the past, to our knowledge no direct examination of mediated versus direct priming at very short SOAs has been conducted to date. Examining priming effects over several SOAs when the prime and the target have different degrees of relatedness, both Lorch (1982) and Ratcliff and McKoon (1981) found no difference in the onset of priming as a function of the type of prime-target relation. However, Lorch used strong versus weak categorical relations, as well as high versus weak associative relations, all of which are direct relations and, thus, irrelevant to the comparison between direct and indirect priming. Ratcliff and McKoon's materials were experimentally linked to each other during a learning phase in which the stimuli pairs appeared in paragraphs, separated by a varying degree of interleaved words. Whatever the sources of these priming effects were, the possibility that ''indirectly'' linked words created in this paradigm are in fact directly related was not sufficiently controlled for. Indeed, the lack of interaction and the consequent conclusion that SA is almost ''instantaneous'' is actually in complete agreement with the priming pattern expected to emerge using related pairs with various degrees of direct relatedness (Plaut,

1995). Therefore, a true comparison of the time coarse of direct and mediated priming still awaits well-controlled examinations.

### *8.2. Comparison to other attractor models*

Most of the previously suggested attractor network models of word recognition and semantic priming did not properly account for several behaviorally established effects such as mediated priming and differences between associative and semantic priming (e.g., Masson, 1995; McRae, de Sa, & Seidenberg, 1993; Sharkey & Sharkey, 1992). Two models fare better in that respect. One has been proposed by Plaut (1995; Plaut & Booth, 2000; see also Moss et al., 1994). Plaut's model suggests a basic distinction between semantic and associative relations, which is expressed in the way memory patterns are learned by the network (and, consequently, in the eventual connectivity between units). This model, therefore, provided a reasonable distinction between the priming effects resulting from the two types of relations. It also demonstrated how inhibition-dominance patterns of priming at long SOA could be readily explained by neuronal properties without assuming the involvement of controlled processes (Plaut & Booth, 2000). However, Plaut's model does not easily handle mediated priming without assuming direct links between indirectly related pairs, and it generally predicts a decrease in priming with SOA for purely semantically related primes and targets, contrary to recent findings (Lucas, 2000). Both these restrictions might be accounted for by the limited long-term dynamics that characterizes Plaut's model, as well as most other attractor network models. Nevertheless, the principles of Plaut's model do not contradict the mechanisms displayed in the current study and the two models could actually be combined. For example, Plaut's model describes in some detail the learning mechanisms that allow patterns to become attractors in the network. It may very well be that introducing adaptation mechanisms into Plaut's model would lead his network to present dynamical behavior resembling the dynamics of our current model and, consequently, allow investigating how different learning schemes contribute to the transition probabilities which the network eventually exhibits.

A second attractor neural network model of priming that addressed some of the limitations mentioned above was recently introduced by Lavigne and Darmon (2008) and Brunel and Lavigne (2009). These authors attempted to relate the semantic priming phenomena to simple associative priming in monkeys during a paired-associate task. Developing a realistic integrate-and-fire neural network inspired by real-time recordings from monkeys' cortical neurons, the authors showed how similar mechanisms may operate in both cases. Most relevant to our discussion, this network could actually be conceived as directly implementing SA behavior: Concepts are characterized in their network by specific populations of neurons which, when active, send excitatory inputs to other populations representing related concepts. Although some degree of attractor dynamics is evident in that model (observing retrospective activation of neurons even after an external stimulus is shutdown), the network allows several patterns to be activated simultaneously and thus resembles the parallel activation of nodes in traditional SA models. Specifically, priming stems from the preactivation of the appropriate neural population prior to the appearance of the target. Mediated priming is also apparent much in the same way as in the SA model.

Although a full appreciation of Lavigne and colleagues' cortical network model is beyond the scope of our current study, it is nevertheless important to note a fundamental difference between that model and most other attractor networks (including ours). The cortical network model allows each neuron to participate in the coding of only one memory pattern. Therefore, in contrast to other attractor models of semantic priming, the representations in the cortical network model are actually not distributed and, as a result, allow for several concepts to be concurrently activated to the same degree. In other words, this model is best seen as a working memory system that indistinguishably holds several items together, with no conservation principle to force the activation to dissipate with semantic distance. This dynamics contradicts some basic premises of older semantic memory models, like Anderson's ACT, as well as most versions of the Collins and Loftus' SA model. It also prevents the network from simulating associative thinking, where one association replaces another in a serial manner. Finally, no distinction between semantic and associative connections is evident in the model.

Finally, another model that calls for comparison to our network is Huber and O'Reilly's nROUSE (''neural mechanism for responding optimally with unknown source of evidence''; Huber & O'Reilly, 2003). Like us, these authors used synaptic depression in their model as the basic mechanistic account for several cognitive phenomena. nROUSE assumes a network architecture consisting of visual, orthographic, and lexical/semantic layers with excitatory feed-forward connections between layers and inhibitory connections within each layer. Additional excitatory feedback from the semantic to the orthographic level produces a simple attractor dynamics in which these two layers converge on consistent activity that corresponds to the visual input presented at that time. The gist of the model is that all connections are subject to synaptic depression; therefore, with time, any activity induced in the network by external inputs gradually decreases and renders the activated nodes less sensitive to additional activation by new inputs. This mechanism thus allows producing many cognitive adaptation effects ranging from word repetition priming (Huber & O'Reilly, 2003) to primed face detection (Rieth & Huber, 2005). Basic semantic priming results are also accounted for in the model (Huber & O'Reilly, 2003), showing an advantage for related over unrelated targets which peaks at a certain SOA but decreases with longer prime durations (similar to some of the reported effects for non-associated prime-target pairs discussed in the present paper). Unlike the current model, however, nROUSE was not aimed at accounting for processes which depend on elaborate semantic structure and, therefore, its semantic network has not been developed to simultaneously contain many concept patterns producing rich attractor dynamics. Nevertheless, as the basic mechanism of both models relies on synaptic depression, they are essentially completing each other and can likely be combined to produce a comprehensive architecture that could account for many cognitive effects in different domains.

### *8.3. Generalizations*

We move now to consider some of the more general properties of the current model. First, we should point out that although we chose to implement our model with a Hopfield-like network and a latching mechanism based on synaptic depression, the results do not

depend on these choices. Hopfield networks may be seen as a prototype of associative networks with distributed representations and gain their power from their simplicity and their straightforward Hebbian-derived connectivity. In principle, however, other types of attractor networks could also be used as long as concepts in these networks are represented in a distributed manner and relatedness is translated to correlations between patterns of activity. Similarly, synaptic depression is only one way of implementing adaptation mechanisms leading to latching dynamics and was chosen for its biological plausibility and experimental support (Tsodyks & Markram, 1997). Yet other biological mechanisms can lead to the same effect (e.g., Herrmann et al., 1993) and could be implemented if supported by evidence.

Second, the vector representations stored in our network were handcrafted to demonstrate the network's dynamics; therefore, they represented concepts, as well as relatedness between concepts, only at the most abstract level. Surely, real representations in the brain have a far more elaborate structure. However, it is precisely the general character of our vectors which gives the model its strength and allows it to relate to past conceptualizations of distributed representations (e.g., feature-based representations and representations based on lexical co-occurrence measures). The emphasis in this study is on the way distributed representations; however, they are defined, develop with time due to adaptation mechanisms of the active units—not on the representations themselves. Nevertheless, some aspects in the model might go along more comfortably with representations that do not require correlations to be dependent on common features (at least in their simple sense of distinguishable traits such as *has four legs* being a feature of *dog*); this is because correlations between concept patterns, as assumed in Simulation 3, can reflect relations that are not necessarily based on common features (e.g., *dog* and *beware*). In that respect, the model's viewpoint is more consistent with vector-space representations such as BEAGLE, which assume that correlations can represent any kind of relations, including different types of semantic and associative relatedness (e.g., Jones et al., 2006).

Third, from a purely mathematical point of view, our model suggests that transitions in the content of associative thought may be approximated by a Markov chain process (see, e.g., Meyn & Tweedie, 1993) where each component in the state vector of the Markov model corresponds to the probability of a certain concept being activated. The whole state vector would therefore correspond to the probability of the semantic network being converged on any of its stored concepts at a given time, and the initial state (where only one particular concept is activated due to an external stimulus) can be viewed as a binary state vector with ''1'' in the appropriate component and ''0'' otherwise. The transition matrix of such a Markov chain model should reflect the basic transition probabilities between any two concepts and could derive its exact values from association norms (for a related idea concerning the origin of language, see Kropff & Treves, 2007; also, for a demonstration of the relation between hidden Markov models and synaptic depression in a different context, see Huber, 2008). Such process can then be combined with a measure of similarity, serving as a proxy for the degree of the correlation between concepts, to provide a crude approximation of the progress of semantic activation over time. For example, assume that a certain concept, designated as $x_k$, is activated at time 0 and our goal is to estimate its mean activation after $n$ time steps (over all realizations of the possible transitions occurring during this time). Using a transition matrix $P$ (driven by association norms) and a similarity measure between two

concepts $S(x, y)$ (taken, e.g., from vector-space representations such as LSA or BEAGLE), the state vector representing the distribution of the network states over the stored concepts after these $n$ time steps is given by $P^n\vec{x_0}$ with $\vec{x_0}$ representing the binary vector which corresponds to the initial state where the network is converged on concept $x_k$. The mean activation of this concept after $n$ steps, based on the similarity measure and the distribution, will therefore be given by $\sum_i S(x_k, x_i) \cdot [P^n\vec{x}_0]_i$, with $i$ going over all stored concepts. Such computation is not unique to the initially activated concept and can be carried out for any other stored concept. Thus, a simple formalization of this sort allows making crude predictions on the degree of activation of concepts, starting from an agreed initial state, which can be compared to findings in paradigms such as semantic priming. It may also serve as a simple way for comparing the basic principles of our model and those of other models, such as BEAGLE. However, it is important to emphasize that this abstract form of the model should be regarded with caution. First, association norms lack significant information regarding the matrix $P$ as they do not contain the probability of the network remaining converged on the current state (i.e., the diagonal of $P$). Second, and even more important, the full network performance is not completely analogous to a stationary Markov chain, as the probabilities of transitions are not stable over time. It is necessary, therefore, to carefully choose the correct time window as the equivalent of one step in the Markov chain to make the analogy as precise as possible. A more accurate analogy of the associative transitions would require using a high-order Markov chain (see, e.g., Russo, Pirmoradian, & Treves, 2010). In other words, the transition probabilities may depend not only on the last attractor but also on preceding attractor states. In this case, the state vector should be extended to a spatiotemporal pattern that takes into account several time steps.

The acceleration of RTs when the activity in the semantic network matches the bottom-up input to the lexical network reflects the integration of prior knowledge about semantic relations (as revealed by the transition probabilities) with instantaneous information. Such an integrative process may form, for example, the foundation of top-down facilitation in reading, in which a word in a sentence is processed faster and more accurately given the previous context. Statistically speaking, the probability for the lexical network to converge to a certain word pattern is proportional to both the conditional probability of this word given the external input, and the conditional probability of this word given the previous word. In this sense, the model implements a basic form of Bayesian inference in a neuronal architecture (cf., Knill & Pouget, 2004).

Finally, it is important to emphasize that we made no assumptions about conscious awareness while proposing this network. Specifically, associative transitions in the semantic network are not necessarily coupled with a transition in the content of conscious thought. To date, consciousness is still an ill-defined term and does not need to be a binary trait (e.g., Mandler, 2005). One could argue that ''full'' conscious awareness of a concept may depend on its mutual activation by different subnetworks in the brain, and particularly linguistic modules like the mental lexicon (see, e.g., Gazzaniga, 2000). In that case, in terms of our model, it is not enough for a certain concept to be activated by the semantic network to reach full consciousness; rather, it might also require the activation of its lexical representation in the lexical network, as well as other neural mechanisms.

## 9. Conclusion

We presented a distributed attractor neural network model that simulates semantic memory and accounts for semantic priming effects. Assuming synaptic adaptation mechanisms, we have shown that the semantic network may engage in associative transitions from one pattern to another (latching dynamics), which may be analogous to the SA mechanism widely accepted as an account for automatic priming. Our model can also be captured as forming a foundation for associative thinking in general. For example, in a recent study, we showed how the current model, operating under accelerated rate of semantic transitions, can also account for the typical patterns of semantic priming in schizophrenia as well as for several thought disorders characterizing the disease (Lerner, Bentin, & Shriki, 2012). While it is clear that various results involving complex interactions within semantic memory need further elaborations, we believe the model presented here can constitute firm grounds for the investigation of semantics in general and modeling semantic activity in the attractor neural network framework in particular.

## Acknowledgments

This manuscript is dedicated to the memory of our dear friend, colleague, and collaborator, Shlomo Bentin, Ph.D. His scientific rigor, open-mindedness, and endless devotion will forever serve as our inspiration.

Oren Shriki was supported in part by the Intramural Research Program of the National Institute of Mental Health.

## Notes

1. Although the above analysis of priming in attractor networks describes how correlated memory patterns facilitate one another, the term ‘‘facilitation’’ does not coincide with its usual meaning in semantic priming literature, where facilitation and inhibition are defined in comparison with a neutral condition in which the target is preceded by either a word that does not bias the recognition of the target (such as BLANK), a pseudoword (a phonologically legal but meaningless sequence of letters), or by a nonword (a phonologically illegal string of letters). It is, therefore, crucial to define how the neutral condition is constructed in attractor network simulations of semantic priming to define whether the effect is purely facilitatory or also includes an inhibitory component. In neutral trials, the network is presumed to remain on some baseline state until target onset. This baseline could be conceived in two ways: One is to assume that the network activity in the neutral condition is distributed randomly rather than being converged on any attractor. If this is so, then some neuronal activation in this baseline state might be somewhat correlated to the upcoming target simply by chance, and therefore, the transition from baseline to the target would be faster than the transition

from an unrelated activity pattern, thus resulting in inhibition. However, as a random baseline would not constitute an attractor of the dynamics, there will be no driving force encouraging it to maintain its activity. This may allow changes to occur more rapidly even compared to the related case where the network needs to escape the basin of attraction of the prime's memory pattern. Hence, such a neutral baseline may lead to the unwarranted result of convergence being fastest in the neutral condition (see Dalrymple-Alford & Marmurek, 1999). A better way to conceive the baseline state is to assume that in the neutral condition, the starting position of the network is not random but rather forms an attractor of the dynamics whose activity pattern is not correlated to the memory patterns representing real words. In this case, the transition of the network state from such an attractor to the activity pattern which represents the target will not be different from the transition when starting from a ''real'' unrelated memory pattern and so no inhibition will should evident.

2. Moreover, we do not assume a strict serial chain of events. In fact, all or some of these levels may be activated in cascade.
3. To this basic structure of the word-recognition system, two other modules could be added. One is a decision module receiving input solely from the lexical network and responsible for the distinction between yes/no binary choices (in a lexical decision task). The other module is responsible for the generation of phonetic codes required for pronunciation and receives information from both the lexical and the orthographic networks. These two modules might be necessary to make accurate predictions regarding RT distributions in response to words and nonwords, which depend on task requirements and response strategies (e.g., Balota, Yap, Cortese, & Watson, 2008). However, as the current model focuses on semantic processing rather than the production of responses, these modules will not be addressed further.
4. The Appendix presents the values of all the parameters used in this study that, whenever relevant, were chosen from a biologically plausible range.
5. Indeed, this asymmetry is most probably task dependent and, if needed, might be reversed (e.g., when a person needs to transform thoughts into words during conversation). The way subjects control such connectivity scaling is not fully understood, but it has been suggested to be the product of attention (e.g., Büchel & Friston, 1997) and carried out by neuromodulators such as Dopamine (Cohen & Servan-Schreiber, 1992; Coull, 1998).
6. It might be noteworthy that the sparseness value used in our simulations (0.06 in the semantic network, see Appendix) creates a ratio between uncorrelated memories and units that is well within the range that allows scaling the network to the number of concepts in the entire human semantic memory. For example, in order to hold 30,000 differentiable concepts (a quantity based on estimations made by Biederman, 1987), our network would require around 500,000 units. Holding a million different concepts requires fewer than 20,000,000 units. These approximated values are well below the estimated number of neurons in the medial-temporal lobe where semantic memory is thought to reside (e.g., Harding, Halliday, & Kril, 1998; Quian Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). Correlations between memories lower these estimates even further.

7. As all patterns were binary vectors with sparse representations, the overlapping active units were actually the major contributors to the correlation. Unrelated concepts were created as patterns with no overlapping active units. Given the finite size of our network, this is the optimal approximation of noncorrelation. As it turns out, such an approximation actually led to a small negative correlation between unrelated patterns. This small bias from 0 had no significant effects on either the behavior of the network or the results.
8. It is well acknowledged that the choice of baseline in neutral trials of priming experiments (e.g., rows of X's; the words ''BLANK'' or ''READY''; or pseudowords driven from unrelated concepts; see Neely, 1991; Plaut & Booth, 2000) can modulate their resulting RTs and should therefore be chosen carefully. Ideally, a neutral trial should have the same alerting properties as related and unrelated primes, but leave the lexical system unbiased with respect to the target. However, it is not self-evident how these requirements are fulfilled in designing neutral trials of attractor networks' simulations (see also note 1). On the one hand, it is clear that in neutral trials the lexical network should not be encouraged to change its activity toward any specific memory pattern. On the other hand, our choice of baseline, which was encoded simply as another attractor of the dynamics, was essentially not different than any unrelated pattern. An alternative could have been to initiate the lexical network to a random activity state with the same mean activity level as any of the word patterns. One of the problems with the latter was discussed in note 1. Another problem, conspicuous when the SOA is not very short, is that as this random pattern would not form an attractor, as soon as the external input diminish, the network would converge to one of the encoded patterns, hence not allowing such a trial to be really neutral. Another possibility could be to initialize the network to a state which partially resembles one of the unrelated concepts (with respect to the target), though not completely equivalent to it. This might be seen as the correlate of a ''pseudoword'' prime which has been recommended for human experiments (e.g., Plaut & Booth, 2000). In that case, the network would most likely converge to this unrelated concept as soon as the external effect diminishes (as it bares strong correlation with the initial state) and effectively yield an unrelated trial. The results, therefore, would be roughly the same as with our choice of baseline. Theoretically, however, they might be less justified as pseudowords are not meant to fully elicit the unrelated words from which they were derived.
9. The absolute RTs in our simulations (in contrast to our relative measures such as the priming effect) are obviously much shorter than experimental RTs of real subjects as they do not include time-consuming processes such as stimulus perception, decision making, and motor response.
10. The small, unexpected difference between the average RTs of the related and unrelated conditions might be explained, admittedly, post hoc. In neutral conditions, when the prime is not different from baseline, the semantic network does not make any transitions until the onset of the target. In contrast, when any non-neutral prime is presented, the network responds and converges to the corresponding pattern. Such a transition works as a ''resetting'' device for the accumulating noise in the silent

units: The newly active units in the new pattern reached by the network depress the noise in these silent units. In neutral trials, this depression is weak due to the synaptic adaptation mechanisms. Noise in these neutral trials can raise the activity levels of those units a little above 0 and, therefore, accelerate the convergence of the lexical network as soon as the target, any target, is presented. Such acceleration is almost nonexistent in non-neutral trials where ''resetting'' occurs. This phenomenon may create a kind of inhibition for unrelated targets, which was reflected in the slightly (2 simulated milliseconds) longer RTs in the unrelated compared with the neutral trials. Note, however, that this effect is more than an order of a magnitude smaller than the major relatedness effect and is not expected to be salient in environments with large amount of variability (such as real RT measurements in priming experiments with humans). Therefore, we do not consider this effect as violating our premise that the basic priming effects in our model are purely facilitatory.

11. In fact, the influence of the correlation structure is even more complicated. The probability of a transition from a cue to a target is not only influenced by their correlation strength and the number of patterns which are correlated to the cue but also depends on the correlation density of the target such that targets connected to many other concepts tend to resist the network from jumping to them.
12. Associations that occur very rarely (e.g., *kitten–dog*) provide very few data points for the calculation of their average latency. Therefore, all such associations were combined together in one group. The criterion for this inclusion was association strength of less than 0.02. The association strength of this group was determined as the weighted average of the association strengths of each of the particular associations. Forty-four data points from 19 different associations were combined in this way. Similarly, the data points of associations to concept patterns which are symmetric with respect to the stimulus (e.g., *friend* and *house*, with respect to *dog*) were also combined, as any difference between them must stem from noise. Qualitatively similar results are also achieved without these procedures, though with reduced statistical significance.
13. A complete cessation of latching dynamics might also be necessary when the network learns to acquire the representation of new concepts. As learning is usually a gradual process that requires the repeated presentation of relevant inputs, it is reasonable to assume that network transitions should be avoided to allow the formation of a reliable representation of those inputs without the risk of them being associated to a concept which the network has randomly jumped to. The control over transitions speculated in our additional work (I. Lerner, S. Bentin, & O. Shriki, unpublished data) may, therefore, serve such purpose as well.

## References

Amster, H. (1964). Semantic satiation and generation: Learning? Adaptation? *Psychological Bulletin*, *62*, 273–286.

Anderson, J. R. (1974). Retrieval of prepositional information from long-term memory. *Cognitive Psychology*, *6*, 451–474.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*, 261–295.

Balota, D. A., & Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 336–345.

Balota, D. A., Yap, M. J., Cortese, M. J., & Watson, J. M. (2008). Beyond mean response latency: Response time distributional analyses of semantic priming. *Journal of Memory and Language*, *59*, 495–523.

Bibitchkov, D., Herrmann, J. M., & Geisel, T. (2002). Pattern storage and processing in attractor networks with short-time synaptic dynamics. *Network: Computation in Neural Systems*, *13*, 115–129.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*, 115–147.

Borowsky, R., & Besner, D. (1993). Visual word recognition: A multistage activation model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 813–840.

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review*, *116*, 220–251.

Brunel, N., & Lavigne, F. (2009). Semantic priming in a cortical network model. *Journal of Cognitive Neuroscience*, *21*, 2300–2319.

Büchel, C., & Friston, K. J. (1997). Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex*, *7*, 768–778.

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

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

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240–247.

Coull, J. T. (1998). Neural correlates of attention and arousal: Insights from electrophysiological, functional neuroimaging and psychopharmacology. *Progress in Neurobiology*, *55*, 343–361.

Dalrymple-Alford, E. E., & Marmurek, H. H. C. (1999). Semantic priming in fully recurrent network models of lexical knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 758–775.

Deese, J. (1962). On the structure of associative meaning. *Psychological Review*, *69*, 161–175.

Flekkoy, K. (1973). Associative frequency and response latency. *Scandinavian Journal of Psychology*, *24*, 199–202.

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: Bradford/MIT press.

Franklin, M. S., Dien, J., Neely, J. H., Waterson, L. D., & Huber, L. (2007). Semantic priming modulates the N400, N300, and N400RP. *Clinical Neurophysiology*, *118*, 1053–1068.

Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition?. *Brain*, *123*, 1293–1326.

Goldstein, M. J. (1961). The relationship between anxiety and oral word association performance. *The Journal of Abnormal and Social Psychology*, *62*, 468–471.

de Groot, A. M. B. (1984). Primed lexical decision: Combined effects of the proportion of related prime-target pairs and the stimulus-onset asynchrony of prime and target. *Quarterly Journal of Experimental Psychology*, *36A*, 253–280.

de Groot, A. M. B. (1985). Word-context effects in word naming and lexical decision. *The Quarterly Journal of Experimental Psychology*, *37A*, 281–297.

de Groot, A. M. B., Thomassen, A. J., & Hudson, P. T. (1986). Primed lexical decision: The effect of varying the stimulus-onset asynchrony of prime and target. *Acta Psychologica*, *61*, 17–36.

Harding, A., Halliday, G., & Kril, J. (1998). Variation in hippocampal neuron number with age and brain volume. *Cerebral Cortex*, *8*, 710–718.

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

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science, USA*, *79*, 2554–2558.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Science, USA*, *81*, 3088–3092.

Horn, D., & Usher, M. (1989). Neural networks with dynamical thresholds. *Physical Review A*, *40*, 1036–1040.

Huber, D. E.. (2008). Causality in time: Explaining away the future and the past. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for rational models of cognition* (pp. 351–376). Oxford, England: Oxford University Press.

Huber, D. E., & O'Reilly, R. C. (2003). Persistence and accommodation in short-term priming and other perceptual paradigms: Temporal segregation through synaptic depression. *Cognitive Science*, *27*, 403–430.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? *Psychonomic Bulletin & Review*, *10*, 785–813.

Jones, L. L. (2012). Prospective and retrospective processing in associative mediated priming. *Journal of Memory and Language*, *66*, 52–67.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*, 534–552.

Kahan, T. A., Neely, J. H., & Forsythe, W. J. (1999). Dissociated backward priming effects in lexical decision and pronunciation tasks. *Psychonomic Bulletin & Review*, *6*, 105–110.

Kawamoto, A. H., & Anderson, J. A. (1985). A neural network model of multistable perception. *Acta Psychologica*, *59*, 35–65.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*, 712–719.

Kropff, E., & Treves, A. (2007). The complexity of latching transitions in large scale cortical networks. *Natural Computing*, *6*, 169–185.

Lambert, W. E., & Jakobovits, L. A. (1960). Verbal satiation and changes in the intensity of meaning. *Journal of Experimental Psychology*, *60*, 376–383.

Lavigne, F., & Darmon, N. (2008). Dopaminergic neuromodulation of semantic priming in a cortical network model. *Neuropsychologia*, *46*, 3074–3087.

Lerner, I., Bentin, S., & Shriki, O. (2012). Excessive attractor instability accounts for semantic priming in schizophrenia. *PLoS ONE 7*(7): e40663. doi:10.1371/journal.pone.0040663

Lorch, R. F. (1982). Priming and search processes in semantic memory: A test of three models of spreading activation. *Journal of Verbal Learning and Verbal Behavior*, *21*, 468–492.

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review*, *7*, 618–630.

Lupker, S. J. (1984). Semantic priming without association: A second look. *Journal of Verbal Learning and Verbal Behaviour*, *23*, 709–733.

Mandler, G. (2005). The consciousness continuum: From 'qualia' to 'free will'. *Psychological Research*, *69*, 330–337.

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

Mato, G. (1999). Stochastic resonance using noise generated by a neural network. *Physical Review E*, *59*, 3339–3343.

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

McKoon, G., & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1155–1172.

McNamara, T. P.. (2005). *Semantic priming: Perspectives from memory and word recognition*. New York: Psychology Press.

McNamara, T. P., & Holbrook, J. B.. (2003). Semantic memory and priming. In A. F. Healy & R. W. Proctor (Eds.), *Handbook of psychology: Experimental psychology* (pp. 447–474). New York: Wiley.

McRae, K., de Sa, V., & Seidenberg, M. S.. (1993). Modeling property interactions in accessing conceptual memory. In J. W. Kintsch (Ed.), *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 729–734). Hillsdale, NJ: Erlbaum.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234.

Meyn, S. P., & Tweedie, R. L.. (1993). *Markov chains and stochastic stability*. New York: Springer-Verlag.

Moss, H. E., Hare, M. L., Day, P., & Tyler, L. K. (1994). A distributed memory model of the associative boost in semantic priming. *Connection Science*, *6*, 413–427.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*, 22–254.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading* (pp. 264–336). Hillsdale, NJ: Erlbaum.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A.. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/. Accessed October 14, 2010.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*, 402–407.

Peterson, R. R., & Simpson, G. B. (1989). Effect of backward priming on word recognition in single-word and sentence contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1020–1032.

Pfister, J. P., Dayan, P., & Lengyel, M. (2010). Synapses with short-term plasticity are optimal estimators of presynaptic membrane potentials. *Nature Neuroscience*, *13*, 1271–1275.

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. In J. F. Lehman & J. D. Moore (Eds.), *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 37–42). Hillsdale, NJ: Erlbaum.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of naming and lexical decision. *Language and Cognitive Processes*, *12*, 767–808.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, *107*, 786–823.

Plaut, D. C., & McClelland, J. L. (2010). Locating object knowledge in the brain: A critique of Bowers' (2009) attempt to revive the grandmother cell hypothesis. *Psychological Review*, *117*, 284–288.

Quian Quiroga, R., & Kreiman, G. (2010). Measuring sparseness in the brain: Comment on Bowers (2009). *Psychological Review*, *117*, 291–299.

Quian Quiroga, R., Reddy, R., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*, 1102–1107.

Ratcliff, R., Gómez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*, 142–165.

Ratcliff, R., & McKoon, G. (1981). Does activation really spread?. *Psychological Review*, *88*, 454–462.

Rieth, C. A., & Huber, D. E.. (2005). Using a neural network model with synaptic depression to assess the dynamics of feature-based versus configural processing in face identification. In B. G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 1856–1861). Hillsdale, NJ: Erlbaum Associates.

Rolls, E. T., Loh, M., Deco, G., & Winterer, G. (2008). Computational models of schizophrenia and dopamine modulation in the prefrontal cortex. *Nature Reviews Neuroscience*, *9*, 696–709.

Russo, E., Namboodiri, V. M., Treves, A., & Kropff, E. (2008). Free association transitions in models of cortical latching dynamics. *New Journal of Physics*, *10*, 015008. doi: 10.1088/1367-2630/10/1/015008.

Russo, E., Pirmoradian, S., & Treves, A.. (2010). Associative Latching Dynamics vs. Syntax. In R. Wang & F. Gu (Eds.), *Advances in Cognitive Neurodynamics (2): Proceedings of the 2nd International Conference on Cognitive Neurodynamics* (pp. 111–115). New York: Springer.

Schlosberg, H., & Heineman, C. (1950). The relationship between two measures of response strength. *Journal of Experimental Psychology*, *40*, 235–247.
Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.
Sharkey, A. J., & Sharkey, N. E. (1992). Weak contextual constraints in text and word priming. *Journal of Memory and Language*, *31*, 543–572.
Silberman, Y., Miikkulainen, R., & Bentin, S. (2005). Associating unseen events: Semantically mediated formation of episodic associations. *Psychological Science*, *16*, 161–166.
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.
Spitzer, M. (1997). A cognitive neuroscience view of schizophrenic thought disorder. *Schizophrenia Bulletin*, *23*, 29–50.
Thompson-Schill, S. L., Kurtz, K. J., & Gabrieli, J. D. E. (1998). Effects of semantic and associative relatedness on automatic priming. *Journal of Memory and Language*, *38*, 440–458.
Tian, X., & Huber, D. E. (2010). Testing an associative account of semantic satiation. *Cognitive Psychology*, *60*, 267–290.
Treves, A. (2005). Frontal latching networks: A possible neural basis for infinite recursion. *Cognitive Neuropsychology22*, 276–291.
Tsodyks, M. V. (1990). Hierarchical associative memory in neural networks with low activity level. *Modern Physics Letters B*, *4*, 259–265.
Tsodyks, M. V., & Markram, H. (1997). The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proceedings of the National Academy of Science, USA*, *94*, 719–723.
Tsodyks, M., Pawelzik, K., & Markram, H. (1998). Neural networks with dynamic synapses. *Neural Computation*, *10*, 821–835.
Warren, R. M. (1968). Verbal transformation effect and auditory perceptual mechanisms. *Psychological Bulletin*, *70*, 261–270.
Waydo, S., Kraskov, A., Quian Quiroga, R., Fried, I., & Koch, C. (2006). Sparse representation in the human medial temporal lobe. *Journal of Neuroscience*, *26*, 10232–10234.
Zador, A. (1998). Impact of synaptic unreliability on the information transmitted by spiking neurons. *Journal of Neurophysiology*, *79*, 1219–1229.

## Appendix

A. The semantic and lexical network parameters that were used in the numerical simulations of the model (units are indicated in brackets whenever relevant):

| Parameter | Semantic Network | Lexical Network |
|---|---|---|
| Number of units, $N$ | 500 | 500 |
| Sparseness, $p$ | 0.06 | 0.04 |
| Correlation strength (% of overlapping active units out of total active units in a pattern) | 0.1 (Strong)<br>0.066 (Moderate)<br>0.033 (Weak) | 0 |
| Unit-gain, $T$ | 0.05 | 0.05 |
| Unit's time constant, $\tau_n$ | 7 [ms] | 13 [ms] |
| Unit activation threshold, $\theta$ | 0.02 | 0.17 |
| Regulation parameter, $\lambda$ | 14.75 | 27.75 |
| Maximal firing rate, $x_{\max}$ | 100 [spks/sec] | 100 [spks/sec] |

## Appendix

A. Continued:

| Parameter | Semantic Network | Lexical Network |
|---|---|---|
| Utilization of synapses within each network, $U_{[\text{within}]}$ | 0.206 [1/spks] | 0 [1/spks] |
| Utilization of synapses between networks, $U_{[\text{between}]}$ | Lexical-to-semantic: 0.087 [1/spks] | Semantic-to-lexical: 0 [1/spks] |
| Synaptic recovery time within each network, $\tau_{r\ [\text{within}]}$ | 93 [ms] | – |
| Synaptic recovery time between networks, $\tau_{r\ [\text{between}]}$ | Lexical-tosemantic: 1,333 [ms] | Semantic-to-lexical: – |
| Input gain between networks (Raw values. Actual values were normalized by the number of pre-synaptic active units in a pattern) | Lexical-to-semantic: 2 | Semantic-to-lexical: 0.21 |
| External input gain | 0.56 | – |
| Input threshold, $\theta_{\text{ext}}$ | 1 | 0.25 |
| Noise amplitude, $\eta_{\text{amp}}$ | 0.05 | 0.025 |
| Noise temporal correlations, $\tau_{\text{corr}}$ | 17 [ms] | 17 [ms] |
| Convergence threshold | 0.95 | 0.95 |

B. The temporal correlations in the noise were generated by filtering the noise using a low-pass filter, which, for two time points separated by $\tau$ ms, took the form: $f(\tau) = \eta_{amp} \cdot e^{-\frac{\tau}{\tau_{corr}}}$