COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 38 (2014) 1562–1603

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12133

Integrating the Automatic and the Controlled: Strategies in Semantic Priming in an Attractor Network With Latching Dynamics

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 October 2012; received in revised form 6 September 2013; accepted 22 October 2013

Abstract

Semantic priming has long been recognized to reflect, along with automatic semantic mechanisms, the contribution of controlled strategies. However, previous theories of controlled priming were mostly qualitative, lacking common grounds with modern mathematical models of automatic priming based on neural networks. Recently, we introduced a novel attractor network model of automatic semantic priming with latching dynamics. Here, we extend this work to show how the same model can also account for important findings regarding controlled processes. Assuming the rate of semantic transitions in the network can be adapted using simple reinforcement learning, we show how basic findings attributed to controlled processes in priming can be achieved, including their dependency on stimulus onset asynchrony and relatedness proportion and their unique effect on associative, category-exemplar, mediated and backward prime-target relations. We discuss how our mechanism relates to the classic expectancy theory and how it can be further extended in future developments of the model.

Keywords: Word recognition; Semantic priming; Neural networks; Distributed representations; Latching dynamics; Controlled processes; Expectancy; Semantic matching

1. Introduction

Among the most familiar distinctions in cognitive science is the one made between automatic and controlled processes. Findings from various experimental paradigms suggest that some cognitive processes exhibit little to no sensitivity to experimental

Correspondence should be sent to Itamar Lerner, Center for Molecular and Behavioral Neuroscience, Rutgers University-Newark, Newark, NJ 07102. E-mail: itamar.lerner@gmail.com
[†]Deceased, July 2012.

manipulations of task conditions such as the time allowed to process stimuli, the salience of target stimuli compared to distractors, or the statistical contingencies occurring over the course of an experimental session; other processes, in contrast, exhibit great dependency on these manipulations (Posner & Snyder, 1975; Shiffrin & Schneider, 1977). Such findings gave rise to the notion that certain cognitive processes occur automatically, without the voluntary involvement—or even awareness—of the individual and are thus indifferent to various experimental manipulations, whereas other processes are intentionally produced when encouraged by the appropriate experimental conditions and typically reflect a strategic attempt to maximize performance in a given situation (see Neumann, 1984, for a review). According to this view, automatic processes are fast acting, do not require attention, and always occur as long as the appropriate stimuli appear (e.g., a written word will automatically elicit reading in a literate adult). Controlled processes, in contrast, are slow to act, require attention, and are used only when an individual observes that certain procedures may aid him/her to improve performance in a given task (e.g., focusing attention on whether a presented word contained the letter "T" when the task demands it). This view of automatic and controlled processing resonates the more intuitive distinction, rooted in phenomenology, in which some cognitive operations are directly influenced by one's own free will whereas others (the most obvious example being, perhaps, reflexive motor responses) occur "by themselves" without any conscious or willful intervention on one's part.

One of the domains in which the automatic/controlled dichotomy was intensively investigated is semantic processing. Specifically, semantic priming—the facilitation in word recognition occurring when a word is presented following the presentation of a related semantic concept (e.g., *Doctor–Nurse*)—has been shown to reflect automatic contributions of the dynamics and structure of semantic memory, as well as strategic planning executed by subjects attempting to optimize the recognition process. These two types of contributions were differently treated by theoretical accounts. Although automatic semantic priming has been subject to quantitative modeling based on neural networks (e.g., Collins & Loftus, 1975; Plaut, 1995), theories of controlled priming were mostly descriptive, basing their explanations on qualitative terms not embedded in a precise mathematical framework (e.g., Becker, 1980; Neely & Keefe, 1989). As automatic and controlled processes have often been treated as orthogonal in the semantic priming literature (as evident, for example, in the influential hybrid three-process model by Neely & Keefe, 1989), detecting common mathematical principles for these mechanisms was not considered a priority. Nevertheless, it was later shown that, in fact, some controlled processes do interact with automatic ones during semantic priming experiments (e.g., Balota, Black, & Cheney, 1992; Neely, O'Connor, & Calabrese, 2010). Moreover, the dichotomy between automatic and controlled processes was put into question by other studies, casting doubt on whether pure automatic processes truly exist (e.g., Besner, 2001; Stolz & Besner, 1999). These studies, as well as the lack of a principled way to define what makes a process automatic from a neuronal point of view, have left this absence of common grounds for the two types of processes undesirable. Models of controlled processes in semantic priming, especially, now seem to be lagging behind the more biologically oriented accounts for automatic mechanisms.

Lately, we have published a novel neural-network account of semantic priming which attempted to address many findings related to automatic processes thought to be involved in this effect in both healthy subjects and schizophrenic individuals (Lerner, Bentin, & Shriki, 2012a,b). Here, we show how an extension of this model can naturally address several important findings typically attributed to controlled processes utilized during performance in this task. Specifically, we show how basic findings regarding the type of relation between the prime and target words (associative, category-exemplar, mediated, backward), the stimuli list (high vs. low relatedness proportion), and the type of task (pronunciation vs. lexical decision) can be naturally accommodated within our model when combined with a simple reinforcement learning rule which attempts to optimize network performance. From these simulations, we derive a new hypothesis about the underlying differences between automatic and controlled mechanisms in general, and in semantic processing in particular. Although our approach does not directly address the more phenomenological aspects of automatic and controlled processes (e.g., the relation to free will), we do attempt to draw a more distinct, qualitative line, based on computational principles, between those processes which can be intentionally modulated (i.e.,, processes that subjects can intentionally choose to activate if encouraged by environmental or internal motivators) and those which are inherently automatic.

The article has the following structure: First, we review some of the main findings regarding controlled processes in semantic priming. Second, we briefly describe how these findings were modeled by two previous qualitative theories. Then, we turn to describe our model and how it accounts for these findings as well as solves some problematic issues which have not been well treated in the past. Finally, we describe how the model relates to previous theories of automatic and controlled priming, present several predictions stemming from it, and mention how it could be further extended. Not all central findings regarding controlled processes are covered; the literature in this domain is vast and also relates to mechanisms—especially decision-making mechanisms—which are not part of our model. However, we partially address these additional findings in the General Discussion and suggest how they, too, may be accommodated in future expansions of the model using the same principles of reinforcement learning and optimization.

2. Semantic priming

2.1. Basic findings supporting involvement of controlled processes

In a typical priming experiment (Meyer & Schvaneveldt, 1971; Neely, 1977; see Neely, 1991, for a review), subjects are presented with two words in succession, the prime and the target. Frequently used procedures involve reading the prime silently and either naming the target (pronunciation task), or deciding whether it is a real word or not (lexical decision task [LDT]). The target could either be semantically (or associatively) related or unrelated to the prime, or a nonword in case of the LDT. The semantic priming effect refers to the finding that the average reaction time (pronouncing the second word

or deciding it is a real word) is shorter and error rates are lower when the prime and the target are semantically related to each other, compared to when they are unrelated. When examined compared to a neutral stimulus (e.g., a nonverbal prime such as "XXXXX"), this effect is revealed to be composed of facilitation (a related word is recognized faster than a word following a neutral stimulus) and, sometimes, also inhibition (an unrelated word is recognized more slowly than a neutral stimulus).

Semantic priming is believed to be based, first and foremost, on automatic mechanisms that rapidly activate the semantic neighborhood of identified words. Contemporary accounts assert that after a prime word is recognized, some of its related semantic concepts are immediately activated in memory before target appearance, thus providing an automatic "head start" to the subsequent recognition of targets which are related to the prime compared to targets which are unrelated to it (see, for example, the spreading activation theory; Collins & Loftus, 1975). However, there are strong indications that controlled strategies also play a part. For example, when the experimental prime-target pairs are embedded in a stimuli list that contains many semantically related items, priming is augmented compared to when related items are sparse (e.g., Neely, Keefe, & Ross, 1989). This Relatedness Proportion effect (RP; the ratio of related prime-target pairs out of all word pairs) is often salient only when the SOA between the prime and target is long (>500 ms; Neely, 1991; but see Bodner & Masson, 2003; and Feldman & Basnight-Brown, 2008; for some evidence of short-SOA stimuli-list effects). Specifically, the pattern of facilitation versus inhibition in priming is directly affected by RP and SOA: Although facilitation already occurs at short SOAs and increases at long SOAs when the RP is high, inhibition seems to be nonexistent at short SOAs and appears only at long SOAs. Moreover, the increase in facilitation is mostly evident for pairs which are strongly associated with each other such that the target is relatively predictable from the prime (e.g., *dog–cat*). These pairs usually do not produce inhibition at all. In pairs which are semantically related but not necessarily associated (e.g., category – exemplar pairs such as *animal – cat*), facilitation does not increase with SOA, whereas inhibition appears at long SOAs (Neely, 1991; McNamara, 2005; see Table 1 for a summary of some of these effects). Such results convinced researchers that subjects attempt to fit their expectations to the commonalities of the task to reduce their response times and elevate their accuracy, but they are able to do so only when they readily detect that many primes are related to the target (i.e., a high RP) and only when there is enough time between prime and target appearances to develop such expectations (i.e., long SOA; Becker, 1980). If expectations are met, facilitation increases; when they are not, inhibition occurs. Two of the most popular theories addressing these findings are the expectancy mechanism (e.g., Becker, 1980) and semantic matching (Neely & Keefe, 1989). These are described below.

2.2. The expectancy mechanism

The expectancy hypothesis suggests that participants create a set of expected words based on the prime and prioritize this set while searching for the target. It is assumed that if participants identify that in many trials the target is semantically and/or associatively

Table 1
Classic- versus current-model accounts of controlled priming

Effect	Classic Account (e.g., Hybrid Model)	Current Model's Account
Existence of an RP effect	High RP causes subjects to expect that primes will be followed by related targets, thus facilitating their recognition compared to when a low RP is used	High RP, in contrast to low RP, supplies opportunity to learn how to adjust network noise such that semantic transitions can be optimally modulated to decrease RTs to related targets
Increased facilitation of targets by high RP compared to low RP when the stimuli list contains many associatively related pairs	Subjects expect the target to be one of the few words that are strongly associated with the prime; if correct, the target is recognized faster compared to when no expectations are elicited	The network learns to jump to the most associated concept and "stay there" (which is optimal for associative pairs). If the target is indeed this associated concept, its recognition is accelerated due to the strong influence of the semantic network on the lexical network
Modest increase in facilitation of targets by high RP compared to low RP when stimuli list contains many category-exemplar pairs	Subjects expect the target to be one of many possible words related to the prime. There is a small benefit of going over all of them compared to when no expectations are elicited	The network learns to avoid transitions (which is optimal for related but unassociated pairs). Thus, the network prevents the decrease in semantic facilitation occurring due to transitions to concepts unrelated to the target
RP effects mainly evident at long SOAs	Short SOAs do not allow enough time to build expectations	Short SOAs do not allow enough time for semantic transitions to occur frequently, thus preventing the network from efficiently learning the benefits of certain regions in the noise-parameter space
Mediated priming does not occur in standard LDT; occurs in LDT with unmixed stimuli lists which contain only mediated pairs	Post-lexical processes such as semantic matching, which occur only in LDT, cancel out mediated priming effects caused by spreading activation. With unmixed lists, semantic matching is not attempted	Default noise in standard LDT does not allow semantic transitions that are necessary for mediated priming. With unmixed lists, the network is sufficiently exposed to the benefits of transitions (similar to high RP effects) and learns to reverse this default tendency

(continued)

Table 1. (continued)

Effect	Classic Account (e.g., Hybrid Model)	Current Model's Account
Backward priming is reduced in long compared to short SOAs in pronunciation tasks but is stable across SOAs in LDT	—	In Pronunciation tasks, the default noise values allow transitions. At short SOAs, transitions are scarce, whereas at long SOAs they are common. Backward priming is eliminated by transitions; therefore, long SOAs reduce this effect. Default noise in LDT prevents transitions to begin with; therefore, backward priming is maintained

Note. Central findings in semantic priming involving controlled processes (see reviews in Neely, 1991; and McNamara, 2005). The current model's account for these effects is contrasted with classic accounts given by the hybrid model. Note that this table provides only a partial description of the findings and their explanations and disregards other effects, such as inhibition in priming. Please see text for more details. LDT, lexical decision task; RP, relatedness proportion; RT, reaction time; SOA, stimulus onset asynchrony.

related to the prime, they tend to develop a set of expected targets from the prime's immediate semantic "neighborhood." When the target appears, this "expectancy set" is scanned first, while the general lexicon is scanned only if the presented target is not included in it. If the target is found in the expected set, its recognition time is considerably accelerated. If it is not found there, however, its recognition is delayed by this initial screening procedure. Hence, the application of an expectancy strategy can account for both the facilitation and inhibition components of the priming effect (cf., the "verification model," Becker, 1976, 1980). Once an expectancy strategy is applied, the magnitude of both facilitation and inhibition depends on the size of the expectancy set: Large expectancy sets should reduce facilitation and augment inhibition while the inverse pattern should emerge when the expectancy set is small. Becker (1980) demonstrated that the size of the expectancy set depends (among other factors) on the subject's experience-based expectations about the prime–target relation. If the target is highly predictable (as in the case of a word list with many strongly associated word pairs), then the expected set will be small, containing only those words which are highly related to the prime, and, consequently, the expectancy-based priming will be facilitation-dominant (any inhibition stemming from the initial screening procedure would be negligible). If, however, more than a few targets could be expected (as, for example, in the case of category-exemplar lists which typically contain exemplar targets with varying degree of relatedness to their corresponding category primes and none are strongly associated to the prime), a larger expectancy set will be produced and consequently priming will be inhibition-dominant. Eisenberg and Becker (1982) referred to these two substrategies as "prediction" and

"expectancy," respectively. The theory further points to several conditions that need to be fulfilled in order for expectancy to occur. First, the SOA must be sufficiently long for an expectancy set to be formulated. With short SOAs, subjects do not seem to have enough time to create such a set before the target appears. Second, the proportion of related trials in the stimulus set (the RP) must be sufficiently large. If the RP is low, subjects do not have enough opportunity to observe the occasional relatedness between prime and target and, therefore, do not engage in expectations at all. Given the above, the expectancy theory naturally explains why high RP increases priming, and it does so only at long SOAs; why inhibition appears at long SOAs but not at short SOAs; and why this inhibition depends on the stimuli list containing many category-exemplar pairs (which encourage the "expectancy" strategy) rather than associatively related pairs (which encourage the "prediction" strategy).

2.3. Semantic matching

Semantic matching refers to a hypothetical post-lexical mechanism, which is activated only during LDT and is used primarily to facilitate the rejection of nonwords and the recognition of related target words. According to Neely and Keefe (1989; see also de Groot, 1983), when subjects are required to reach a lexical decision, they can exploit any semantic relatedness between a target stimulus and its prime to accelerate the binary word/nonword response. Priming experiments employing LDT typically contain an equal number of word and nonword targets to prevent a bias toward one of the two required responses. As the real target words are divided to those related and those unrelated to their primes, there are, overall, more nonword targets than unrelated real-word targets in the stimuli list. The ratio of nonword targets of the total nonrelated targets (called "nonword ratio") is thus typically above 0.5, meaning that if no relatedness between prime and target exists, the target is most likely a nonword (and, of course, if a relation does exist, the target is obviously a real word). This information can thus allow participants to facilitate responses to nonwords, as well as to related target words. However, when the target is an unrelated real word, the "nonword" decision is obviously wrong and the need to reverse the initial tendency toward it (on the basis of additional bottom-up information) delays the response. Therefore, the facilitation of nonword responses by the semantic matching mechanism comes at the cost of delaying responses to real unrelated words, which, in turn, contributes to the inhibition component of priming. It was also shown that semantic matching, like expectancy, acts mostly at long SOAs (Neely, 1991) but the reasons for this SOA dependency are not entirely clear (McNamara, 2005).

As the presumed mechanism of semantic matching is post-lexical and hence less sensitive to the order of prime and target presentation (in contrast to forward-associative mechanisms such as expectancy), it is often considered to be responsible for the phenomenon of backward priming (Neely, 1991). In backward priming, the prime and target are only related through an association whose direction is opposite to the order of their presentation (e.g., prime – *baby*, target – *stork*; as *baby* is an associate of *stork* but not vice versa, the direction of association is from target to prime whereas the presentation order

is opposite). Many models of priming consider both automatic mechanisms and expectancy to be inappropriate for accounting for this effect as they require the prime to activate the target based on either a forward association or a clear semantic relation (none of which exists between backward-related primes and targets). Processes that are insensitive to the specific direction of association, such as semantic matching, were therefore suggested as the likely contributors to the effect. Accordingly, backward priming was initially detected only in LDT, where semantic matching is supposedly used, and not in pronunciation, where this mechanism is inactive.

In addition, semantic matching is thought to eliminate mediated priming (McNamara, 1992; Shelton & Martin, 1992). Mediated priming refers to the priming effect achieved by primes and targets which are only related through a mediating word (e.g., *dog – milk*, mediated by *cat*). As this effect can already occur at short SOAs, it is usually attributed to automatic mechanisms. However, this effect is not robust and is typically observed only when semantic matching is not likely to be applied (e.g., in pronunciation tasks, or in lexical decisions when no directly related pairs are present in the stimulus list, thus creating a noninformative 0.5 nonword ratio). The absence of mediated priming when semantic matching is applied was explained by the assumption that the inhibitory effects of semantic matching may "cancel out" the indirect facilitatory effects of automatic mechanisms as indirectly related words are recognized as unrelated by the semantic matching mechanism (Neely, 1991; see Table 1).

2.4. Caveats

Although the combination of expectancy and semantic matching (which, together with the automatic mechanism of spreading activation, was titled "the hybrid three-process theory" by Neely & Keefe, 1989) seems to satisfyingly account for most controlled priming effects (Neely, 1991), it has nevertheless run into several difficulties and contradictions. First, whereas both mechanisms should be responsible for facilitation and inhibition, the role of expectancy in producing inhibition has been challenged by findings showing no inhibition in pronunciation tasks even at long prime-target SOAs using category-exemplar pairs with high RP (Neely, 1991). Addressing this caveat, Keefe and Neely (1990) suggested that inhibition may actually stem entirely from the semantic matching mechanism, which is activated only in LDT and is irrelevant in pronunciation where no decisions are needed. According to this hypothesis, expectancy, at typical priming conditions, contributes only to the facilitation component of semantic priming (i.e., only Eisenberg and Becker's "prediction" strategy is actually used) and not to inhibition. Accordingly, it has been suggested that semantic matching is actually responsible mostly for inhibition rather than facilitation (Lorch, Balota, & Stamm, 1986; McNamara, 2005). In other words, semantic mechanisms such as expectancy only induce facilitation, whereas inhibition is entirely dependent on decision mechanisms outside the semantic system. More challenging to the hybrid model was the finding that expectancy and automatic mechanisms interact whereas according to the theory they should be independent (Balota et al., 1992; Neely et al., 2010). In addition, the contribution of semantic matching to backward

priming and to the elimination of mediated priming was called into question. First, in contrast to the early results, significant backward priming effects were found in pronunciation tasks at short SOAs (Hutchison, 2003; Kahan, Neely, & Forsythe, 1999; see Table 1). As semantic matching is not used in pronunciation, it cannot account for such effects. Second, although semantic matching is supposed to act only at long SOAs, mediated priming under typical LDT is already abolished at short SOAs (Balota & Lorch, 1986). In addition, some recent results have shown that semantic matching may actually *contribute* to certain mediated priming effects and that the existence of mediated priming in LDT may be more dependent on whether the strength of the associative connections between prime to mediator and mediator to target are sufficiently high rather than on decision-making factors (Jones, 2012).

All in all, some of the most illuminating findings in the priming literature, including the list context effects, the influence of the prime–target relation type, and the effect of task-type and SOA, have been roughly accounted for by the hybrid model; however, this ad hoc combination of different automatic and controlled mechanisms has been nothing but obvious. We turn now to describe our previous network model of priming and how it can be extended to account for many of the above findings, replacing the perspective taken by the hybrid model, as well as solve some of the difficulties. Moreover, we show how the resulting network dynamics may constitute a mechanistic ground for certain aspects of the expectancy theory. In the current article, however, we focus only on the facilitation results; the way inhibition might arise is deferred to the General Discussion.

3. The model

3.1. Basic architecture and automatic mechanisms

The currently proposed network is based on the model presented in Lerner et al. (2012a). Here, we present only the main attributes of that network (see more details in the Appendix and in Lerner et al., 2012a). The model contains two computational layers, lexical/phonological and semantic (Fig. 1). Visual input representing a word is fed into the lexical/phonological layer where the word is recognized. The activity elicited in the lexical layer is fed forward to the semantic layer where the word's meaning is stored. Importantly, these processes are interactive, so that, in addition to the feed forward transmission from the lexical to the semantic layer, the semantic layer can influence the lexical layer by feedback.

The lexical and semantic layers are modeled as attractor neural networks with sparse representations and continuous-time dynamics (see Hopfield, 1982, 1984; Tsodyks, 1990). Each network is a fully connected recurrent network composed of 500 neurons. Memory patterns encoded to each network are binary vectors of size 500, with "1" indicating a maximally active neuron, and "0" an inactive one. The representations are sparse (i.e., a small number of neurons are active in each pattern). When an external input is fed into neurons which are part of a specific memory pattern, the activity of the entire network is

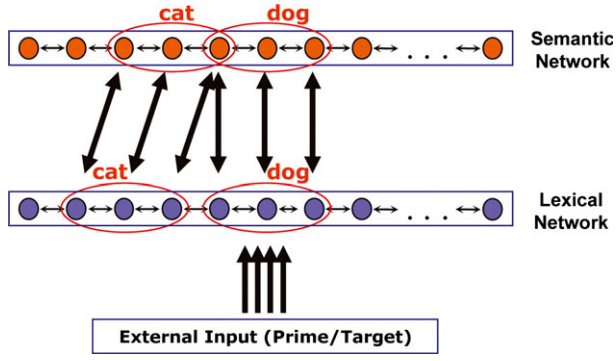

Fig. 1. Architecture of the network model. Patterns representing related concepts are correlated in the semantic network but uncorrelated in the lexical network. Active units of two toy example patterns representing "dog" and "cat" are marked. Connections between networks are from active units of a pattern in one network to all corresponding active units in the other network. For simplicity, only some of these connections are drawn.

driven by the internal connectivity to gradually converge to this pattern. The connectivity between the neurons is set according to the Hopfield weight matrix for sparse representations (see Appendix), which assures the stability of the patterns. The neurons are analog within the range [0,1] and reach binary values when converged to one of the memory patterns. They obey a logistic transfer function of their local input, which represents the total influence of both the lateral connections coming from the other neurons in the same layer as well as external inputs from other layers. External inputs are always excitatory. Gaussian noise with temporal correlations is added to the local input, inserting some degree of stochasticity to the system.

In the semantic layer, memory patterns represent concepts. Relatedness between concepts is implemented as correlations between memory patterns (reflecting the degree of overlap between them). For example, in Fig. 1, the concepts *dog* and *cat* are sharing one active neuron, making them correlated. The more two concepts are related, the stronger their correlation is; unrelated patterns have a correlation near 0.

In addition to the typical stable-state dynamics, the semantic network is also crucially influenced by adaptation mechanisms, which prevent neurons from maintaining a steady firing rate and make the network unable to hold its stability infinitely. As a consequence, with time, the network autonomously leaves the present attractor and converges to a different one. The process may repeat again and again, with the network "jumping" from one attractor to another. Such jumps between attractor states, hypothetically reflecting associative thought chains, were termed "Latching Dynamics" by Treves (2005). It was found that there is a higher probability of network transitions between correlated patterns rather than between uncorrelated ones, as the former require fewer changes in the overall activity (Herrmann, Ruppin, & Usher, 1993). Critically, this process depends on the degree of noise in the system: If the noise is very low, the destabilization caused by the adaptation mechanism would be weak and latching dynamics would not be evident; if,

however, the noise is not too low, latching dynamics would appear as described. In our model, we assumed the default noise is not low and allows for latching to take place (Lerner et al., 2012a).

In the lexical layer, encoded memory patterns represent words. The dynamics are similar to those governing the semantic network, with two important differences: There are no correlations between the word patterns in the lexical network (indicating no lexical relations between the words, such as "bat"-"rat" and "cable"-"table," mimicking the lack of such relations in typical stimuli of semantic priming experiments), and there are no adaptation mechanisms which cause latching dynamics (resulting in simple steady-state behavior with no associative transitions). The links between the lexical and semantic networks are based on connections between active neurons in corresponding patterns (see Fig. 1). An activated neuron in a certain word pattern in the lexical network sends excitatory connections to all active neurons in the corresponding concept pattern of the semantic network and vice versa. Therefore, the activation of one word pattern in the lexical network activates to different extents all related concept patterns in the semantic network, and vice versa. The bottom-up input to the lexical network, which represents visually presented words, is also excitatory and activates only the neurons that are included in the corresponding word pattern.

Lexical-to-semantic connections are strong but are also subject to synaptic depression with slow recovery time. This allows the lexical network to have a fast, short-lived influence on the semantic network, allowing it to quickly converge to the appropriate concept pattern and engage in latching dynamics with no further interference (until a new bottom-up external input arrives and the lexical network converges to a new word pattern). Semantic-to-lexical connections are weak and are not suppressed, allowing the semantic network to have a slow and enduring effect on the lexical network. This top-down influence adds to the bottom-up external influence and allows priming effects to appear: If the meaning of a newly processed word (target) is related to a concept already activated in the semantic network (prime), the lexical network will recognize this word faster than if the target is not related to the prime, because both the bottom-up and the top-down (correlation-dependent) streaming contribute to the recognition process (see Stolz & Besner, 1996, for a similar conceptualization in an interactive-activation model). Lastly, the bottom-up input to the lexical network is constant for as long as a word is visible to the system and is extinguished when the visual word disappears.

As demonstrated in Lerner et al. (2012a), the model simulates the typical priming patterns characterizing normal subject performance in semantic priming experiments under conditions favoring automatic mechanisms. Specifically, we have shown how the top-down influence of the semantic network causes directly related prime-target pairs to yield shorter convergence times in the lexical network, taken to indicate reaction times (RTs) of subjects, compared to unrelated or neutral pairs, at both short and medium SOAs, hence demonstrating priming. Moreover, we have shown how the transitions in the semantic network cause this top-down influence to constantly change, leading to modifications in the sensitivity of the lexical network to external input over time. Specifically, the transition probability from one concept pattern to another partially reflected

association strength between concepts, yielding SOA-dependent priming and allowing for indirectly related pairs, which do not have correlated representations, to nevertheless shorten RTs and yield mediated priming. Assuming backward-related pairs share correlated representations (as also hypothesized by Plaut & Booth, 2000), the model was also shown to yield backward priming which diminishes when the SOA is long, replicating common results in pronunciation tasks (but not in LDT). Finally, the semantic transitions were shown to correspond to the classic automatic mechanism of spreading activation (SA).

Although successful in demonstrating many automatic priming results, the above model did not address findings that are usually attributed to controlled mechanisms. First, as transitions between attractors caused by latching dynamics were an inherent part of the model, long SOAs could lead the semantic network to engage in "too many" transitions such that it reached concepts completely unrelated to the initial prime and thus resulted in a significantly diminished priming effect—contrary to experimental evidence. Second, as the network mechanisms were completely insensitive to regularities in the statistics of stimuli over trials, list context effects were not possible. Finally, the model did not differentiate between pronunciation tasks and LDT and therefore disregarded their different influence on priming.

3.2. Incorporating controlled processes: Basic assumptions

The primary conjecture we introduce in the current article is that in contrast to the basic model presented above, latching dynamics is not a fixed characteristic of the system but, rather, can be responsive to environmental conditions of the kind usually taken to modulate control processes in priming. Indeed, from a strictly dynamical perspective, even with synaptic depression mechanisms operating, latching dynamics is not obligatory. Rather, semantic transitions depend on the amount of noise in the semantic network, with high noise accelerating transitions and low noise delaying or even preventing them. Other parameters can have similar effects.[1] As noted above, reduction in the noise to sufficiently low values causes the immediate cessation of transitions and, consequently, forces the network to maintain its convergence on a certain memory state. Elevating the noise, on the other hand, accelerates transitions (see Fig. 2A for a demonstration of these effects). As the effective strength of the feedback from the semantic to the lexical network depends on which concept the semantic network rests on, the level of noise can have a substantial influence on the priming effect. This influence, as we will show, closely matches many of the controlled priming effects in the literature. We, therefore, propose that the level of noise in the semantic network can be modulated before and during a trial, reflecting what could be seen as the underlying neuronal mechanisms governing "focusing" and "defocusing" of attention on specific concepts and their neighborhoods, with attention acting as a signal-to-noise regulation mechanism. Focusing attention, in this scheme, is reflected by lowering the noise, while defocusing attention is reflected by elevating the noise. This control represents subjects' attempt to increase the efficiency of information processing in the network. Indeed, the notion that attention is involved in

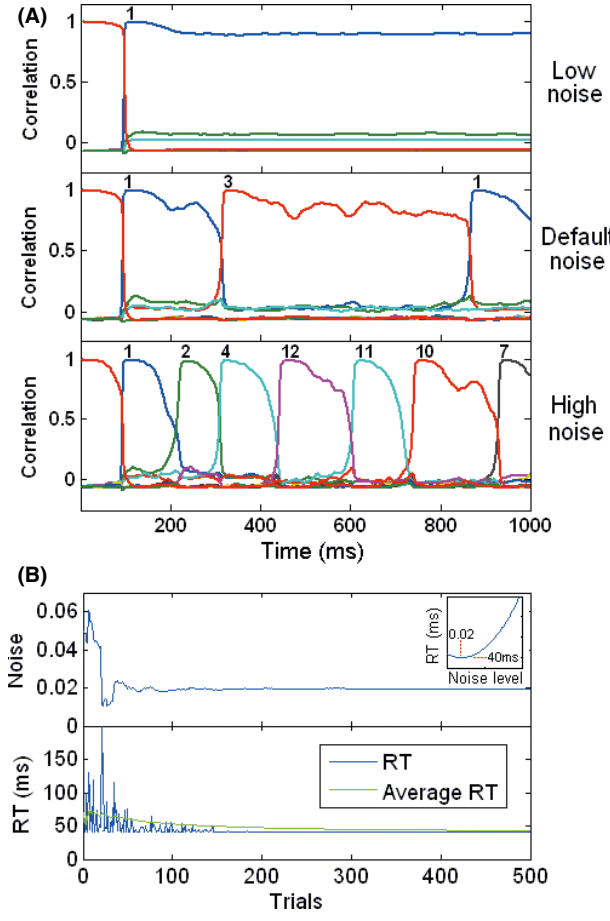

Fig. 2. (A) Correlation of the semantic network state with its stored memory patterns as a function of time, demonstrating the differences in typical transition times under various noise levels. Each pattern is indicated by a line with a different color (not all correlation lines are visible at all times, as often they coincide). The network was presented with an external stimulus representing pattern 1 for 100 ms and then allowed to run freely, jumping from one pattern to another. Moment of convergence to a specific pattern is indicated by the corresponding pattern number above the appropriate line (low noise: $\eta_{amp} = 0.01$; default noise: $\eta_{amp} = 0.05$; high noise: $\eta_{amp} = 0.08$). (B) Results of a simplified simulation demonstrating RT minimization performed by the learning rule. In this simplified simulation, RT is hypothesized to be a parabolic function of the noise with a minimum at 40 ms reached at a noise value of 0.02 (shown in the inset). In the simulation, the noise is initialized to the value 0.05 and is then allowed to change incrementally using the learning rule, with RTs corresponding to the noise value at each trial drawn from the parabolic function. Learning drives the noise to converge to its minimum value (upper panel), thus bringing the system to constantly produce the minimum RT (lower panel).

increasing signal-to-noise ratios in neurons which are active in the processing of stimuli has been suggested in the past based on an aggregate of electrophysiological and imaging studies in humans (Coull, 1998) and its feasibility has been shown in both in vivo recordings (Haider, Hausser, & Carandini, 2013) and computational modeling (e.g.,

Servan-Schreiber, Printz, & Cohen, 1990). The level of subjects' awareness of these noise modulations, however, could be partial and indeed might be translated to only a general feeling of "attendance" to certain presented or expected stimuli (see also McNamara, 2005; about the unnecessity of full conscious involvement in expectancy, as well as Rose, Haider, Weiller, & Buchel, 2002, for an example of discrepancies between neuronal and conscious measures of learning, with the latter lagging behind the former).

An additional working hypothesis we make is that subjects can identify the moment at which convergence to a semantic concept occurs and adjust their control parameters accordingly (this identification can be implemented in various ways but to remain focused on the current objective, we will not specify its exact mechanistic realization). In practice, this hypothesis is manifested by allowing the noise level to change during a trial after a transition has occurred.

3.3. Learning in the network

3.3.1 General framework

During semantic priming experiments, strategies are employed to improve task performance. In most semantic priming experiments, subjects are instructed to respond as quickly and as accurately as possible; therefore, improvement in performance is reflected by quicker and less erroneous responses. This implies that after each trial, subjects are able to monitor the outcome of their response (both the time it took and whether the reply was correct) and make corresponding adjustments in their information processing so that they would benefit from experience in subsequent trials—a classic reinforcement framework. RTs have an advantage in this scenario: As errors are binary (either a correct response was given, or not) and occur only sporadically, they provide an impoverished feedback compared to RT information. Therefore, in the following simulations, we considered RT minimization to be the main goal of subjects' strategies.

From a general perspective, if noise can be regulated during a trial, finding an optimal solution which minimizes RTs requires solving a highly complicated optimization problem with many degrees of freedom. This is neither practical nor reasonable. Therefore, for simplicity, we assume that only two control parameters are adjusted during the experimental session: the amount of noise in the system at the beginning of each trial (which is maintained until the first transition occurs; termed here, "initial noise") and the amount of noise from the first transition until the end of the trial (termed "late noise").[2] These settings, although restricted, could still allow a variety of strategies to be employed. For example, subjects could lower the initial noise in the system and avoid transitions (which, in cognitive terms, corresponds to focusing attention on the prime; in terms of the classic controlled processes, this could be thought of as actively avoiding engagement in any kind of expectancies regarding the upcoming target word); they could lower the noise after one transition and thus maintain the activation of the new concept which the system has jumped to (focusing attention on an associate of the prime, which, in expectancy terms, reflects making a prediction that this pattern would be the target); they could increase the noise values and accelerate transitions (actively avoid attending any specific

concept by "rushing" through associations); or they could withhold any significant manipulations of the noise and allow transitions to flow undisturbed, as in the default case (allowing natural "stream of consciousness"). Whether these or any other strategies are actually applied depends on the learning mechanism and the specific characteristics of the task.

3.3.2 Learning rule

As discussed above, discovering which noise adjustments are required to minimize RTs to targets during a priming task may be seen as a reinforcement learning problem. Subjects are assumed to employ, after each trial, a reinforcement learning rule that modulates the noise parameters based on the data available to them. While many reinforcement learning methods may accomplish this objective, we have chosen to use a very simple rule to prevent results from being highly sensitive to the details and efficiency of the learning process.

The local input of each neuron in our network includes a noise term drawn from a Gaussian distribution with 0 mean, and standard deviation termed "noise amplitude" (see Lerner et al., 2012a). Learning in the following simulations was based on a reinforcement rule that adjusts this noise amplitude in the semantic network (separately for the initial and late noise) to minimize the average reaction times of the system. The noise amplitude of the lexical network remained constant as in the original model.

The learning process assumed the following pattern: Each trial began with the noise amplitude (independently for the initial and late noise) set to a certain value, $\eta_{amp}(n)$ (n being the trial number). The system then "decided" to try a somewhat different noise value, $\eta_{amp}(n)+\varepsilon(n)$, with $\varepsilon(n)$ being an exploration parameter drawn from a Gaussian distribution with zero mean and variance that decays with trials: $\varepsilon(n) \sim N(0, Ae^{-\beta n})$. This exploration term allows the system to examine the effect of variable noise amplitudes around the current one, with a large average magnitude of exploration in the beginning of the experimental session and becoming insignificant as trials advance (see Appendix for specific values of the parameters). The network dynamics was then set in motion, and yielded a reaction time for that trial, RT(n). This RT was compared to the average RT of the previous trials (it is assumed that this average RT value is accessible to the system, representing an intrinsic evaluation of accumulated past performance) and induced a change in the current noise value according to the following learning rule:

$$\eta_{amp}(n+1) = \eta_{amp}(n) + \alpha\left(\overline{RT}_{n-1} - RT(n)\right)\varepsilon(n)$$

Here, $\eta_{amp}(n+1)$, is the new noise amplitude, $\overline{RT}_{n-1}$ the average reaction time of the previous trials (trials 1 to $n-1$), and α the learning rate. In a nutshell, this learning rule implies that if the recent RT was better (i.e., shorter) than the average RT, the system tends to change the noise value in the direction of the exploration parameter $\varepsilon(n)$ (note that this parameter can be both positive or negative); if it was worse (longer), it tends to change the noise value in the opposite direction. The rationale for comparing the latest RT to the (average of) previous RTs is that this difference constitutes a simple metric for

estimating local network performance: When the current trial improves results (i.e., yields shorter RT than the average RT of previous trials), it indicates that the random noise change in the beginning of the trial was probably beneficial and should be maintained. If, however, the current trial yielded degraded performance, the random change should be avoided. As a result, this learning mechanism pushes the network toward a greedy search of the noise parameter space. After the learning rule was applied, the next trial started with the new noise amplitude, η_{amp} $(n + 1)$, and the process repeated. As trials progressed, the tendency to "explore" different noise values decreased and the system settled to the state it reached without further learning. It is also worth noting that these trial-by-trial modulations are not assumed to require fully conscious involvement (see Lerner et al., 2012a; for discussion) and, therefore, should not necessarily be accompanied by a subjective feeling of "strategy change" (see also McNamara, 2005, for an argument against the necessity of conscious processes in expectancy mechanisms).

Fig. 2B displays this learning process in simplified settings that were artificially created, for the sake of demonstration, without using the neural network. In these simplified settings, we made the ad hoc assumption that reaction time is a parabolic function of the noise with a minimum at $\eta_{amp} = 0.02$ (RT(0.02) = 40 ms). Setting the default noise to 0.05 and applying the learning mechanism, with RT values of each trial derived directly from the parabolic function, we see that the noise settles down to approximately 0.02, thus bringing the system to constantly produce the minimum RT (and, consequently, the average RT) value of 40 ms. As evident in the figure, learning is not monotonic and noise values jitter up and down due to the exploratory behavior; however, the general tendency is to move toward the "correct" value and eventually settle on it.

3.4. Encoded patterns

The encoded patterns in the semantic network were chosen such that various types of relations between concepts (e.g., directly related; indirectly related; backward related; strongly associated; weakly associated) could be reflected.[3] The general structure resembled simulations in Lerner et al. (2012a), containing four semantic neighborhoods with four patterns each (patterns 1–4, 5–6, 9–12, and 13–16). Correlation strengths could be weak, medium, or strong. In the first neighborhood, all correlations had medium strengths with the exception of one pair having a strong correlation (patterns 1 and 2). In the second neighborhood, patterns 5–6 had a strong correlation, patterns 5–7 and 5–8 a medium correlation, and patterns 6–7 and 7–8 a weak correlation (patterns 6 and 8 were not correlated). In the third neighborhood, all correlations were weak and equal across all pattern pairs, and in the fourth neighborhood all correlations were equal and strong. Finally, there were two correlations between patterns belonging to different neighborhoods: Pattern 2 was strongly correlated with pattern 5, and pattern 9 was strongly correlated with pattern 13 (see Fig. 3A and B). The resulting structure created the following types of relations between pairs of patterns: strongly directly related (e.g., patterns 1–2 or 2–5); medium directly related (e.g., 5–7); weakly directly related (e.g., 10–11); and indirectly related (e.g., 1–5, mediated by pattern 2, or 2–6, mediated by

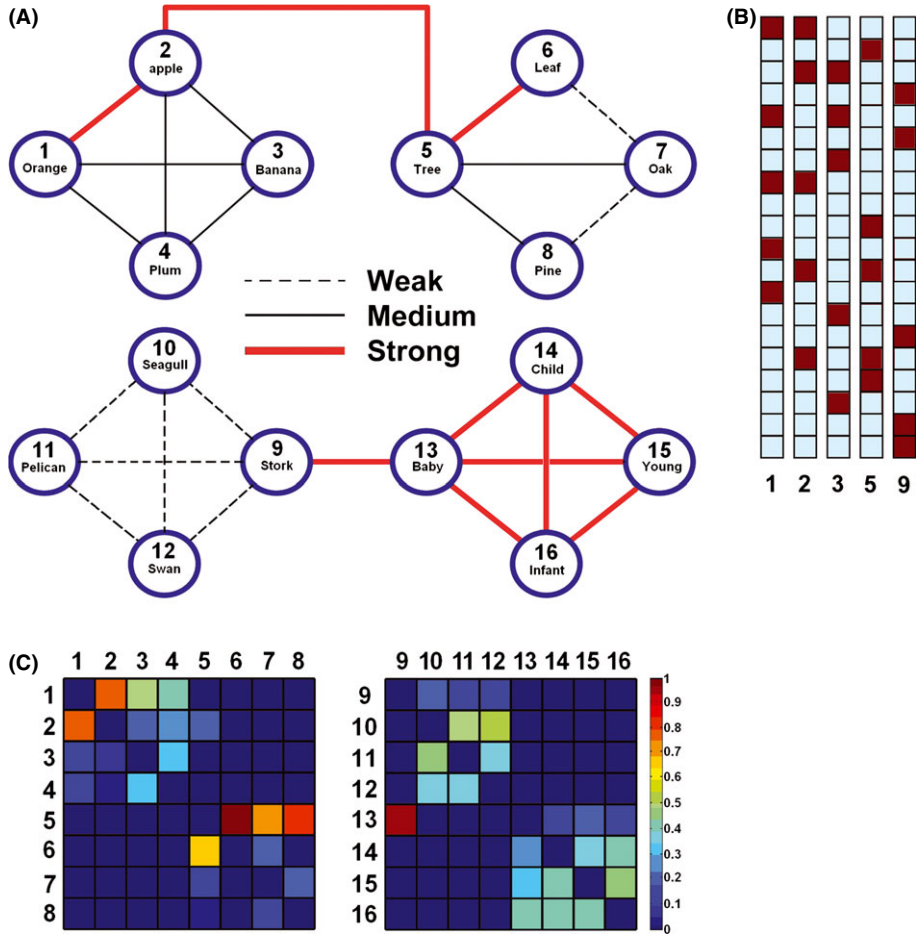

Fig. 3. (A) Structure of the semantic memory used in the simulations. Specific words are attached to the concept numbers for easier conceptualization. (B) A simplified illustration of the relatedness between concepts as represented by vector-correlations in the network, for several representative concepts (brown/light blue colors representing values of 1/0). For purposes of illustration only 20 components are shown instead of 500. (C) First-transition probabilities between the concepts appearing in A (No transitions existed between upper and lower neighborhoods; therefore, they are not shown). Probabilities are indicated by colors ranging from 0 (dark blue) to 1 (red). Columns represent the presented words and rows represent their associations.

5). In addition, as shown in Fig. 3C (which was calculated by examining which transitions occur in the network starting from a specific pattern, averaged over 100 repetitions; see Lerner et al., 2012a, for details), this structure led to different transition probabilities between certain pairs: Patterns 1–2, for example, have a strong mutual transition probability and are therefore strongly associated in both directions; patterns 9–13, in contrast, are associated only in the forward direction (9→13), with the backward direction (13→9) having a transition probability close to 0. Therefore, patterns

13–9 form a backward-related pair. Patterns 5–7 and 5–8 belong to the same neighborhood but have weak forward associations; therefore they represent, for our current concern, typical "category—exemplar" relations.

Sixteen "word" patterns were encoded in the lexical network, corresponding to the 16 "concept" patterns in the semantic network. These patterns were all uncorrelated. In addition, a seventeenth pattern was encoded in each network, uncorrelated to the rest. This pattern served as the initial state of each network at the beginning of the simulations (see Lerner et al., 2012a).

4. Simulation 1: Effect of stimuli list

4.1. Simulation 1a: Learning with strongly associated items

In this simulation we examined whether noise adjustments could be the mechanism accounting for RP effects on priming when strongly associated pairs are used. These effects have been attributed to controlled processes in human experiments, and specifically to the "prediction" strategy of expectancy (Becker, 1980).

4.1.1. Methods

Each trial consisted of the presentation of two inputs, a prime followed by a target, each being one of the pre-encoded lexical patterns. Related trials could consist of any of the strongly correlated pairs with strong transition probability from prime to target (e.g., 1–2; 5–6; 9–13). Unrelated trials included any prime from the first or second neighborhood and a target from the third or fourth neighborhood, or vice versa. Prime-target pairs were randomly chosen from within the possible combinations for each condition.

The procedure followed Lerner et al. (2012a): Each trial started with the presentation of an external input to the lexical network, which served as "prime." After 100 ms (150 numerical time steps; see Appendix), this external input was removed, and a new external input corresponding to the target was presented to the lexical network with either a short 200 ms or a long 1000 ms SOA (cf. Neely, 1991). The reaction time to a target was measured from its onset and until the convergence of the lexical network ("correct" convergence to the target attractor was always achieved). Convergence was defined as the network's state reaching a 0.95 correlation with the relevant memory pattern. In addition, the noise-amplitude values of the semantic network at the beginning of each trial (initial noise) and after the first transition in each trial (late noise) were changed according to the learning rule.[4] The simulation was run with an RP of either 0.25 or 0.75 (cf. Neely, 1991), yielding, together with the two SOAs, four independent conditions. Five hundred trials were run in each condition, with related and unrelated trials at the appropriate proportions randomly scattered across the session. This whole design was repeated 10 times to assure that learned trajectories in each condition are representative and results are robust.

4.1.2. Results

Fig. 4 displays the two noise-amplitude values along trials for each SOA and RP (mean values for each trial over the 10 repetitions are represented by the blue and green lines; the matched-color areas represent one standard deviation above and below the mean). At low RPs, these values changed only slightly compared to their starting values at the beginning of the learning process. At long SOAs with high RP, however, the system converged to a state with low amplitude value of the late noise, indicating a learned strategy of allowing a single transition in the semantic network after which the activity is maintained without further transitions. At short SOAs with high RP, there was a tendency to converge to a state with somewhat higher initial noise, indicating a strategy to accelerate the first transition (although this result was not as consistent as the one achieved at the long SOA, indicated by the larger standard deviations). Fig. 6A displays a running average of RTs over related trials (using a 50-bins window size) for one representative run of the simulation in each of the four conditions (here, and in the rest of the article, we do not uniquely address reaction time results for unrelated trials as they have not been

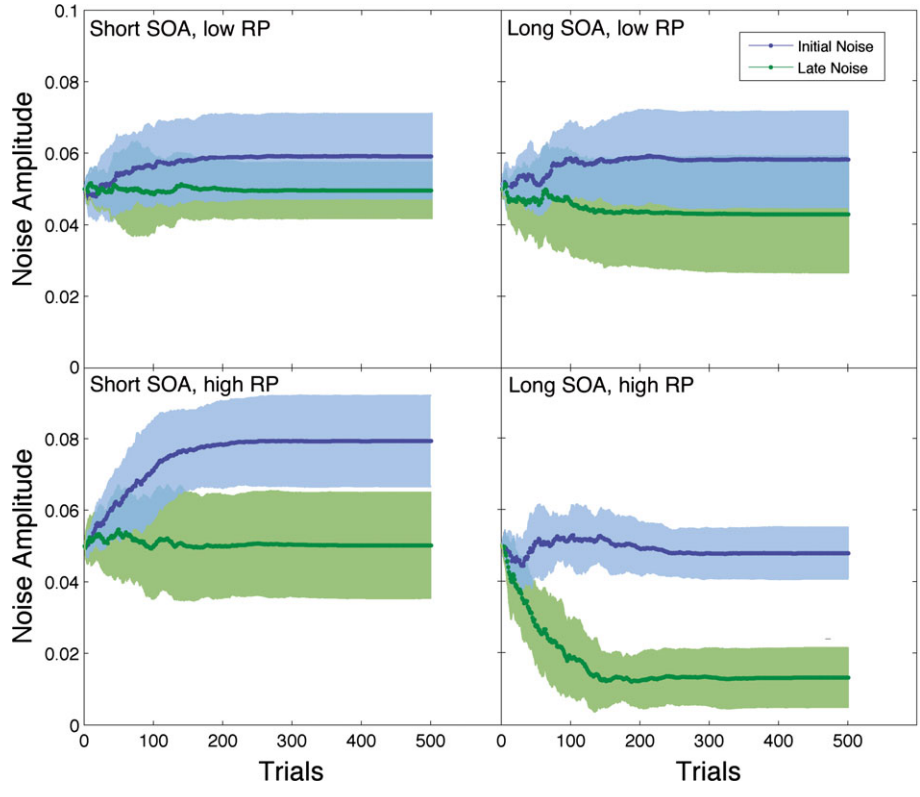

Fig. 4. Initial and late noise as a function of trial number for the various SOA and RP conditions in Simulation 1a. Lines represent average over 10 repetitions. Matched-color bands represent ±1 standard deviations above and below the average.

significantly modulated by any of our manipulations; indeed, our network produces mainly facilitatory effects, not inhibitory. See Lerner et al., 2012a, as well as the General Discussion, for discussion of this issue). Only the long SOA, high-RP condition yielded a significant modulation of RT, with the values decreasing (i.e., the facilitation magnitude increasing) as learning progresses. Fig. 6B displays, for the same representative runs of the simulation as in 6A, a running average over trials (again, using a 50-bins window size) of the average number of transitions in a trial for each RP condition at short and long SOAs. At short SOAs, transitions were rare and were not modulated throughout the session, regardless of RP. At long SOAs, the number of transitions tended to decrease toward 1 at the high-RP condition, confirming the reduction in late noise, whereas with low RP, the number of transitions increased a bit. Finally, mirroring the RT results, the priming effect (computed as the average RT of unrelated trials minus RT of related trials over the entire session) was modulated by RP only at the long SOA. At the short SOA, priming was 33 and 35 ms for the low and high RP, respectively. At the long SOA, in contrast, priming was 33 ms at the low RP and 46 ms at the high RP (see Fig. 7 for comparison between short- and long-SOA results at high RP and representative experimental findings in humans).

4.2. Simulation 1b: Learning with category-exemplar items

The current simulation, as the former one, examined how noise regulation affects priming under different RP and SOA conditions, but this time using weakly associated pairs. Experimentally, such pairs usually produce an equivalent amount of facilitation at short and long SOAs when the RP is high. In addition, when the RP is low, facilitation is often reduced at long SOA compared to high RP (Neely, 1991).

4.2.1. Methods

Primes and targets of related trials in the current simulation mostly belonged to the second neighborhood and could either be 5–8, 5–7, or 5–2 (see transition probabilities in Fig. 3C). The existence of a correlation between these pairs which is not accompanied by a high transition probability is analogous to category–exemplars relationship, commonly used in human experiments, which are typically only weakly associated in the forward direction (i.e., from category to exemplar; see McNamara, 2005). The procedure was similar to that used in Simulation 1a.

4.2.2. Results

Fig. 5 displays the mean noise values and standard deviations for the four SOA/RP conditions. As can be seen, replicating Simulation 1a, at both RP conditions at short SOAs and for low RP at the long SOA, there were no robust noise modulations during the session. However, with long SOA and high RP, the initial noise was significantly decreased across trials, indicating the selection of a strategy that minimizes transitions and maintains focused semantic activation on the prime. Fig. 6C and D display a running average of the RTs of related trials and of the number of transitions for one

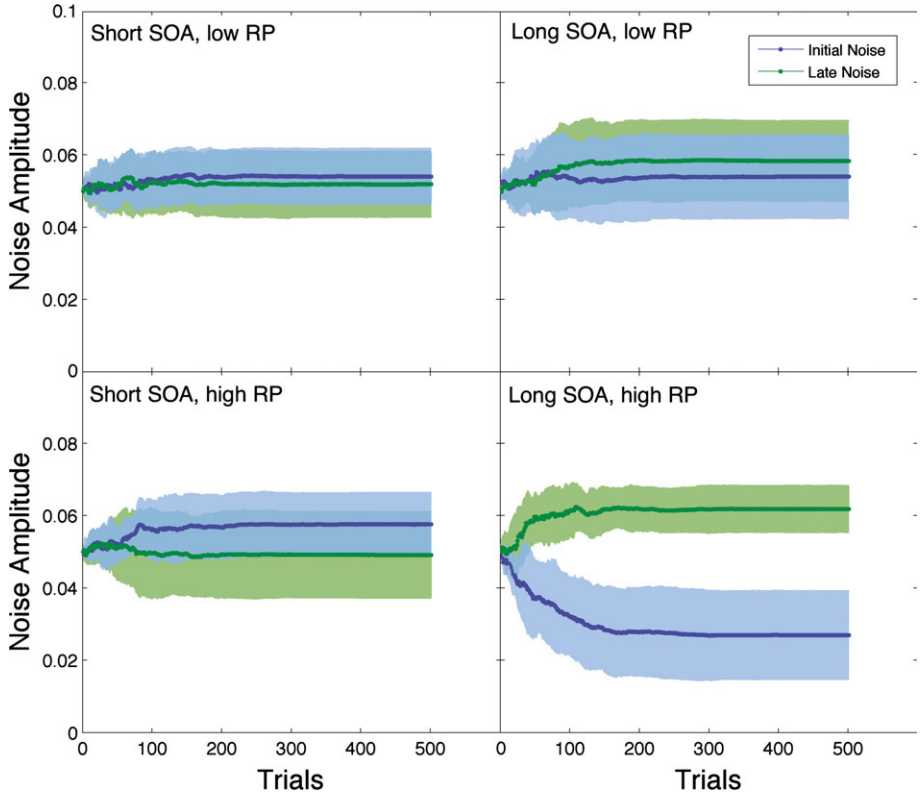

Fig. 5. Initial and late noise as a function of trial number for the various SOA and RP conditions in Simulation 1b. Lines represent average over 10 repetitions. Matched-color bands represent ±1 standard deviations above and below the average.

representative run of the simulation in each of the four conditions. There was a small but significant decrease in RT at the long SOA when the RP was high, and no modulation of RT in the other conditions. Similarly, the number of transitions was modulated only with long SOA and high RP; in this condition, there was a significant decrease toward 0 in the number of transitions along the experiment. The priming effect was modulated by RP in the long SOA condition (14 ms compared to 23 ms for the low vs. high RP, respectively) but not in the short SOA condition (24 ms vs. 22 ms for low and high RPs). Fig. 7 compares some of these results to representative human findings under the same conditions, showing similar trends.

4.3. Discussion of Simulation 1

The results of Simulations 1a and 1b showed how control over the probability of semantic transitions, targeted at minimizing RTs, can yield significant modulation of the priming effect that corresponds to experimental results in humans under the same

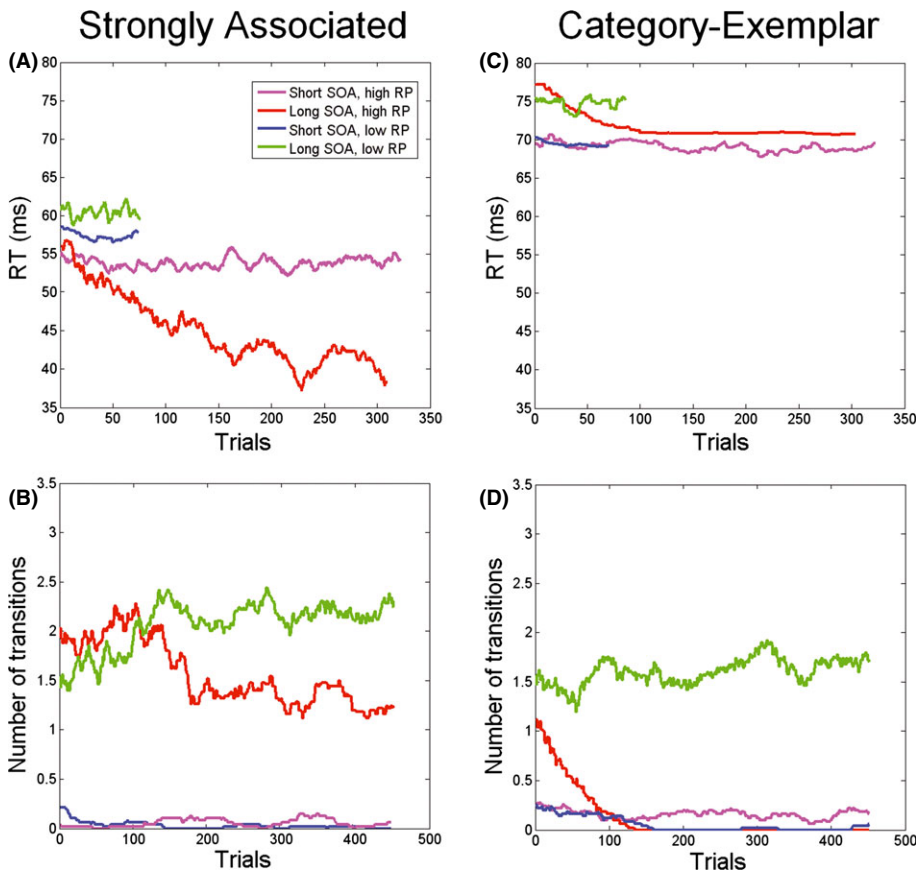

Fig. 6. (A) Reaction times of one representative run of the simulation in each condition for the strongly associated related targets in Simulation 1a as a function of trial number. Reaction times are computed using a moving average over a 50-trials window; consequently, simulations with high RP (containing many related pairs) have more data points compared to simulations with a low RP. (B) Number of transitions in a trial in Simulation 1a as a function of trial number (computed for the same representative trials in A, using the same moving average). (C) The same as A, for the category-exemplar pairs in Simulation 1b. (D) The same as B, for Simulation 1b.

conditions. When the items list contained pairs with a high prime-to-target transition probability (as in Simulation 1a), the system learned to allow one transition and reduce the probability of additional transitions. Using this strategy maximized the benefits from "predicting" the correct target. When, in contrast, weakly associated pairs such as category-exemplars were used (as in Simulation 1b), transitions were not beneficial and could have even been disadvantageous; therefore, the system learned to avoid transitions altogether. Successful optimization in both simulations, however, was not robust and required a sufficient frequency of relevant pairs from which the system could learn. Unrelated pairs cannot be optimized one way or the other and therefore do not help in finding the optimal strategy. Only related trials allow optimization and so RP became an essential

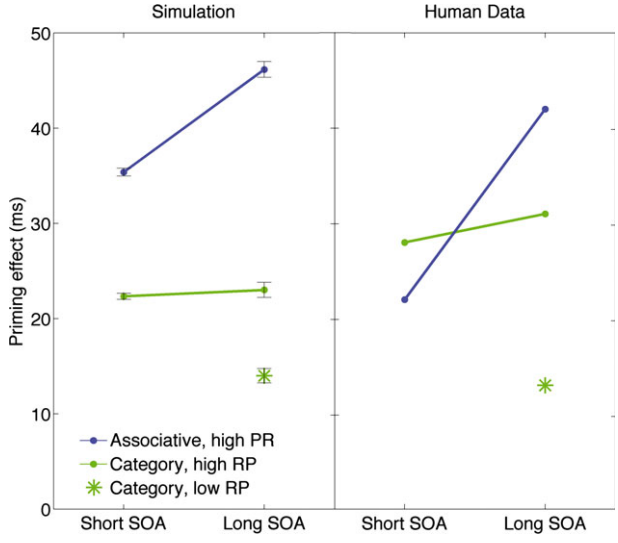

Fig. 7. Main priming effects of Simulation 1 alongside corresponding findings from human studies. Human results are taken from Neely, 1991, table 6 (facilitation effects) and table 8 (low RP, high-dominance exemplars in pronunciation). Error bars represent ±1 standard error of the mean.

modulator of learning. Likewise, short SOAs are often not lengthy enough to allow the system to commit any transition whether the noise is high or not; therefore, the system cannot efficiently scan through the state space at short SOAs and is thus struggling to find a better solution than the one presented by the default noise values. Only sufficiently long SOAs, which permit transitions, allow consistent optimization to take place. This dependency of optimization on high RP and long SOA in the model mirrors the common findings in the literature, which indicate these very same conditions are required for controlled processes to operate (Neely, 1991). Despite some numerical differences, the simulation results clearly match the patterns of priming found in human studies (Fig. 7). In addition, when only high RP is considered, RT optimization has a somewhat different role comparing associative and category-exemplar pairs: Although priming for associated pairs actually increased from short to long SOA, the priming effect for category-exemplar pairs was merely prevented from decreasing at long SOAs (see the corresponding RTs in Fig. 6). Thus, even when optimization takes place, long SOAs may increase facilitation for associated pairs compared to short SOAs but do not affect facilitation of category-exemplar pairs, consistent with experimental data (cf. Neely, 1991).

Although short SOAs did not generally yield RT modulation, there was some indication that an optimization process does occur under short SOA when a high RP and associative pairs were used. This was evident by the fact that the system increased the initial noise value under these conditions. This result is not surprising: Since transitions are beneficial when associative pairs are used, and since short SOAs barely allow transitions with the default noise value, increasing this noise can compensate for the lack of sufficiently long SOA. However, this modulation did not result in a substantial RT minimization

because even with a somewhat higher noise value, transitions were still not common enough. Therefore, whereas our model does not completely rule out controlled priming effects at short SOAs, it indicates that these may often prove to be inefficient. This state of affairs may be seen as analogous to a case where subjects notice that a target word is often related to the prime, attempt to predict the target on each trial based on the prime, but face difficulty in actually doing so due to the short lag between the two stimuli.

It is important to note that the learning rule which we chose to use is neither optimal nor necessary. While being quite straightforward and simple, it requires subjects to estimate the average RTs of previous trials and necessitates small incremental adjustments of the noise parameters. Other reinforcement learning rules may lead to faster and more efficient learning, possibly requiring fewer trials to take effect and, as a result, perhaps yielding larger differences between conditions. However, it is clear that low RP and short SOAs would still be inferior to high RP and long SOAs in producing RT optimization as they allow only little opportunity to learn. In addition, it may be argued that despite our 10 repetitions of each condition in the simulation, the fact that the system learns a certain solution does not mean that it is optimal, and, indeed, the solution which was found could be merely a local minimum of the problem domain. Particularly, it is possible that different starting values of the noise at the beginning of the learning process (as is actually hypothesized in the next simulation) might have led to different, possibly better solutions. To examine this possibility, we took advantage of the fact that the parameter space is just two-dimensional, allowing for direct scanning of various initial and late noise values without a learning process. Fig. 8 shows the average RTs across 100 trials for a range of initial and late noise values using the same network and stimuli as in Simulations 1a and 1b, but without applying the learning rule. As can be seen, minimum RTs are achieved for initial and late noise values which are approximately the same as the ones discovered by the learning network, indicating that the results indeed represent a global minimum. Hence, the network did not merely learn an incidental local solution stemming from our choice of using a 0.05 default noise at the start of the simulations, but, rather, learned—when the RP and SOA allowed it—the best configuration of noise parameters which minimizes RTs in the relevant condition, as we speculated to be the role of controlled processes in priming.

5. Simulation 2: Mediated and backward priming in LDT and pronunciation

Priming effects in LDT are usually stronger than in pronunciation tasks (Neely, 1991; but see Hutchison, Balota, Cortese, & Watson, 2008). The main reasons for this difference are usually considered to stem from outside the semantic system: First, LDT employs decision-making mechanisms that may contribute to the priming effect whereas pronunciation does not; second, pronunciation, in contrast to LDT, may involve direct orthographic-to-phoneme conversion which bypasses the lexicon, therefore reducing the effect of the semantic level on information processing and diminishing the semantic priming effect accordingly. However, several important types of priming effects which

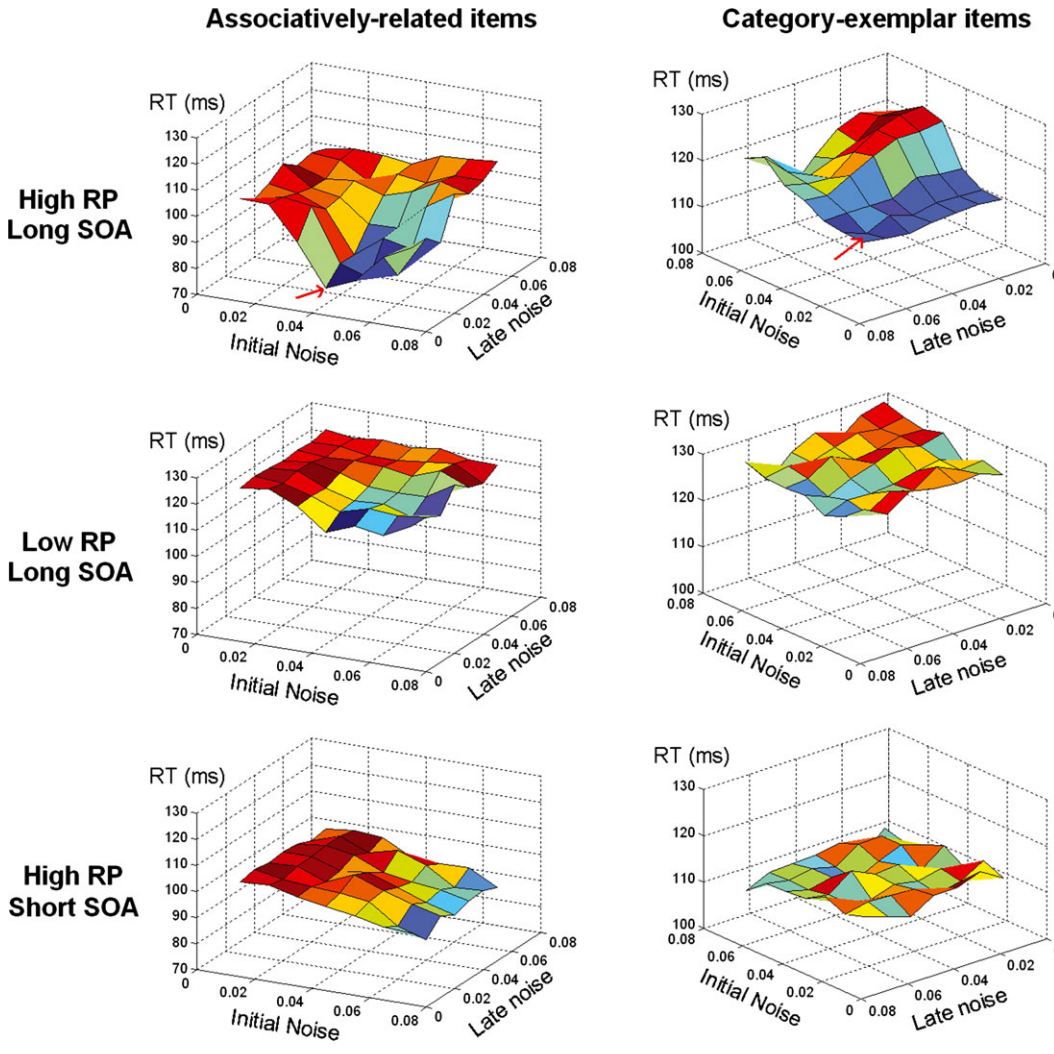

Fig. 8. Mean RTs (over both related and unrelated trials) for various RP, relatedness, and SOA conditions over a range of initial- and late-noise values. Optimal values are clearly observed in the high RP, long SOA condition: Lists containing associatively related items achieve minimum at a medium initial-noise value combined with a low late-noise value, whereas lists with category-exemplar items achieve minimum at a low initial-noise value without a strong preference for any late-noise values (minimum RTs are marked by red arrows). When the RP is low or when the SOA is short, RTs are much less modulated by the noise, with only a slight advantage to high initial-noise values for associatively related items at the short SOA, high RP condition (bottom left figure).

distinguish LDT and pronunciation cannot be accounted for by these explanations: One is the absence of mediated priming in LDT under typical procedures, although it is significant in pronunciation when similar procedures are used (Balota & Lorch, 1986). Mediated priming does appear in LDT, however, if no directly related pairs are included in the

stimuli list (Hutchison, 2003). Another difference is that backward priming appears in both tasks at short prime-target SOAs, but at long SOAs it appears only in LDT (Hutchison, 2003). As reviewed earlier, previous explanations which have been provided for these findings are not satisfactory. The current model, in contrast, can account for these patterns assuming that in LDT, unlike pronunciation, the default value of the noise at the beginning of the experimental session (before learning initiates) is low. In other words, we assume that in pronunciation subjects begin the experimental session with unfocused semantic activation which allows semantic transitions, whereas in LDT they begin the session with focused activation and no transitions. In what follows, we first present a simulation which demonstrates how such an assumption yields priming effects which fit well with the experimental results using mediated and backward-related pairs. We then discuss in detail possible sources supporting this assumption.

5.1. Methods

We examined the influence of the default noise on mediated priming starting the priming simulation with two different noise values. In the pronunciation-analog condition, the default noise values were set to 0.05, as in the earlier simulations. In the LDT-analog condition, the default noise was set to 0.01, a value which does not encourage semantic transitions (see Fig. 2). Similar to previous human experimental procedures (cf. Balota & Lorch, 1986), the SOA was either 250 or 500 ms and the RP was low (0.33). Items consisted of three pair types ("mixed" list condition): Directly related pairs that were not strongly associated (e.g., 5–8; 5–2), mediated pairs (e.g., 9–14; 1–5), and unrelated pairs. All pair types were randomly distributed across the session. In addition, to explore whether omitting directly related pairs influences mediated priming in LDT, we ran another simulation in the LDT-analog condition using a list of items which contained only the mediated and unrelated pairs (a mediated-only, "unmixed" list condition). All other conditions in this latter simulation were similar to the mixed list.

Another set of simulations was run to examine the influence of the default noise on backward priming. The SOAs in those simulations were 150 and 500 ms, with a high RP of 0.8 (cf. Kahan et al., 1999). The stimuli list was mixed, consisting of 20% backward-related pairs (which were always 13–9, the only real backward pair in the network; see Fig. 3C), 60% were related pairs (chosen from the same pairs as in the mediated-priming simulations), and 20% unrelated pairs. In addition, although not tested in human experiments in the past, we explored the possibility that inclusion of direct pairs might be the source for absence of backward priming at long SOAs in pronunciation tasks (as they hamper mediated priming in LDT; see Discussion). To test this assumption, another simulation was run in the pronunciation-analog condition using a stimuli list consisting of 80% backward-pairs and 20% unrelated pairs (a backward-only, unmixed list).

The procedure was identical to the previous simulations, with the learning rule active throughout all sessions. Only results of a single run in each condition are presented as there were almost no differences between learning trajectories over multiple runs.

5.2. Results

Fig. 9 displays the initial and late noise values for the mediated and backward priming simulations. Noise values were not modulated significantly in any of the mixed lists and remained in the vicinity of the starting values (0.01 in the LDT-like condition, 0.05 in the pronunciation-like condition). The initial noise was modulated, however, in the unmixed lists: The mediated-only list caused the initial noise in LDT to rise up, allowing semantic transitions to occur. The backward-only list, in contrast, had an effect only at the long SOA, where it showed the opposite pattern: The initial noise in pronunciation

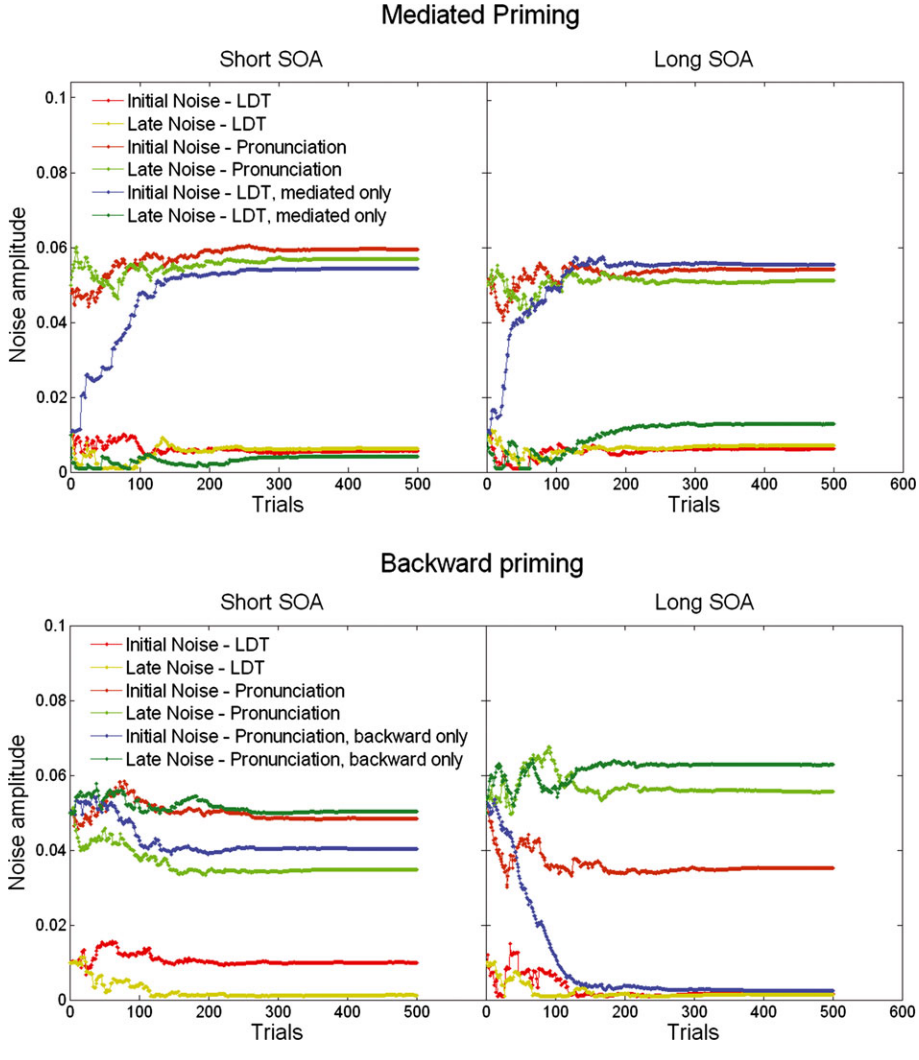

Fig. 9. Initial and late noise as a function of trial number, SOA, and task for the mediated and backward priming conditions in Simulation 2.

decreased to LDT-like values, indicating a strategy to cease transitions and focus semantic activation on the prime.

Priming effects in each condition are presented in Fig. 10 alongside their respective patterns in corresponding human experiments[5] (human results are presented by dashed lines or by single star marks). In the mixed list, mediated priming was significant in the pronunciation-like condition and also increased slightly from short to long SOA. In the LDT-like condition, in contrast, mediated priming was near zero at both SOAs. These effects closely resembled the human experiment results. When the mediated-only list was used, mediated priming effects in the LDT-like condition were similar to those achieved using the mixed list in the pronunciation-like condition. The backward priming effects using the mixed list were robust and roughly equal at the short and long SOAs in the LDT-like condition; in the pronunciation-like condition, in contrast, there was a significant decrease from short to long SOAs. These patterns, once again, resembled the corresponding human experiments, although the strength of the backward priming effect at short SOA using pronunciation was higher in the simulation. When the backward-only list was used, the backward priming effects in pronunciation resembled their corresponding effects using LDT in the mixed list.

5.3. Discussion

The results of Simulation 2 show that task-related differences in mediated and backward priming can be generated by the model under the assumption that subjects approach LDT and pronunciation differently: LDT with focused semantic activation which prevents transitions and pronunciation with unfocused semantic activation which allows transitions.

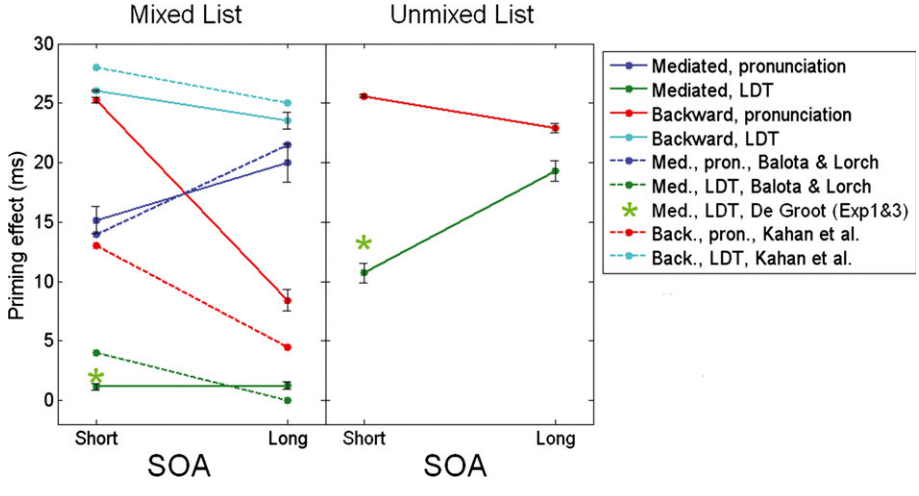

Fig. 10. Priming effects for the various conditions in Simulation 2 (solid lines), alongside corresponding results from human studies (dashed lines, and star marks). Human results are taken from Balota and Lorch (1986), Kahan et al. (1999), and de Groot (1983). Error bars represent ±1 standard error of the mean.

These two initial tendencies may then be altered during the session if encouraged by the stimuli list. However, whereas mixed lists may not be consistent enough to induce any strategy, unmixed lists, in which all related items can benefit from the same manipulation, do encourage them. For example, mediated-only lists contain (indirectly) related primes which can be exploited to reduce RTs only in the condition a semantic transition has occurred; therefore, the system, when starting from a low noise value (as in LDT), learns a strategy in which the initial noise is raised to induce transitions. This effect, in contrast to simulation 1, was already evident at short SOAs, suggesting that although short SOAs provide rare opportunities to learn (only occasional transitions can occur when the SOA is short), when the stimuli list is consistent enough the system can take advantage of these rare opportunities (note, however, that the short SOA in this simulation was longer than the short SOA in simulation 1, 250 ms vs. 200 ms, respectively, which, by itself, provides some more opportunity for transitions to occur). Backward-only lists, in contrast, contain primes which lose their facilitatory effect if transitions occur. Consequently, the system learns, starting from a high noise value (as in pronunciation), to minimize transitions by lowering the noise and hence maintain focused activation on the prime.

The priming results closely matched corresponding experimental findings in the literature. One difference between the simulation and the human results which did occur, though, was that the simulation produced a higher backward priming effect in pronunciation at short SOAs compared to the corresponding result in humans. However, the actual size of the priming effect in humans is influenced by several processes that are not covered by the current model; specifically, the model does not address the possible use of direct orthographic-to-phoneme conversion which bypasses the lexicon and, consequently, reduces the priming effects in pronunciation. As a result, the exact magnitude of the priming effect resulting from the simulations is less important in the current context than the various patterns of priming which the model yielded under different experimental manipulations (SOA, the nature of the stimuli list, the task, etc.). As was shown, these patterns clearly matched the human data. In the unmixed lists, the mediated-only condition also yielded the experimentally established effect of mediated priming in LDT when no directly related pairs are included in the stimuli list (e.g., Hutchison, 2003). Backward-only lists, however, were not examined in the past. Therefore, the results of the simulation in the pronunciation-like condition using backward-only list set up a novel prediction that can be tested in future experiments.

The results of Simulation 2 were based on the assumption that LDT and pronunciation differ in the default focus of activation with which subjects approach the task. However, why should such a difference exist? There may be several reasons. First, the pronunciation task is considered to engage more "automatic" processes compared to LDT as the subjects' task—naming written words—is more natural and mirrors everyday reading behavior. LDT, in contrast, requires an uncommon response (deciding whether a word is real or not) which might encourage subjects to take a cautious approach at the beginning of the session and avoid defocusing semantic activation off the prime. Another possibility, however, may be related to post-lexical decision-making processes operating in LDT but not in pronunciation. There is considerable evidence that after target appearance, sub-

jects tend to examine whether it has any semantic relations to the prime to accelerate their decision regarding its lexical validity (cf., the semantic matching strategy; Neely & Keefe, 1989). In particular, subjects interpret a lack of such relations as indicating that the target is most likely a nonword. In that case, unfocused semantic activation may compromise this effort if the semantic network jumps from a related prime (which eases the decision process) to a less related or unrelated concept (which hampers it). Consequently, from the point of view of RT optimization, it may be more beneficial for subjects to suppress semantic transitions when post-lexical strategies are involved. Indeed, mediated priming—which necessitates semantic transitions—does appear in several variations of LDT in which prime-target comparisons are discouraged (e.g., the continuous-LDT; the go/no go task. See Neely, 1991). Future developments of the model may allow an accurate expression of this hypothesis by incorporating a full decision-making mechanism that would use learning rules that take under consideration post-lexical strategies.

6. General discussion

Priming effects at long SOAs often appear to involve controlled strategies employed by subjects to optimize their performance in the task. These strategies have been posited to explain observed priming effects that are not easily accounted for by automatic mechanisms. In two simulations, we have shown that some of the key phenomena attributed to such strategies can be interpreted in our network model as the modulation of semantic transitions aimed to minimize RTs using simple reinforcement learning. The principles which govern this learning are identical in spirit to those suggested by previous theories, and thus result in a similar sensitivity of priming to RP, SOA, and type of prime–target relationship (see Table 1). However, the current implementation gives an exact and quantitative formulation to these principles and connects them to a mechanistic neuronal network model of semantic memory in which the controlled strategies get an accurate interpretation. In addition, the way automatic and controlled processes combine is clearly stated in the current work, suggesting that both operate on a common representation scheme within an attractor neural network whose default dynamics can be influenced by subject-controlled signal-to-noise modulations.

The interaction between automatic and controlled processes in our model may have nontrivial consequences. Essentially, it suggests that controlled processes are greatly influenced by the automatic dynamics of the network. More particularly, it implies that controlled strategies cannot compensate entirely if these dynamics are to become dysfunctional. For example, we have previously shown that the behavior of schizophrenic patients in semantic priming experiments can be explained by an increased rate of transitions in the network, caused by a deficient synaptic depression mechanism (Lerner et al., 2012b). This increased rate, while stemming from the automatic components of the network, also results in a reduced ability of the system to efficiently exert fine control of the dynamics through noise modulations, which in turn may manifest as a deficit in the subject's attention and ability to use contextual information for applying strategies (Braver, Barch, & Cohen, 1999; Lerner

et al., 2012b). Consequently, what seems to appear as a deficiency in control might actually hide an underlying deficit in automatic mechanisms. In this sense, our model does not support distinguishing between automatic and controlled processes purely on the basis of task performance under given conditions (see also Logan, 1988). A more fundamental distinction, based on computational principles, is offered at the end of the current discussion.

It is tempting to interpret the suggested control over transitions as partially reflecting the expectancy mechanism, or at least the "prediction" strategy suggested by Becker (1980). First, similar to the current model, the expectancy mechanism was originally formulated as operating on the same representations elicited by automatic mechanisms, increasing the activation of expected concepts by means of selective attention and inhibiting unattended concepts (Posner & Snyder, 1975). Second, the "prediction" strategy suggested by Becker (1980) as part of the expectancy mechanism claims that facilitation, unlike inhibition, is caused mainly by the activation of a single or very few predicted concepts based on the characteristics of the stimuli pairs in previous trials. This resembles the association with a single "predicted" concept in the current model. Third, a recent study by Neely et al. (2010) suggests that under conditions of cognitive load, activation of concepts is eliminated within 1,200 ms after the target onset unless expectancy mechanisms are initiated by the use of a high RP, similar to the way the loss of activation of a semantic neighborhood due to repeated transitions in the current model is reversed by controlled noise modulations. If, however, no cognitive load is applied, priming was found to be persistent at long SOAs even with a low RP, similar to the way priming is maintained in our network due to the default tendency to avoid transitions in LDT (cf. Hutchison, Neely, & Johnson, 2001; Neely et al., 2010). Finally, there are indications that spreading activation and expectancy effects interact (Balota et al., 1992), a result which fits well with the view that both are based on the same mechanism, as in the current framework. Nevertheless, the proposed theory resembles the expectancy mechanism only to a certain extent: Although the expectancy theory suggests that in some circumstances expecting wrong targets results in inhibition of responses, the presented model speaks only of facilitation. In addition, the current model describes control only over the rate of transitions; which transitions occur, in contrast, is entirely determined by the correlation structure. Some studies, however, found more subtle expectancy effects directed only toward certain types of prime–target relations (e.g., McKoon & Ratcliff, 1995). Such effects cannot be reproduced by the model in its current form. Some natural extensions of the model, however, may allow it to express these effects, as discussed later.

It is important to emphasize that the suggested theory of controlled processes does not only constitute a partial mechanistic implementation of past theories but also offers a novel contribution to the understanding of mediated and backward priming. As previously described, two important results concerning these priming effects were not sufficiently accounted for by previous automatic and controlled models of priming: the absence of mediated priming in LDT at short SOAs; and the existence of backward priming in pronunciation at short SOAs but not at long SOAs. To reiterate, the absence of mediated priming in LDT is usually attributed to its being concealed by the employment of a semantic matching strategy. However, there is evidence that this strategy, like

expectancy, requires a long SOA to develop (Neely, 1977; Neely & Keefe, 1989), thus casting doubt on its ability to comprise an important modulating factor at short SOAs. The existence of backward priming in pronunciation, on the other hand, can be explained by previous theories only if backward-related items are believed to share some sort of semantic relations; but in that case, the effect should also persist at long SOAs—contrary to experimental findings. The current model solves these two difficulties by suggesting which control parameters might be effective even at short SOAs (i.e., initial noise) and how LDT and pronunciation differ from one another such that task-specific long-SOA outcomes are produced (i.e., different default noise values). This way, the two types of priming effects are accommodated within a general framework describing the typical processes occurring in semantic memory.

Our model also offers some insight into the typical differences in priming patterns found in LDT and pronunciation when high- versus low-dominance exemplars are used (Neely, 1991). As it turns out, when the experimental stimuli list is made of category-exemplar pairs, both high-dominance exemplars (e.g., *flower – rose*) and low-dominance exemplars (e.g., *flower – lily*) lead to semantic priming in LDT, but only the high-dominance exemplars consistently show priming in pronunciation. Moreover, RP effects are not evident for low-dominance exemplars in pronunciation (Neely, 1991). These results are accounted for in our model by the presumed low correlation between the representations of low-dominance prime-target pairs in the semantic network, as well as by the presumed differences in initial noise between LDT and pronunciation. First, a low correlated pair is prone to exhibit small priming effects. As semantic priming effects in lexical decision are stronger than in pronunciation (but see Hutchison et al., 2008), the combination of an insensitive task and a target with weak correlations to the prime may explain why priming is not significant under these conditions (for a related account, see McNamara, 2005). Moreover, since in pronunciation the system indulges in transitions by default, and these transitions would almost never be to a low-dominance exemplar, any effect of the prime on the low-dominance exemplars would decrease even further. In LDT, in contrast, the system does not make transitions by default; therefore, the small priming effect produced by such pairs would be preserved. Finally, as RP can only make a difference when there are priming effects to begin with, no RP effect is expected for low-dominance prime-target pairs in pronunciation.

The model's account for mediated and backward priming leads to three important predictions that have not been examined so far in semantic priming research with human subjects. First, the model predicts that mediated priming in typical LDT should be evident when expectancy to highly associated targets is strongly encouraged, compared to its absence when such expectancy is not induced. This prediction comes from the fact that mediated priming in the model strictly depends on a semantic transition and such a transition is encouraged when subjects expect strongly associated pairs. Note that this prediction lies in contrast to the more common view (e.g., Neely, 1991) stating that mediated priming is concealed when controlled processes are encouraged. In addition, if this prediction is validated, it would support another major premise of the model, namely, that controlled processes such as expectancy, and automatic processes such as spreading

activation, do, in fact, interact (in contrast to their additive effect hypothesized by the hybrid three-process theory; see also Balota et al., 1992).

A second prediction is that when expectancy to highly associated pairs is encouraged, backward priming in lexical decision should mimic the pattern seen in pronunciation, namely, a significant reduction in the effect when moving from short to long SOAs. This prediction is based on the hypothesis that backward priming at short SOAs reflects the existence of correlated representations between primes and targets while its absence at long SOAs in pronunciation is the result of semantic transitions. Inducing expectancy to strong associates encourages transitions and so backward priming effects should diminish —just like in pronunciation tasks. Finally, a third and complementary prediction, which was mentioned earlier, is that using a backward-only list in a pronunciation task (i.e., a list which strongly encourages withholding transitions) should result in backward priming at both short and long SOAs, since without transitions the prime-target correlations are preserved. For the same reasons, repetition priming may also be strengthened using a stimuli list with many backward-related pairs compared to a list with many highly associated pairs as the former will push the system to maintain prime activation (however, repetition priming is affected by many other levels of processing, including orthographical and phonological; therefore, any predictions regarding this effect should be treated cautiously).

6.1. Caveats and future developments of the model

A phenomenon that our model cannot easily address in its present form occurs when subjects in a priming experiment are led to expect specific kinds of targets rather than targets that are generally associated with the prime. For instance, if the stimuli list in a priming task contains mostly antonyms, it was shown that the priming effect is significantly higher for antonym pairs compared to prime-target pairs with a different (and therefore unexpected) type of semantic relation (McKoon & Ratcliff, 1995). Similarly, if subjects are instructed by the experimenter to expect, for example, instances of body parts after seeing the prime "building," pairs which are congruent with the instructions will produce higher priming effects than incongruent pairs (Neely, 1977), although no prior semantic relation exists between the prime and target. What these examples have in common is that instead of relying on frequent associative relations learned during lifetime, subjects base their expectations on episodic relations that were acquired during the experimental session. Such expectancies could, in fact, be implemented in the model if we assume that temporary, episodic connections are formed between active units of certain patterns in the semantic network. For example, if memory pattern μ is episodically connected to memory pattern υ, the connectivity matrix should be reformulated to $J_{ij}^{\mathrm{NEW}} = J_{ij}^{\mathrm{OLD}} + \kappa \epsilon_i^{\upsilon} \epsilon_j^{\mu}$, with κ being a positive coefficient indicating the relative weight of the episodic connections compared to the original ones. With such connections added, Herrmann et al. (1993) showed that the network, when converged to pattern μ, started having significant transition probabilities to pattern υ, even if the two patterns were uncorrelated, and these probabilities increased with κ. Many idiosyncratic connections can be added this way and they will all influence the transition probabilities in the

network. Other memory patterns would not be affected and the general behavior of the network should remain the same. If, in a priming experiment, two such episodically connected patterns are presented as prime and target, a significant priming effect would emerge due to the prime-to-target transitions. In the more delicate case, when a certain *type* of relation needs to be learned rather than an idiosyncratic connection between unrelated concepts, a more complex structure of the semantic storage must be assumed in which representations of concepts include a portion of units which reflect functional relations (e.g., antonymity). In this case, formation of specific episodic connections between certain functional units (rather than between the entire units active in two concepts) can elevate the probability of transitions between concepts with the corresponding relation, thus allowing the instantiation of complex expectancies.

Naturally, the way episodic connections are actually instantiated in the brain may be more elaborate than the raw illustration presented above. Episodic learning is assumed to involve the hippocampus, which, in turn, exchanges information with the neo-cortex where semantic knowledge is presumably stored (McClelland, McNaughton, & O'Reilly, 1995; see also Winocur, Moscovitch, & Sekeres, 2007). Therefore, implementing episodic associations as direct synaptic connections between concept patterns may be an oversimplification. In reality, information might be transferred by a readout mechanism from the semantic network to other areas (such as the hippocampus), where episodic connections based on any relevant data are formed, and then sent back as input to the semantic network. In other words, episodic connections might actually involve indirect routes. Indeed, former studies show that the hippocampus might be involved in such higher order relational learning (e.g., Howard, Fotedar, Datey, & Hasselmo, 2005) and some evidence for the plausibility of a readout mechanism for synonyms and antonyms has also been provided (Chen, Lu, & Holyoak, 2010). It may further be noted that when episodic effects are considered, other, nonsemantic aspects of the stimuli can also, in principle, affect the expectancy mechanism. For example, Hutchison (2007) showed that subjects could control whether or not they applied expectancy, on a single trial basis, according to the color of the prime, which was predictive of the target relatedness. Such result can be incorporated within our framework assuming that noise modulations are affected by contingencies in the physical properties of the stimuli (e.g., color) in addition to their sensitivity to RTs. If, for example, a certain color becomes associated with lower RTs when transitions are attempted, the system may learn to apply a large noise (leading to transitions) each time the prime is presented in that color, but not when the prime is in a different color. In other words, the system will learn to expect a related target only when the prime is in the right color, leading to the observed experimental findings. However, as this mechanism is almost completely within the realms of episodic learning (rather than the interaction between episodic and semantic processes), we do not develop it further in the current article.

Another important effect that was not addressed by the current model is inhibition in priming. Indeed, in addition to the facilitating effects of related primes that the present study has focused on, unrelated primes sometimes lead to slower recognition of targets compared to neutral primes, demonstrating inhibition. Since in most cases the inhibitory effects are unique to LDT and do not appear in pronunciation, they are often considered

to be the product of decision-making processes external to the semantic system (Neely, 1991). Indeed, several other findings unique to semantic priming in LDT, such as the nonword facilitation effect (Neely, 1977), are also best accounted for by speculating the existence of an independent decision-making module (e.g., Neely et al., 1989). However, some studies clearly show that under certain conditions, at least part of the inhibition effects are a direct result of wrong expectancies (e.g., Neely,1977).

As it turns out, expectancy-related inhibition can be achieved in our model, assuming that the connections from the semantic to the lexical layer are strengthened after a transition from the prime pattern (a transition which, in effect, represents the prediction). Since in typical semantic priming tasks there is no explicit need to use top-down information for accurate performance, these connections were set to be subthreshold in our model, affecting the dynamics only when combined with a corresponding bottom-up input arriving to the lexical network (see Lerner et al., 2012a; for details). However, if, after a transition, these connections are strengthened beyond a certain value, they will, independently of external input, press the lexical network to converge to the word pattern corresponding to the concept pattern which the semantic network has jumped to. In related trials, this semantic transition is usually to a pattern related to the upcoming target (or even the target itself); therefore, the additional pressure will mostly strengthen the typical facilitation effects of these trials. In unrelated trials, however, the transition is to an unrelated pattern and so the additional pressure will delay convergence to the target (compared to neutral trials in which semantic-to-lexical connections do not have an effect; see Lerner et al., 2012a), thus demonstrating inhibition. Indeed, such expectancy-induced enhancement of the semantic feedback to the lexical layer has been suggested independently by several authors to account for other priming-related phenomena (e.g., Brown, Stolz, & Besner, 2006; Robidoux, Stolz, & Besner, 2010; Stolz & Neely, 1995). Future developments of the model could examine whether such strengthening can be learned by a mechanism resembling the one used to adjust the noise in the present model, and clarify what are the experimental conditions which encourage such a strategy (see Lerner & Shriki, 2014, for more information).

6.2. A new perspective on automaticity

As a final note, returning to the automatic/controlled dichotomy, it is important to emphasize that the current model, which combines autonomous neural-network mechanisms with "controlled" manipulations of network parameters, may suggest a fundamental distinction between automatic and nonautomatic processes. As we have suggested in the present work, semantic transitions may be sensitive to controlled processes. In addition, as briefly described above based on previous studies (e.g., Stolz & Neely, 1995), the connectivity strength between the semantic and lexical network may also be intentionally modulated, depending on whether bottom-up or top-down processes are emphasized in the task. In other words, the underlying principle of the current model is that every process which involves a change in the network state could potentially be under cognitive control. However, correlations between stored patterns are an inherent trait of the

network's connectivity structure and therefore when one pattern is activated, its correlated patterns are also activated to some degree and this simultaneous semi-activation cannot be prevented. The only way that such co-activation could be changed is by modifying the representations themselves (a lengthy procedure which includes unlearning the present representations and relearning new ones). A good example of this attribute is the priming effect of directly versus indirectly related pairs: Directly related pairs have correlated representations and therefore elicit automatic priming which cannot be prevented as long as the prime concept is accessed. Indirectly related pairs, on the other hand, require a transition in the semantic network. This transition can be avoided (as we hypothesized, for example, in LDT) and, therefore, such priming is not automatic per se (see also, Jones, 2012). Hence, the fundamental distinction between automatic and nonautomatic processes, which our model implies, is the distinction between processes that depend on overlap of representations and processes that depend on transitions between representations. Encoding representations in a correlated manner inevitably leads to their automatic dependency on each other. Processes that involve a change from one active representation to another are not necessarily automatic and can be affected by controlled manipulations. This distinction echoes, to a certain extent, a well-known proposal by Logan (1988) stating that automaticity depends on the accumulation of domain-specific knowledge that enables replacing long algorithmic computations with quick retrieval of relevant exemplars. However, in our view, it is not simply the accumulation of knowledge per se that makes the difference but, rather, its organization as representations with a correlative structure. Such a perspective can be related to the classic attributes of automaticity described by Shiffrin and Schneider (1977): Automatic processes are fast (in our model, they do not require a change in the network activity state, such as a transition from one pattern to another, as they rely on correlations), do not suffer from capacity limitations and therefore run in parallel (an activated pattern in the model is simultaneously activating all of its correlated patterns), are unavoidable (a correlation-based encoding necessarily means a pattern cannot be activated without the partial activation of its correlated patterns), and inflexible (the only way to change the mutual dependency between correlated concepts is by relearning their representations). Whether this formulation of automaticity can be broadened to include other, nonsemantic procedures, remains to be discussed elsewhere.

7. Conclusion

The current work extended our previous network model of automatic semantic priming (Lerner et al., 2012a) to capture some of the most important traits characterizing controlled processes involved in this cognitive phenomenon. Assuming that latching dynamics in the semantic network can be controlled by noise adjustments, we have shown how different types of priming effects (semantic, associative, mediated, backward) are distinctively influenced by these adjustments, and how different modulators of the task such as SOA, prime–target relations, and RP yield separate optimization solutions which fit known experimental results. Our approach allows for automatic and controlled priming

effects to be interpreted within one framework and shows how the interplay between them can be understood in neural network terms. Although the current work did not address some important aspects of controlled priming such as inhibition and the formation of complex expectancies, we have sketched how these effects may be integrated with our approach in future extensions of the model. Our model also provides several testable predictions regarding backward and mediated priming which may be examined in future semantic priming experiments. Finally, based on the mechanisms of our network, we suggested a novel way of defining automatic versus nonautomatic processes that may be further expanded to include other cognitive domains.

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. The reasons why these parameters modulate the rate of transitions have to do with the dynamical properties of attractor networks and are beyond the scope of the current article.
2. Although choosing the noise values before and after a transition to be the parameters of control may seem too simplified compared to the general problem (where noise is modulated at each instant of time during a trial), this choice, in fact, leads to approximately the same results as the general problem. The reason is that noise, as long as it is not extreme (as in our model), affects the network dynamics mainly at specific points in time when the network is ready to move from one attractor to another (due to the adaptation mechanisms reaching a critical level, making the attractor unstable). At these critical phases, a sufficiently high level of noise will allow a transition to occur while a low level of noise will delay it. At other times, when the network is safely converged to an attractor that is not yet unstabilized by adaptation, noise does not have much effect as the nature of attractor dynamics acts to resist interference and keep the network state static. Moreover, after a transition has occurred, the adaptation mechanisms are "reset," so to speak, because new units will now be active, and so the next noise value to have an effect would be the one just before the network is ready to make another transition. Consequently, the noise values that really make a difference are the ones before each transition, meaning that only a few control parameters—one before each potential transition—can significantly change behavior. Moreover, in separate simulations we have performed (data not shown), adding another control parameter, after the second

transition and before the third, did not result in further modulation of performance as the SOAs used in the simulations do not allow enough time for the network to frequently complete more than two transitions; therefore, the region in the parameter phase space where more than two transitions occur is not sufficiently explored by the learning mechanism and is thus irrelevant. Acknowledging these facts, optimizing just two noise values—one before the first transition and one after it (i.e., before the potential second transition) represents a faithful approximation to the general optimization problem.

3. The semantic structure in our model determines the strength of correlations between pairs and the types of associative transitions occurring in the network (see Lerner et al., 2012a,b), which, in turn, determine the way priming is affected by the learning mechanism. It does not, however, affect results other than through its role in setting the correlations and transitions. Therefore, to study the effect of learning on priming, it is necessary to compare pairs with different correlative and transitional properties rather than compare different semantic structures. Using an alternative semantic structure with stimuli pairs that have the same correlative strength and transition probabilities as the one studied here would thus yield the same results. Indeed, running the learning mechanism on the various semantic structures used by Lerner et al. (2012a,b) yielded equivalent pattern of findings to those of the current study when pairs with similar correlations and associative transition probabilities were compared (data not shown).
4. The noise values were not allowed to become negative. If the learning rule drove any of the values below 0, the value was clipped to 0 instead.
5. Values from human experiments are presented only in the case where conditions approximately matching the current simulations could be found. To the best of our knowledge, there are no published experiments using typical LDT and mediated-only lists at long SOAs or pronunciation tasks using backward-only lists at any SOA.

References

Balota, D. A., Black, S. R., & Cheney, M. (1992). Automatic and attentional priming in young and older adults: Reevaluation of the two-process model. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 485–502.

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.

Becker, C. A. (1976). Allocation of attention during visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 556–566.

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

Besner, D. (2001). The myth of ballistic processing: Evidence from Stroop's paradigm. *Psychonomic Bulletin & Review*, *8*, 324–330.

Bodner, G. E., & Masson, M. E. (2003). Beyond spreading activation: An influence of relatedness proportion on masked semantic priming. *Psychonomic Bulletin and Review*, *10*, 645–652.

Braver, T. S., Barch, D. M., & Cohen, J. D. (1999). Cognition and control in schizophrenia: A computational model on dopamine and prefrontal function. *Biological Psychiatry*, *46*, 312–328.

Brown, M., Stolz, J. A., & Besner, D. (2006). Dissociative effects of stimulus quality on semantic and morphological contexts in visual word recognition. *Canadian Journal of Experimental Psychology*, *60*, 190–199.

Chen, D., Lu, H., & Holyoak, K. J. (2010). Learning and generalization of abstract semantic relations: Preliminary investigation of bayesian approaches. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32rd Annual Conference of the Cognitive Science Society* (pp. 871–876). Austin, TX: Cognitive Science Society.

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

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

Eisenberg, P., & Becker, C. A. (1982). Semantic context effects in visual word recognition, sentence processing, and reading: Evidence for semantic strategies. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 739–756.

Feldman, L. B., & Basnight-Brown, D. M. (2008). List context fosters semantic processing: Parallels between semantic and morphological facilitation when primes are forward masked. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *34*, 680–687.

de Groot, A. M. B. (1983). The range of automatic spreading activation in word priming. *Journal of Verbal learning and Verbal Behavior*, *22*, 417–436.

Haider, B., Hausser, M., & Carandini, M. (2013). Inhibition dominates sensory responses in the awake cortex. *Nature*, *493*, 97–100.

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.

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

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

Hutchison, K. A. (2007). Attentional control and the relatedness proportion effect in semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 645–662.

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *Quarterly Journal of Experimental Psychology*, *61*, 1036–1066.

Hutchison, K. A., Neely, J. H., & Johnson, J. D. (2001). With great expectations, can two "wrongs" prime a "right"? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *27*, 1451–1463.

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

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.

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

Lerner, I., Bentin, S., & Shriki, O. (2012a). Spreading activation in an attractor network with latching dynamics: Automatic semantic priming revisited. *Cognitive Science*, *36*, 1339–1382.

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

Lerner, I., & Shriki, O. (2014). Internally- and externally-driven network transitions as a basis for automatic and strategic processes in semantic priming: Theory and experimental validation. *Frontiers in Psychology*, *5* 314. doi:10.3389/fpsyg.2014.00314

Logan, G. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.

Lorch, R. F., Balota, D. A., & Stamm, E. G. (1986). Locus of inhibition effects in the priming of lexical decisions: Pre-or postlexical access? *Memory & Cognition*, *14*, 95–103.

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

McKoon, G., & Ratcliff, R. (1995). Conceptual combinations and relational contexts in free association and in priming in lexical decision and naming. *Psychonomic Bulletin & Review*, *2*, 527–533.

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

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

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.

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.

Neely, J. H., & Keefe, D. E. (1989). Semantic context effects on visual word processing: A hybrid prospective/retrospective processing theory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory*. Vol. 24 (pp. 207–248). New York: Academic Press.

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

Neely, J. H., O'Connor, P. A., & Calabrese, G. (2010). Fast trial pacing in a lexical decision task reveals a decay of automatic semantic activation. *Acta Psychologica*, *133*, 127–136.

Neumann, O. (1984). Automatic processing: A review of recent findings and a plea for an old theory. In W. Prinz & A. F. Sanders (Eds.), *Cognition and motor processes*. Berlin: Springer-Verlag.

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). Mahwah, NJ: Erlbaum.

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.

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information Processing and Cognition: The Loyola symposium* (pp. 55–85). Hillsdale, NJ: Erlbaum.

Robidoux, S., Stolz, J., & Besner, D. (2010). Visual word recognition: Evidence for global and local control over semantic feedback. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 689–703.

Rose, M., Haider, H., Weiller, C., & Buchel, C. (2002). The role of medial temporal lobe structures in implicit learning. An event-related FMRI study. *Neuron*, *36*, 1221–1231.

Servan-Schreiber, D., Printz, H. W., & Cohen, J. D. (1990). A network model of catecholamine effects: Gain, signal-to-noise ratio and behavior. *Science*, *249*, 892–895.

Shelton, J. R., & Martin, R. C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1191–1210.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, *84*, 127–190.

Stolz, J. A., & Besner, D. (1996). Role of set in visual word recognition: Activation and activation blocking as nonautomatic processes. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 1166–1177.

Stolz, J. A., & Besner, D. (1999). On the myth of automatic semantic activation in reading. *Current Directions in Psychological Science*, *8*, 61–65.

Stolz, J. A., & Neely, J. H. (1995). When target degradation does and does not enhance semantic context effects in word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 596–611.

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

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

Winocur, G., Moscovitch, M., & Sekeres, M. (2007). Memory consolidation or transformation: Context manipulation and hippocampal representations of memory. *Nature Neuroscience*, *10*, 555–557.

Appendix

This appendix provides the main equations governing the network dynamics and the specific parameter values that were used in the numerical simulations of the model. Please refer to Lerner et al. (2012a,b) for more details. Units are indicated in brackets whenever relevant. In all numeric simulations, the time step represented $\Delta t = 0.66$ ms.

a. The activity of the i-th neuron at time t, $x_i(t)$, was a logistic function of its local input $h_i(t)$ which obeyed:

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

With η_i being the noise term. The synaptic depression of the connection weight between the i-th and j-th neuron obeyed:

$$J_{ij}(t) = \frac{J_{ij}^{\max} - J_{ij}(t)}{\tau_r} - U x_{\max} x_i(t) J_{ij}(t)$$

With $J^{\max}$ being the Hopfield connectivity matrix for sparse patterns (Tsodyks, 1990):

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

The temporal correlations in the noise were generated by filtering the noise using a low-pass filter, which, for two time points separated by τ ms, took the form:

$$f(\tau) = \eta_{\text{amp}} \cdot e^{\frac{\tau}{\tau_{\text{corr}}}}$$

b. The semantic and lexical network parameters

Parameter	Semantic Network	Lexical Network
Number of neurons, N	500	500
Sparseness, p	0.06	0.04
Correlation strength (% of overlapping active neurons out of total active neurons in a pattern)	0.1 (Strong) 0.066 (Moderate) 0.033 (Weak)	0
Neuronal gain, T	0.05	0.05
Neuron's time constant, τ_n	7 [ms]	13 [ms]
Neuronal activation threshold, θ	0.02	0.17
Regulation parameter, λ	14.75	27.75
Maximal firing rate, x_{max}	100 [spks/s]	100 [spks/s]
Utilization of synapses within each network, $U_{[within]}$	0.206 [1/spks]	0 [1/spks]
Utilization of synapses between networks, $U_{[between]}$	Lexical to Semantic: 0.087 [1/spks]	Semantic to Lexical: 0 [1/spks]
Synaptic recovery time within each network, $\tau_{r\ [within]}$	93 [ms]	—
Synaptic recovery time between networks, $\tau_{r\ [between]}$	Lexical to Semantic: 1333 [ms]	Semantic to Lexical: —
Input gain between networks (Raw values. Actual values were normalized by the number of pre-synaptic active neurons in a pattern)	Lexical to Semantic: 2	Semantic to Lexical: 0.21
External input gain	—	0.56
Input threshold, θ_{ext}	1	0.25
Default noise amplitude, $\eta_{amp}(0)$ (For both initial and late noise)	Pronunciation-like: 0.05 LDT-like: 0.01	0.025
Noise temporal correlations, τ_{corr}	17 [ms]	17 [ms]
Multiplication factor of the exploration parameter, A	0.06	—
Exponential coefficient of the exploration parameter, β	0.0125	—
Learning rate, α	0.002	—
Convergence threshold	0.95	0.95