Attention & Performance 2002 meeting

Final version, 2/26/2003 18:02 PM

The neural bases of subliminal priming

Stanislas Dehaene

INSERM Unit 562

Service Hospitalier Frédéric Joliot, CEA/DRM/DSV

4 Place du Général Leclerc

91401 Orsay cedex, France

dehaene@shfj.cea.fr

Phone: +33 1 69 86 78 73 Fax: +33 1 69 86 78 16

Abstract

Psychologists have long reported that words that are made invisible by forward and backward masking can nevertheless cause behavioral priming effects. Functional neuroimaging can now be used to explore the neural bases of masked priming. Subliminal priming causes reduced activation in multiple areas (fusiform gyrus, intraparietal sulcus, and motor cortex), in direct correspondence with behavioral manifestations of priming at the orthographic, semantic, and motor level. This implies that a whole stream of processors can operate unconsciously. The neural code in each area can be assessed by varying prime-target relations. A simple mathematical framework is proposed that tentatively relates priming at the voxel level with the shape of the tuning curves of single neurons in the underlying tissue. Priming thus provides a general method to study the fine microcode in each brain region (the 'priming method').

Introduction

Visual masking refers to the reduction of the visibility of a visual stimulus when it is preceded or followed, in close spatial and temporal proximity, by the presentation of another visual stimulus called the 'mask'. Many different sorts of masking exist (Enns & Di Lollo, 2000). In this chapter, I concentrate on pattern masking, a technique which has been used extensively in psychology to study the levels of representation of words. Under appropriate conditions, words that are followed by a masking pattern can be rendered invisible, even to trained observers. Yet, behavioral priming experiments have repeatedly indicated that the masked words are nevertheless processed unconsciously. Behavioral manifestations of masked priming have been observed at various levels including the orthographic, the phonological, and possibly the semantic level (see e.g. Bowers, Vigliocco, & Haan, 1998; Cheesman & Merikle, 1984; Ferrand & Grainger, 1994; Forster & Davis, 1984; Greenwald, Draine, & Abrams, 1996; Marcel, 1983; Neely & Kahan, 2001). Thus, masking appears as a useful paradigm to study two central questions in cognitive neuroscience: the various levels of representations of words, and the nature of the differences between conscious and unconscious processes.

In this chapter, I review a number of studies that have used masked priming to examine the neural activity evoked by subliminal words. I first examine which brain areas are affected by masked priming effects, and what this tells us about the level of coding of words in those areas. I then propose a mathematical formalism to explain how priming at the single-cell level can translate into macroscopic changes in fMRI activation. Finally, I discuss what changes in brain activation patterns may occur when a word crosses the threshold of consciousness and becomes visible.

1. General logic of subliminal priming studies

In a typical subliminal priming experiment (figure 1), each trial consists in the consecutive presentation, at the same screen location, of a random configuration of symbols or geometrical shapes (pre-mask), a first word (the prime), another random configuration of symbols or shapes (post-mask), and a second word (the target). The prime word is presented very briefly, typically 10-50 ms. The target is presented for a much longer duration, typically 500 ms. Finally, the stimulus onset asynchrony between the prime and target is short (typically 60-120 ms). Under those conditions, subjects report seeing only the masks and the target word, but not the prime.

This subjective invisibility can be confirmed by asking subjects to perform an explicit task on the primes. In recent experiments, my colleagues and I have used a two-alternatives forced-choice identification task in which subjects have to select, amongst two alternative words, the one that matches the prime. As illustrated in figure 1, with a 29-ms prime presentation duration, performance typically does not differ from the chance level of 50% success (Dehaene et al., 2001). Other tasks such as prime presence-absence jugdment (Dehaene et al., 1998b) or prime categorization (Naccache & Dehaene, 2001b) yield similar results.

In spite of their subjective invisibility, the prime words impact on the processing of target words. This can be demonstrated by varying the relation between the prime and target. The simplest design compares trials in which the same word is presented twice, as both prime and target, with trials in which different words are presented as prime and target (figure 1). In this condition of repetition priming, response times to the target are consistently shorter on repeated than on non-repeated trials (e.g. Forster, 1998; Forster & Davis, 1984).

By systematically varying the physical, phonological, or even semantic proximity between the prime and target, it is then possible to probe which levels of word processing underlie this

facilitation effect. As described below, in combination with brain imaging, this becomes a powerful method to probe the brain areas that are traversed by the wave of activation induced by subliminal primes, and the level of representation associated with each of them.

2. Subliminal priming in the visual word form system

Dehaene et al. (2001) first examined the cerebral bases of subliminal word repetition priming. FMRI data were collected in a fast event-related paradigm while subjects performed a bimanual semantic classification task on visual words. Unbeknownst to them, each target word was preceded by a subliminal prime (figure 1). I used a 2x2 design in which the prime and target could be the same word or different words, and could appear in the same or different case, thus defining four types of events. The amount of activation in each of those events was identified relative to a fifth « null » event in which only the masks were presented and no response was required.

The behavioral results showed that response times were faster on repeated trials, whether or not the words shared the same case (figure 2). In searching for the cerebral bases of this effect, two distinct types of brain regions were identified. The right extrastriate occipital cortex showed repetition suppression only for physically identical primes and targets, suggesting a role for right visual areas in coding the precise visual features of the letters (Marsolek, Kosslyn, & Squire, 1992). The left fusiform gyrus, however, showed repetition suppression whenever the same word was repeated, whether in the same case or not (figure 2). Thus, this region appears to encode the word string in a case-independent fashion.

This region may provide the cerebral substrate of the visual word form system (Warrington & Shallice, 1980), a structural representation of visual words as an ordered sequence of abstract letter identities or multi-letter graphemes, invariant for size, font, and case (Cohen et al., 2000; Cohen et al., 2002; Dehaene, LeClec'H, Poline, LeBihan, & Cohen, 2002). It is roughly symmetrical to the right-hemispheric fusiform face area, and may play for

visual word recognition the same role that similar or neighboring regions of the fusiform and lingual gyri play for other visual objects such as faces, objects, or places (Haxby, this volume). In adults, it has become partially attuned to a specific script, as shown by its greater response to real words than to consonant strings of similar arbitrary shape (Cohen et al., 2002). Indeed, in children this area activates in direct proportion to the child's reading skills, and its response is absent in dyslexic readers who have not developed expertise in word recognition (Paulesu et al., 2001; Shaywitz et al., 2002; Shaywitz et al., 1998).

In the rest of this paper, I refer to this area of left fusiform activation during word reading as the visual word form area (VWFA). It should be clear that this label does not imply that this area responds only to words, or even maximally to words as opposed to other categories of visual stimuli. Indeed, specificity for words is unlikely given that reading is a recent cultural invention that appears to 'recycle' neural tissue initially engaged in broader objet recognition functions. The issue of category specificity is not addressed by my research, which is rather aimed at understanding what type of code for visual words is present in this region.

In order to further specify the exact nature of the word representation attained by subliminal primes in the VWFA, I recently performed two more word repetition priming experiments, whose results can only be briefly reviewed here (Dehaene et al., in preparation). The first examined whether visual features alone could explain priming in this region. To this end, the visual similarity of upper and lowercase letters was manipulated (figure 2). Half of the prime-target pairs comprised words made of letters that are highly similar in upper and lower case (e.g. Oo, Pp). The other half used only highly dissimilar letters (e.g. Aa, Gg) for which the uppercase-lowercase association is essentially arbitrary. The same strip of left fusiform cortex as in experiment 1 showed replicable subliminal repetition suppression that was present even when the letters were visually dissimilar (similar to what was observed in

response times ; see figure 2) . This confirmed that this region is not solely concerned with visual shapes, but encodes letter strings using a culturally acquired abstract letter-identity code.

A second experiment examined whether single letters or larger units such as graphemes or whole words are encoded in the VWFA. In order to repeat letters without repeating words, anagrams were used. For instance, by priming the French target word « REFLET » with the prime « trefle », almost all of the middle letters (r, e, f, l, e) could be repeated. By moving the prime one letter position relative to the target (e.g. « trefle_ followed by _REFLET, it was even possible to repeat those letters at the same screen location without repeating the same word. By comparing this to a word-repeated trial, with or without a shift in letter position, the nature and position invariance of the neural codes underlying priming could be tested. Would priming depend on letter repetition, word repetition, or both ? The results revealed an interesting dissociation between posterior and anterior areas. The posterior portion of the VWFA (y = -68 in the Talairach coordinate system) showed repetition suppression only when the same letters were repeated at the same location, regardless of their case. This region thus holds a case-invariant but position-selective letter code. More anteriorily (y=-56), location-independent priming was found for both repeated words and anagrams compared to a control, non-repeated condition (figure 2). Thus, this region encodes a case- and position-invariant representation of visual units that are smaller that the whole word. Finally, in a still more anterior fusiform region (y=-48), priming became greater for repeated words than for anagrams, thus revealing a case- and position-independent whole-word code, or at least a code sensitive to the larger graphemic units that distinguish a word from its anagram. Behavioral response times seem to be sensitive only to this whole-word code (figure 2).

Two conclusions may be drawn from those studies at the visual word form level. First, behavioral priming effects provide only a coarse indication of the levels of representations traversed by a subliminal prime. Functional imaging reveals a much richer variety of priming effects, ranging from feature-based to letter or whole-word based priming. Second, subliminal primes can be processed quite far along the ventral visual identification pathway. Although several studies have identified a tight correlation between ventral fusiform activity and the contents of visual consciousness (e.g. Bar et al., 2001; Grill-Spector, Kushnir, Hendler, & Malach, 2000), fusiform activity is not sufficient for conscious reportability. Exactly what more is needed will be addressed in the final discussion.

3. Subliminal priming at the motor level

How far beyond early vision can a subliminal word travel without entering into consciousness? To address this issue, Lionel Naccache and I studied another priming paradigm in which semantic and motor components of priming could be assessed (figure 3). Subjects viewed number words or digits that they had to compare with 5. “Larger” or “smaller” responses were made by depressing a left-hand or right-hand button, with response assignment changing in the middle of the experiment. Unbeknownst to the subjects, each target number was preceded by a numerical prime. Various prime-target relations could be tested with this paradigm (figure 3): the notation of the prime and target could be the same or different; they could represent the same or different quantity; and finally they could yield the same motor response (both larger or both smaller than 5) or a different motor response (one larger and the other smaller).

Dehaene et al. (1998b) showed that the latter factor had a measurable impact on both behavior and brain activation. Behaviorally, subjects were faster by 24 ms when the prime-induced and target-induced responses were congruent than when they differed. At the brain level, measures of motor activation obtained with both fMRI and ERPs revealed activation of

motor cortices on the side that would have been appropriate for responding to the prime. This was revealed by computing a lateralized response index based on the difference in motor activity from the hemisphere involved in programming the correct target-induced response and from the other hemisphere (this measure is called the "lateralized readiness potential" (LRP) in the ERP literature, and we termed it the "lateralized BOLD response" (LBR) for fMRI). This index was larger on motorically congruent than on incongruent trials. With the higher temporal resolution afforded by ERPs, it was possible to identify two successive stages, an early one in which the LRP was only induced by the subliminal prime and a later one in which this small motor bias was overcome by the much larger motor response induced by the supraliminal target. In summary, the activation induced by a masked prime can travel all the way to the regions involved in programming a motor response. Similar results have been obtained using simpler visual stimuli such as arrow shapes (Eimer & Schlaghecken, 1998; Neumann & Klotz, 1994; Schmidt, 2002; see also Jaskowski et al, 2002).

A crucial aspect of those finding is that motor priming is obtained although the assignment of responses to the left and right hands is arbitrary and is changed during the experiment. This implies that subliminal primes are processed along neural pathways that are temporarily established to comply with experimental instructions. Naccache, Blandin and Dehaene (2002) further confirmed the effect of top-down influences on subliminal priming by manipulating subjects' temporal attention in the number-priming paradigm. Across three experiments, the same prime-target pairs were presented in contexts in which their onset was either predictable or unpredictable. In experiment 1, temporal predictability was manipulated either by presenting trials at a fixed or variable lag with respect to trial onset. In experiment 2, visual cues were presented or omitted prior to the prime-target pair. Finally in experiment 3, valid or invalid verbal cues specified when the prime-target pair was likely to appear. In all three situations, priming was found to depend critically on the deployment of temporal

attention. Behavioral priming was reproducibly observed whenever subjects' attention could be focussed on the prime-target pair. However, priming disappeared when the prime-target pair appeared at an unexpected time. Both motor and repetition priming were eliminated.

Those results can be interpreted as showing that the focusing of temporal attention on the target plays a permissive role that benefits the processing of a temporally contiguous prime. In the absence of such attentional amplification, the processing of subliminal primes is considerably reduced or eliminated. Thus, subliminal priming is not independent of attention. This challenges the classical view of subliminal priming, based on automatic spreading activation theory (Neely, 1991), according to which the unconscious processing of primes is passive and automatic (Eysenck, 1984; Posner & Snyder, 1975; Schneider & Shiffrin, 1977). If a processing pathway is prepared by top-down attention and intentions, it can be applied automatically to subliminal stimuli (Dehaene & Naccache, 2001; Dagenbach, Carr & Wilhemsen, 1989; Hommel, 2000). However, if the organism is not prepared to process them, subliminal primes induce little or no spreading of activation.

4. Can subliminal priming occur at the semantic level?

Our finding of motor priming in the number comparison task was initially interpreted as a clear, though indirect, proof of *semantic* processing of subliminal number symbols. That subjects could activate the motor cortex of the hand that would have been appropriate for responding to the prime seemed to imply that subjects had unconsciously categorized the prime as larger or smaller than 5, thus implying semantic access.

However, an alternative interpretation was that the observed motor activation was due to direct motor specification (Neumann & Klotz, 1994). Because a very small number of stimuli -- the digits 1, 4, 6, 9 and the corresponding words – were repeatedly used as both primes and targets, subjects could have learned to associate each visual stimulus with the corresponding response, thus bypassing semantic access. Use of such a direct visuo-motor

route was recently demonstrated by Abrams and Greenwald (2000). In an affective categorization task, they showed that new primes that were made of fragments of previous seen targets yielded subliminal motor priming solely based on visual fragments, not on whole-word meaning. For instance the prime word SMILE, created from the targets words SMUT and BILE, ended up paradoxically priming the negative rather than the positive response. Thus, although the task required semantic categorization, and although a priming effect was observed, the primes only received a shallow, non-semantic analysis of their component letters and the associated motor responses (see also Damian, 2001).

Fortunately, new experiments and reanalyses have now demonstrated that the number priming results do not fall prey to a similar non-semantic interpretation. We have now replicated the original behavioral experiment with novel numbers that are only presented as primes, never as targets (Naccache & Dehaene, 2001b). Because those numbers are never seen consciously and are never responded to, they cannot be associated with motor responses. Yet in two different experiments, those novel primes were found to cause significant motor priming, indicating that at least part of the motor priming effect arises from a genuinely semantic route. This positive effect of novel numerical primes has now been replicated and extended by others (Greenwald, Abrams, Naccache, & Dehaene, 2002; Reynvoet, Caessens, & Brysbaert, 2002), though I am aware of one failure to replicate (Sid Kouider, unpublished PhD thesis).

Further analyses also demonstrated that the motor priming effect was present in the first block trial and that both motor priming and the classical semantic distance effect did not change with practice, which is inconsistent with the idea that the task is increasingly being performed using a non-semantic route.

Most crucially, another priming effect was identified that could only be interpreted at the semantic level. This effect was termed “quantity priming” (figure 3). Subjects were faster

on repeated trials (e.g. prime 9, target 9) than on congruent non-repeated trials (e.g. prime 6, target 9). Because in both cases the response induced by the prime is the same and is congruent with the target, motor priming cannot contribute to this effect. The effect is found with the same magnitude even when the notation is changed (e.g. prime NINE, target 9), suggesting that it occurs at an abstract level. Furthermore, behavioral experiments have demonstrated that the amount of quantity priming varies monotonically with the semantic distance between the prime and the target (e.g. 9-9 versus 8-9, 7-9 or 6-9), thus confirming that it originates at the level of quantity coding (Koechlin, Naccache, Block, & Dehaene, 1999; Reynvoet & Brysbaert, 1999).

Finally, at the neural level, fMRI demonstrated that quantity priming was associated with repetition suppression in the left and right intraparietal sulci (Naccache & Dehaene, 2001a). Those regions are consistently activated in a variety of number processing and calculation tasks, whenever subjects have to manipulate the quantity associated with numerical symbols (e.g. Dehaene, Spelke, Stanescu, Pinel, & Tsivkin, 1999). They are thought to represent numerical quantities on an internal continuum analogous to a mental 'number line'. The fact that their activation is reduced when the same quantity is repeated twice, possibly in different notations, confirms that they encode particular numerical quantities and can be activated unconsciously.

In summary, priming effects with subliminal numbers have been observed at both the semantic (intraparietal) and motor levels. This provides a clear indication that semantic-level processing of masked primes is possible. It should be noted that digits are some of the most frequent visual symbols and are semantically unambiguous. The ease and speed of visual-to-semantic transduction may explain why it seems easier to obtain semantic priming with numbers than with other types of words (Abrams & Greenwald, 2000; Damian, 2001).

5. The "priming method": priming as a neuroimaging tool

The above examples illustrate how priming can be used in combination with neuroimaging to separate out distinct stages of the cortical coding of words. In occipital cortex, priming depends on repeating the same marks at the same location on the retina, suggesting a position-specific feature-level code. In fusiform cortex, the code for words is increasingly more abstract, as priming can be obtained when repeating the same letters in a different case, and sometimes even at a different location. In intraparietal cortex, the code is even more elaborate, as priming can be elicited by repeating the same numerical quantity using different symbols such as TWO and 2.

Inferences based on neuroimaging observations of priming are powerful because they indirectly reveal both the **precision** and the **abstraction** of a neural code. When repetition suppression is observed in the intraparietal area for, say, 9 followed by 9 as opposed to 6 followed by 9, this implies that the stimuli 9 and 6 are distinguished within this area: there are partially distinct populations of neurons encoding those two stimuli. I refer to such stimulus-selectivity as the "precision" of the neural code. On the other hand, when the amount of repetition suppression is found not to differ for, say, NINE followed by 9 as opposed to 9 followed by 9, this implies that the neurons in this area recognize the similarity between the stimuli NINE and 9. Thus, there are populations of neurons that abstract away from the notation used to convey a number. I refer to this as the degree of "abstraction" of the neural code.

Neither precision nor abstraction can be inferred on the basic of other neuroimaging designs such as the subtraction, parametric, or conjunction methods. For instance, direct subtraction of the brain activations to the digits 6 and 9 would most likely result in an absence of detectable differences at the scale of resolution typical of most fMRI studies (2-4 mm). Furthermore, although a statistical conjunction or masking operation can be used to

demonstrate that the word SIX and the digit 6 cause similar patterns of activation in the intraparietal cortex, this would not prove that they are being coded by the same neurons. At present, only the priming method affords such non-invasive inferences about the neural microcode.

Use of the priming method does not require that the primes be subliminal, or that primes and targets appear in close succession. Indeed, a very similar approach, often called the fMRI adaptation method, was initially applied to both blocked and event-related studies of mid- to long-term stimulus repetition, where it yielded detailed information on the coding of words and objects (e.g. Buckner et al., 2000; Grill-Spector et al., 1999; Kourtzi & Kanwisher, 2000; Vuilleumier et al., 2002; Wagner et al., 2000). The use of subliminal primes, however, is advantageous because it ensures that the results are not contaminated by attentional or strategic biases. For instance, novel stimuli may be more interesting or attention-grabbing than old ones. Conversely, reduced fMRI activation on consciously repeated trials may be due to lesser attention.

Stimulus repetition may also result in large-scale task reorganization. For instance, subjects may merely repeat a previous response instead of computing it *de novo*, resulting in large-scale changes in brain activation on repeated trials (Raichle et al., 1994). Crucially, such attentional or strategic activation changes may occur in brain regions distant from those involved in the initial recognition that the stimulus was repeated, and hence may no longer be informative about the local neural microcode. Subliminal primes effectively prevent such strategies.

Finally, subliminal primes have only a very temporary effect, typically lasting less than 500 milliseconds, and so effectively act as a **tracer** of existing representations. As will be argued further below, the priming method is effective under the assumption that neuronal tuning curves are not being changed by exposure to the primes. Consciously visible primes

may lead to long-lasting learning-induced changes that no longer warrant the inference that repetition suppression reflect the pre-existing code in a given area.

A clear drawback of subliminal primes, however, is that they result in small and often hard-to-detect activation. The effects reported earlier were very small and required averaging across subjects. To study the code in higher-level areas, where prime-induced activation barely penetrates, conscious priming may be the only possibility. Still, it is desirable to prevent subject's awareness of repetitions, especially since it has been observed that attention to repetition may cause disrupt the repetition suppression phenomenon (Henson et al., 2002).

6. A mathematical formalism linking priming at the cellular and neuroimaging levels

I now elaborate a minimal mathematical formalism of the priming method, in order to formulate more precisely its predictions according to the nature of the experimental design. Consider the brain-imaging signal $I(s)$ measured, for instance with fMRI, in response to a single stimulus s. In any voxel of a few cubic millimeters, this signal typically reflects the summed activation of several million neurons, and can thus be written as :

[1] $$I(s) = h(\sum_i f_i(s))$$

where $f_i(s)$ is the firing rate of neuron i in response to stimulus s, and h represents the hemodynamic response function which transforms firing rates into measurable changes in blood flow and oxygenation. (For simplicity, I neglect here the time dimension of brain-imaging signals ; a more complete treatment would treat h as a temporal convolution operator).

Firing rate $f_i(s)$ depends on a measure of the distance between the actual stimulus s and the neuron's preferred stimulus s_i^{pref} . This can be written as

[2] $f_i(s) = F(\|s - s_i^{pref}\|)$

where $\|...\|$ is a distance metric, and F is a monotonically decreasing function (possibly a Gaussian) specifying how firing rate varies with proximity in similarity space to the preferred stimulus. The metric $\|...\|$ characterizes the neurons' tuning curves and how they weight different dimensions of stimulus variation. It is expected to vary considerably between different areas. Many neurons in visual area V1, for instance, prefer a certain stimulus location and orientation, but are blind to stimulus variation along other dimensions such as object identity. A different metric would describe neurons in inferotemporal cortex, where a neuron might respond preferentially to a certain face while being insensitive to its exact illumination or location on the retina. Determination of the stimulus preference metric is crucial to understanding the function of a given brain area. In what follows, I examine to what extent brain-imaging measurements $I(s)$ can reveal the preference metric of the underlying neurons.

The subtraction method. In this method, one simply examines the response of a given voxel to different stimuli s_1 and s_2. Then two different cases can occur. In the simplest case, one stimulus may be a better stimulus than the other for a majority of the neurons : $\|s_1 - s_i^{pref}\| < \|s_2 - s_i^{pref}\|$ for most neurons i. Then it follows that $I(s_1)>I(s_2)$: one can measure a stronger response of the whole voxel to the preferred stimulus that to the non-preferred stimulus. This situation occurs when one is probing the neural " macrocode " of a given cortical region, for instance the retinotopy of the primary visual cortex. At some level, it is

generally possible to identify a parameter that is preferred by most if not all neurons in the considered area (e.g. a specific retinal location for a V1 voxel, or faces versus objects for a given fusiform voxel, etc.).

When one is probing the neural " microcode ", however, it is generally not the case that all the neurons have the same global preference. Rather, populations of neurons with different preferences are intermixed at a sub-millimetric spatial scale. In any given voxel, there might be just as many neurons that prefer s_1 over s_2, than there are neurons that prefer s_2 over s_1. It follows that $I(s_1) \cong I(s_2)$: the signals recorded to stimuli s_1 and s_2 cannot be distinguished. It is important to note that this can occur even though the stimuli s_1 and s_2 vary along a dimension which is relevant to the neurons in the considered area, and even though they excite different populations of neurons. The spatial intermingling of these populations, together with the coarse size of the voxels relative to the neural scale, renders them impossible to detect with current imaging methods.

The priming method. The priming method measures the response to the same target stimulus *s*, but varies the context in which it is presented by preceding it with variable primes. For simplicity, we consider here only the simplest case in which a single presentation of a prime stimulus *p* precedes the presentation of the target stimulus *s*. Assuming a linear fMRI response, the measured signal is a simple combination of the signals induced by *p* and *s* :

[3] $I(p,s) = \alpha I(p) + \beta I(s)$

In this equation, the attenuation factors α and β characterize how the measured signal differs from the simple sum of the activations *I*(*p*) and *I*(*s*) that would be found if *p* and *s* were presented in isolation. The α term reflects the fact that the prime-induced activation

can be attenuated, for instance because the prime *p* is presented for a shorter duration, or much in advance of the target *s*, etc. At one extreme, $\alpha = 0$ if the time interval between *p* and *s* is sufficiently long that one can measure the activation induced by *s* without any contamination by the activation induced by *p* (e.g. in long-term priming experiments).

More important, however, is the β term. This term is introduced to take into account the repetition suppression effect of *p* on *s*. I tentatively propose that, at the single neuron level, the neural response of neuron *i* to stimulus *s* preceded by prime *p* can be written as :

[4] $f_i(p,s) = (1 - A(\|p - s\|))\ F(\|s - s_i^{pref}\|)$

where *A* is a function that decreases monotonically from A_0 to zero. A_0 measures the amount of habituation to a repetition of the same stimulus, expressed as a percentage of the normal imaging response to that stimulus when it is not repeated. The function A, which specifies the shape and amount of repetition effect, is expected to vary with the details of the priming paradigm. For instance, it would be expected to decrease exponentially as a function of the prime-target asynchrony in a subliminal priming situation.

Equation [4] makes several assumptions that remain to be verified. First it assumes that repetition suppression has a multiplicative effect on neuronal firing: the curve indexing the tuning of the cell to the target is unchanged by priming, but the intensity of its firing is modulated as a function of the distance between the prime and target. Second, it assumes that repetition suppression is a continuous phenomenon : there can be graded levels of adaptation as a function of the degree of similarity between prime and target, rather than a mere categorical distinction between repeated and non-repeated situations. Third, and crucially, equation [4] assumes that the **same metric** underlies stimulus preference and repetition suppression. The hypothesis here is that what counts as a repetition must be measured along the same dimensions that characterize tuning curves in the area of interest. If, say, an infero-

temporal neuron prefers a given face while being "blind" to variations in illumination, equation [4] predicts that repetition suppression should be sensitive to exactly the same variables, such that the neuron would show exactly the same degree of habituation if the preferred face was repeated with a different illumination or with the same illumination. This putative property of repetition suppression remains to be confirmed at the single-cell level.

Although speculative, equation [4] has the advantage of making explicit quantitative predictions that may guide the much-needed further electrophysiological research into the characteristics of the repetition suppression phenomenon. At present, the biological mechanisms of repetition suppression are unknown. One possibility is that the amount of repetition suppression is determined locally, for instance as a direct reflection of how much the neuron has fired in the recent past. Such a simple habituation mechanism is unlikely to be sufficient, however. It would predict that habituation at the single-neuron level is just a function of the similarity between the prime and the neuron's preferred stimulus, rather than between the prime and the target as proposed in equation [4]. This is a significantly distinction that could be separated experimentally. The present proposal that repetition suppression depends on a global measure of prime-target similarity suggests that one should rather look for a mechanism at the neuronal population level, for instance in the speed with which a set of neurons converges to a stable population code.

From equation [4], one can derive a version of [3] which makes explicit the repetition suppression effect at the voxel rather than at the single-neuron level:

[5] $$I(p,s) = \alpha\, I(p) + (1 - A(\|s - p\|))I(s)$$

This equation shows clearly that the metric underlying the tuning curves of individual neurons in the measured voxel is directly accessible in the brain-imaging measure *I*.

Does repetition priming always result in reduced activation ? As seen from equation [5], *I(p,s)* reflects both stimulus preference and priming. Can we systematically predict the direction of the resulting effect ? Suppose that an experiment is set up to compare the activation induced by repeating the same stimulus twice, $I_{repeated} = I(s,s)$, with the activation induced on non-repeated trials $I_{non\text{-}repeated} = I(p,s)$ ($p \neq s$). Furthermore suppose that non-repeated trials cause no repetition suppression ($A(\|s - p\|) = 0$). The predicted difference then reduces to :

[6] $$I_{repeated} - I_{non-repeated} = \alpha\,(I(s) - I(p)) - A_0\, I(s)$$

Equation [6] shows that, in general, repetition priming need not always result in a reduced activation. Rather, the measured effect of priming on the fMRI signal is the sum of two terms. The first term reflects the stimulus preference of the observed region and can be positive or negative, while the second term reflects repetition suppression and is always negative. In general, there is no guarantee that the second term will always win. If the activation induced by the non-repeated prime, $I(p)$, is smaller than the one induced by the repeated target $I(s)$, then it is possible for activation to be greater on repeated trials than on non-repeated trials.[1]

However, there is a remarkable complementarity between the priming method and the classical subtraction method. As noted previously, the subtraction method does not work when the contrasted stimuli *p* and *s* are such that $I(p) = I(s) = I$. Yet, it is precisely in this situation that the priming method works best. In that case, equation [6] reduces to

[1] Other factors may also contribute to the observation of repetition enhancement rather than repetition suppression. In particular, my equations suppose that the neuronal tuning curves remain unchanged by priming, and that priming acts only as a temporary and reversible multiplicative factor on neural firing rates. If, however, the presentation of the primes leads to learning-induced changes in tuning curves, this may compensate or even overturn repetition-suppression effects (see e.g. Henson, Shallice, & Dolan, 2000).

[7] $I_{repeated} - I_{non-repeated} = -A_0\, I$

which is necessarily negative. More generally, the observed activation on a given trial with prime *p* and target *s*, according to equation [5], reduces to :

[8] $I(p,s) = (\alpha + 1 - A(\|s - p\|))I$

Thus, the observable *I(p,s)* is found to depend *only* on the measure of similarity between prime *p* and target *s* for the voxel under consideration.

The 2x2 priming design. When stimuli *p* and *s* do not cause identical levels of activation ($I(p) \neq I(s)$), or when it is unknown if they do, it is still possible to obtain a brain-activation measure which is proportional to the similarity between *s* and *p*, and which is guaranteed to decrease on repeated trials. This is achieved by using a two-by-two stimulus design in which the same two stimuli s_1 and s_2 serve both as primes and as targets. Four prime-target pairs are possible, two repeated ones s_1-s_1 and s_2-s_2 , and two non-repeated ones s_1-s_2 and s_2-s_1. The following activation levels are predicted :

	Prime s_1	Prime s_2
Target s_1	$I(s_1, s_1) = \alpha I(s_1) + (1 - A_0) I(s_1)$	$I(s_2, s_1) = \alpha I(s_2) + (1 - A(\|s_1 - s_2\|)) I(s_1)$
Target s_2	$I(s_1, s_2) = \alpha I(s_1) + (1 - A(\|s_1 - s_2\|)) I(s_2)$	$I(s_2, s_2) = \alpha I(s_2) + (1 - A_0) I(s_2)$

The interaction term of this matrix then provides a necessarily negative measure proportional to s_1-s_2 similarity :

[9] $$Interaction = I_{repeated} - I_{non-repeated} = -(A_0 - A(\|s_1 - s_2\|))\frac{I(s_1) + I(s_2)}{2}$$

Furthermore, the influence of $I(s_1)$ and $I(s_2)$ can be removed by normalizing the interaction term with respect to the amount of activation obtained on repeated trials :

[10] $$Q(s_1, s_2) = \frac{I_{repeated} - I_{non-repeated}}{I_{repeated}} = -\frac{A_0 - A(\|s_1 - s_2\|)}{\alpha + 1 - A_0}$$

The resulting quantity, Q, is always negative. Note that Q does not depend on $I(s_1)$ and $I(s_2)$, and hence is devoid of all stimulus preference effects. Q provides a direct estimate of how similar the stimuli s_1 and s_2 appear to neurons in the studied voxel. By varying the nature of the relation between s_1 and s_2 , and measuring Q using a 2x2 priming design, it becomes possible to directly evaluate the metric or " neural code " by which neurons in the considered area classify various stimuli as similar or dissimilar. All of the experiments reviewed earlier used such a design, with the same stimuli appearing with the same frequency as primes and targets while only the prime–target relation varied.

Conclusion: subliminal and supraliminal processing

Neuroimaging studies supplement behavioral studies of priming by identifying the neural structures successively activated by the primes, and their associated neural codes. The results indicate that subliminal primes activate abstract visual, semantic and even motor levels of representation. Given such an unexpected depth of processing, why do the subliminal primes fail to reach consciousness? And what additional neural events occurs when the

threshold of consciousness is reached? In this conclusion, I have space only for a few theoretical and empirical considerations on these issues.

The neuronal workspace model (Dehaene, Kerszberg, & Changeux, 1998a; Dehaene & Naccache, 2001) proposes that access to consciousness is associated with the sudden, coordinated firing of many neurons distributed throughout the brain, though particularly concentrated in parietal, prefrontal, and cingulate cortices, and linked by a dense network of long-distance axons. Dense recurrent connections allow information that reaches this level to become quickly available to many different processes including categorization, evaluation, memorization and intentional action. This contrasts with the encapsulation of subliminal information to narrow and highly specialized processors. Because workspace neurons are densely interconnected by multiple recurrent connections, the hypothesis predicts the existence of a dynamical threshold for masked perception. Above a critical level, neural activation becomes self-amplifying and suddenly jumps to a much higher and longer-lasting level ('ignition'). Below this threshold, activation is weak and transient, though it can still propagate in a feedforward manner.

According to this theory, masking acts by shutting down the transient activation of the prime before the interplay of bottom-up and top-down connections has time to amplify it and render it accessible at a whole-brain scale. This view is compatible with neurophysiological recordings in infero-temporal cortex or frontal eye field while awake monkeys were presented with brief masked or unmasked stimuli (Kovacs, Vogels, & Orban, 1995; Rolls & Tovee, 1994; Thompson & Schall, 1999). Masking was found to leave the initial phasic response of the neurons unchanged, while preventing almost entirely the later, more sustained part of the firing train (see also Super, Spekreijse, & Lamme, 2001). I obtained analogous evidence from fMRI and ERPs using masked or unmasked words (Dehaene et al., 2001). Compared to word-absent trials, masked words caused only a small transient activation which was increasingly

smaller as one moved from extrastriate cortex to fusiform gyrus and precentral cortex. Unmasking the words greatly enhanced activation in the same areas and created a large-scale activation which included distant parietal, inferior prefrontal and midline precentral/cingulate cortices. Unmasking also enhanced the long-distance correlation between those sites and, in ERP recordings, was associated with a enhanced late positive complex (P300) that was absent or greatly reduced in the masked situation. Further work should examine whether this predicted correlation between large-scale amplification and conscious access continues to hold on a single-trial basis during masking as well as other paradigms such as change blindness or the attentional blink (see Beck, Rees, Frith, & Lavie, 2001; Vogel, Luck, & Shapiro, 1998).

References

Abrams, R. L., & Greenwald, A. G. (2000). Parts outweigh the whole (word) in unconscious analysis of meaning. Psychol Sci, 11(2), 118-24.

Bar, M., Tootell, R. B. H., Schacter, D. L., Greve, D. N., Fischl, B., Mendola, J. D., Rosen, B. R., & Dale, A. M. (2001). Cortical mechanisms specific to explicit visual object recognition. Neuron, 29, 529-535.

Beck, D. M., Rees, G., Frith, C. D., & Lavie, N. (2001). Neural correlates of change detection and change blindness. Nature Neurosci., 4, 645-650.

Bowers, J. S., Vigliocco, G., & Haan, R. (1998). Orthographic, phonological, and articulatory contributions to masked letter and word priming. J Exp Psychol Hum Percept Perform, 24(6), 1705-19.

Buckner, R. L., Koutstaal, W., Schacter, D. L., & Rosen, B. R. (2000). Functional MRI evidence for a role of frontal and inferior temporal cortex in amodal components of priming. Brain, 123, 620-640.

Cheesman, J., & Merikle, P. M. (1984). Priming with and without awareness. Percept. Psychophys., 36, 387-395.

Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M. A., & Michel, F. (2000). The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. Brain, 123, 291-307.

Cohen, L., Lehericy, S., Chochon, F., Lemer, C., Rivaud, S., & Dehaene, S. (2002). Language-specific tuning of visual cortex? Functional properties of the Visual Word Form Area. Brain, 125(Pt 5), 1054-69.

Dagenbach, D., Carr, T. H., & Wilhelmsen, A. (1989). Task-induced strategies and near-threshold priming: Conscious effects on unconscious perception. J Mem Lang, 28, 412-443.

Damian, M. F. (2001). Congruity effects evoked by subliminally presented primes: automaticity rather than semantic processing. J Exp Psychol Hum Percept Perform, 27(1), 154-65.

Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998a). A neuronal model of a global workspace in effortful cognitive tasks. Proc. Natl. Acad. Sci. USA, 95, 14529-14534.

Dehaene, S., LeClec'H, G., Poline, J. B., LeBihan, D., & Cohen, L. (2002). The visual word form area: A prelexical representation of visual words in the fusiform gyrus. NeuroReport, 13(3), 1-5.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition, 79, 1-37.

Dehaene, S., Naccache, L., Cohen, L., Le Bihan, D., Mangin, J. F., Poline, J. B., & Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. Nature Neurosci., 4, 752-758.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P. F., & Le Bihan, D. (1998b). Imaging unconscious semantic priming. Nature, 395, 597-600.

Dehaene, S., Spelke, E., Stanescu, R., Pinel, P., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. Science, 284, 970-974.

Eimer, M., & Schlaghecken, F. (1998). Effects of masked stimuli on motor activation: behavioral and electrophysiological evidence. J Exp Psychol Hum Percept Perform, 24(6), 1737-47.

Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking. Trends Cog. Sci., 4, 345-352.

Ferrand, L., & Grainger, J. (1994). Effects of orthography are independent of phonology in masked form priming. Q J Exp Psychol [A], 47(2), 365-82.

Forster, K. I. (1998). The pros and cons of masked priming. J Psycholinguist Res, 27(2), 203-33.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. J. Exp. Psychol. Learn. Mem. Cogn., 10, 680-698.

Greenwald, A. G., Abrams, R. L., Naccache, L., & Dehaene, S. (2002). Long-term semantic memory versus contextual memory in unconscious number processing. submitted.

Greenwald, A. G., Draine, S. C., & Abrams, R. L. (1996). Three cognitive markers of unconscious semantic activation. Science, 273(5282), 1699-702.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. Neuron, 24(1), 187-203.

Grill-Spector, K., Kushnir, T., Hendler, T., & Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. Nature Neurosci., 3(8), 837-843.

Henson, R., Shallice, T., & Dolan, R. (2000). Neuroimaging evidence for dissociable forms of repetition priming. Science, 287(5456), 1269-72.

Henson, R. N. A., Shallice, T., Gorno-Tempini, M.-L., & Dolan, R. J. (2002). Face repetition effects in implicit and explicit memory tests as measured by fMRI. Cereb. Cortex, 12, 178-186.

Hommel, B. (2000). The prepared reflex: Automaticity and control in stimulus-response translation. In S. Monsell & J. Driver (Eds.), *Control of cognitive processes: Attention and Performance, Vol. XVIII* (pp. 247-273). Cambridge, MA: MIT Press.

Jaskowski, P., van der Lubbe, R. H., Schlotterbeck, E., & Verleger, R. (2002). Traces left on visual selective attention by stimuli that are not consciously identified. Psychol Sci, 13(1), 48-54.

Koechlin, E., Naccache, L., Block, E., & Dehaene, S. (1999). Primed numbers: Exploring the modularity of numerical representations with masked and unmasked semantic priming. J. Exp. Psychol. Hum. Percept. Perf., 25, 1882-1905.

Kourtzi, Z., & Kanwisher, N. (2000). Cortical regions involved in perceiving object shape. J Neurosci, 20(9), 3310-8.

Kovacs, G., Vogels, R., & Orban, G. A. (1995). Cortical correlate of pattern backward masking. Proc. Natl. Acad. Sci. USA, 92, 5587-5591.

Marcel, A. J. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. Cogn Psychol, 15, 197-237.

Marsolek, C. J., Kosslyn, S. M., & Squire, L. R. (1992). Form-specific visual priming in the right cerebral hemisphere. J. Exp. Psychol. Learn. Mem. Cogn., 18(492-508).

Naccache, L., Blandin, E., & Dehaene, S. (2002). Unconscious masked priming depends on temporal attention. Psychol. Sci., in press.

Naccache, L., & Dehaene, S. (2001a). The Priming Method: Imaging Unconscious Repetition Priming Reveals an Abstract Representation of Number in the Parietal Lobes. Cereb Cortex, 11(10), 966-74.

Naccache, L., & Dehaene, S. (2001b). Unconscious semantic priming extends to novel unseen stimuli. Cognition, 80, 215-229.

Neely, J. H., & Kahan, T. A. (2001). Is semantic activation automatic? A critical re-evaluation. In H. L. Roediger, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), The nature of remembering: Essays in honor of Robert G. Crowder (pp. 69-93). Washington D.C.: American Psychological Association.

Neumann, O., & Klotz, W. (1994). Motor responses to non-reportable, masked stimuli: Where is the limit of direct motor specification. In C. Umiltà & M. Moscovitch (Eds.), Attention and Performance XV: Conscious and non-conscious information processing (pp. 123-150). Cambridge, Mass.: MIT Press.

Paulesu, E., Demonet, J. F., Fazio, F., McCrory, E., Chanoine, V., Brunswick, N., Cappa, S. F., Cossu, G., Habib, M., Frith, C. D., & Frith, U. (2001). Dyslexia: cultural diversity and biological unity. Science, 291(5511), 2165-7.

Raichle, M. E., Fiez, J. A., Videen, T. O., MacLeod, A. K., Pardo, J. V., Fox, P. T., & Petersen, S. E. (1994). Practice-related changes in human brain functional anatomy during non-motor learning. Cereb. Cortex, 4, 8-26.

Reynvoet, B., & Brysbaert, M. (1999). Single-digit and two-digit Arabic numerals address the same semantic number line. Cognition, 72(2), 191-201.

Reynvoet, B., Caessens, B., & Brysbaert, M. (2002). Automatic stimulus-response associations may be semantically mediated. Psychon Bull Rev, 9, 107-112.

Rolls, E. T., & Tovee, M. J. (1994). Processing speed in the cerebral cortex and the neurophysiology of visual masking. Proc. R. Soc. Lond. B Biol. Sci., 257(1348), 9-15.

Schmidt, T. (2002). The finger in flight: real-time motor control by visually masked color stimuli. Psychol Sci, 13(2), 112-8.

Shaywitz, B. A., Shaywitz, S. E., Pugh, K. R., Mencl, W. E., Fulbright, R. K., Skudlarski, P., Constable, R. T., Marchione, K. E., Fletcher, J. M., Lyon, G. R., & Gore, J. C.

(2002). Disruption of posterior brain systems for reading in children with developmental dyslexia. Biol. Psychol., in press.

Shaywitz, S. E., Shaywitz, B. A., Pugh, K. R., Fulbright, R. K., Constable, R. T., Mencl, W. E., Shankweiler, D. P., Liberman, A. M., Skudlarski, P., Fletcher, J. M., Katz, L., Marchione, K. E., Lacadie, C., Gatenby, C., & Gore, J. C. (1998). Functional disruption in the organization of the brain for reading in dyslexia. Proc Natl Acad Sci U S A, 95(5), 2636-41.

Super, H., Spekreijse, H., & Lamme, V. A. F. (2001). Two distinct modes of sensory processing observed in monkey primary visual cortex (V1). Nature Neurosci., 4, 304-310.

Thompson, K. G., & Schall, J. D. (1999). The detection of visual signals by macaque frontal eye field during masking. Nature Neurosci., 2, 283-288.

Vogel, E. K., Luck, S. J., & Shapiro, K. L. (1998). Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. J Exp Psychol Hum Percept Perform, 24(6), 1656-74.

Vuilleumier, P., Henson, R. N., Driver, J., & Dolan, R. J. (2002). Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. Nat Neurosci, 5(5), 491-9.

Wagner, A. D., Koutstaal, W., Maril, A., Schacter, D. L., & Buckner, R. L. (2000). Task-specific repetition priming in left inferior prefrontal cortex. Cereb Cortex, 10(12), 1176-1184.

Warrington, E. K., & Shallice, T. (1980). Word-form dyslexia. Brain, 103, 99-112.

Figure Legends

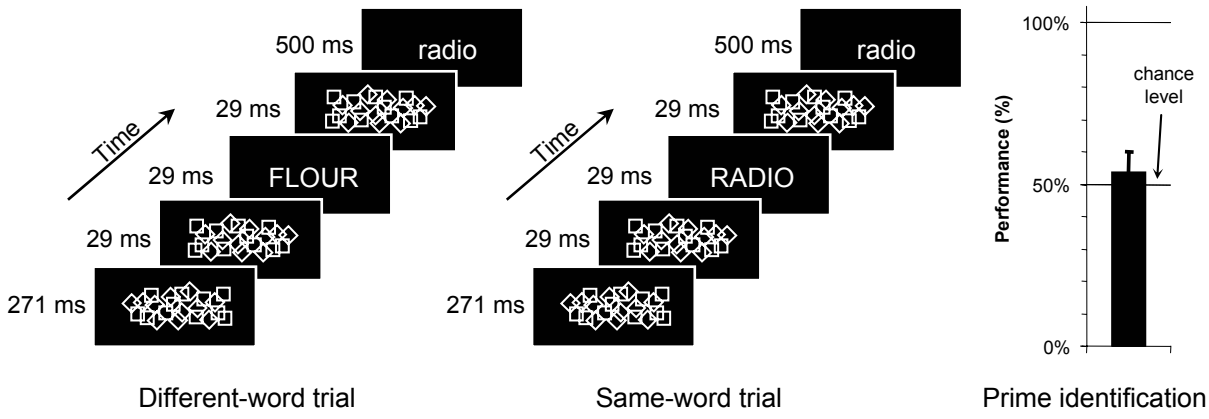
Figure 1. Design of a typical repetition priming study (redrawn from Dehaene et al., 2001). The sequence of stimuli (left) typically includes a long pre-mask (here, two successive random geometrical patterns) as well as an optional brief post-mask surrounding the prime word. Chance-level performance is obtained on a two-alternative forced-choice test of prime identification (right).

Figure 2. Parallels and differences between behavioral and neuroimaging measures of word priming. In the visual word form area of the left fusiform gyrus (inset), repetition suppression resists case change (top) and is independent of letter similarity (middle), paralleling behavioral priming. However, the fusiform gyrus appears insensitive to the whole-word configuration of letters and thus shows priming even for anagrams, whereas behavioral priming shows whole-word selectivity (redrawn from data in Dehaene et al, 2001; Dehaene, in preparation). In this and subsequent figures, error bars represent standard errors calculated across subjects after removal of overall intersubject variability common to all conditions.

Figure 3. A number priming paradigm affords separate analyses of quantity and motor priming (redrawn from Dehaene et al., 1998; Naccache & Dehaene, 2001; see text for details).

Figure 4. Advantages of the priming method over the classical subtraction method in neuroimaging. An fMRI voxel typically contains over a million neurons, each tuned to a particular preferred stimulus in perceptual or abstract space (left). Although two stimuli may be encoded by different populations of neurons within this voxel, activation averaged over the entire population of neurons is likely to fail to reveal any measurable difference between them

(middle). However, thanks to the reduced level of activation associated with repetition suppression, differences may be observed by comparing the activation evoked by the same stimulus S_2 when preceded either by the same stimulus S_2, or by a different stimulus S_1 (right). This is true only if the neural code in this voxel distinguishes between S_1 and S_2.

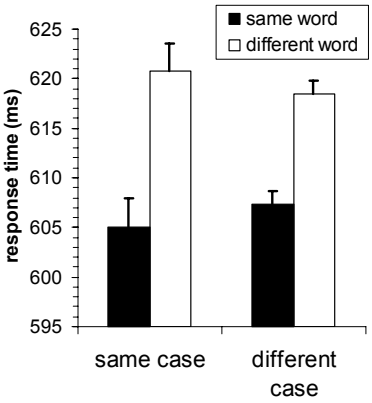
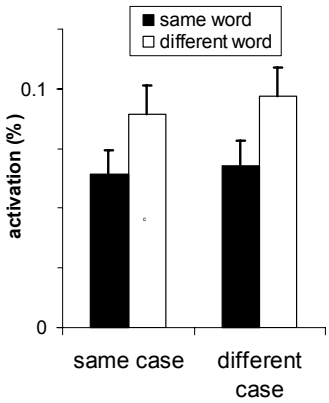
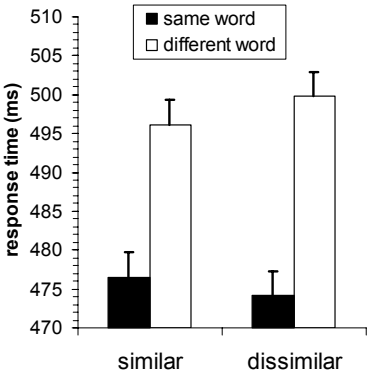
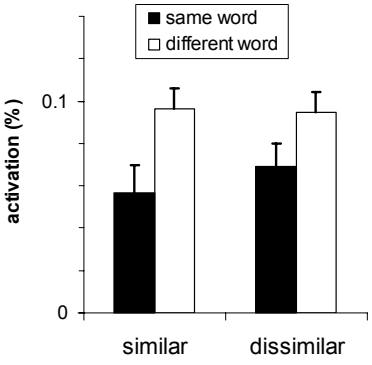
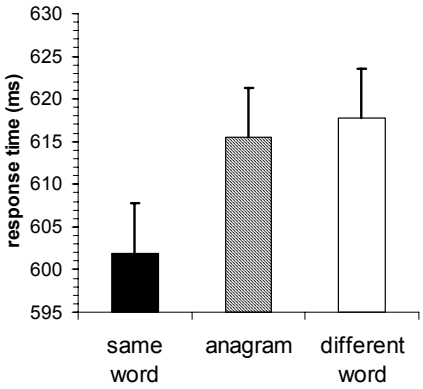
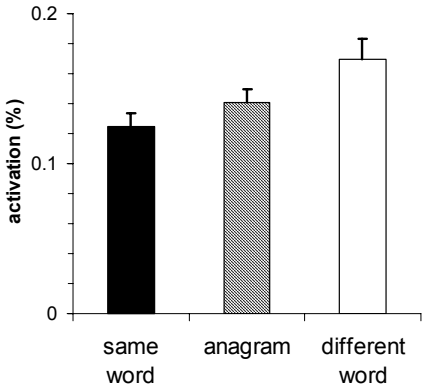


Different-word trial

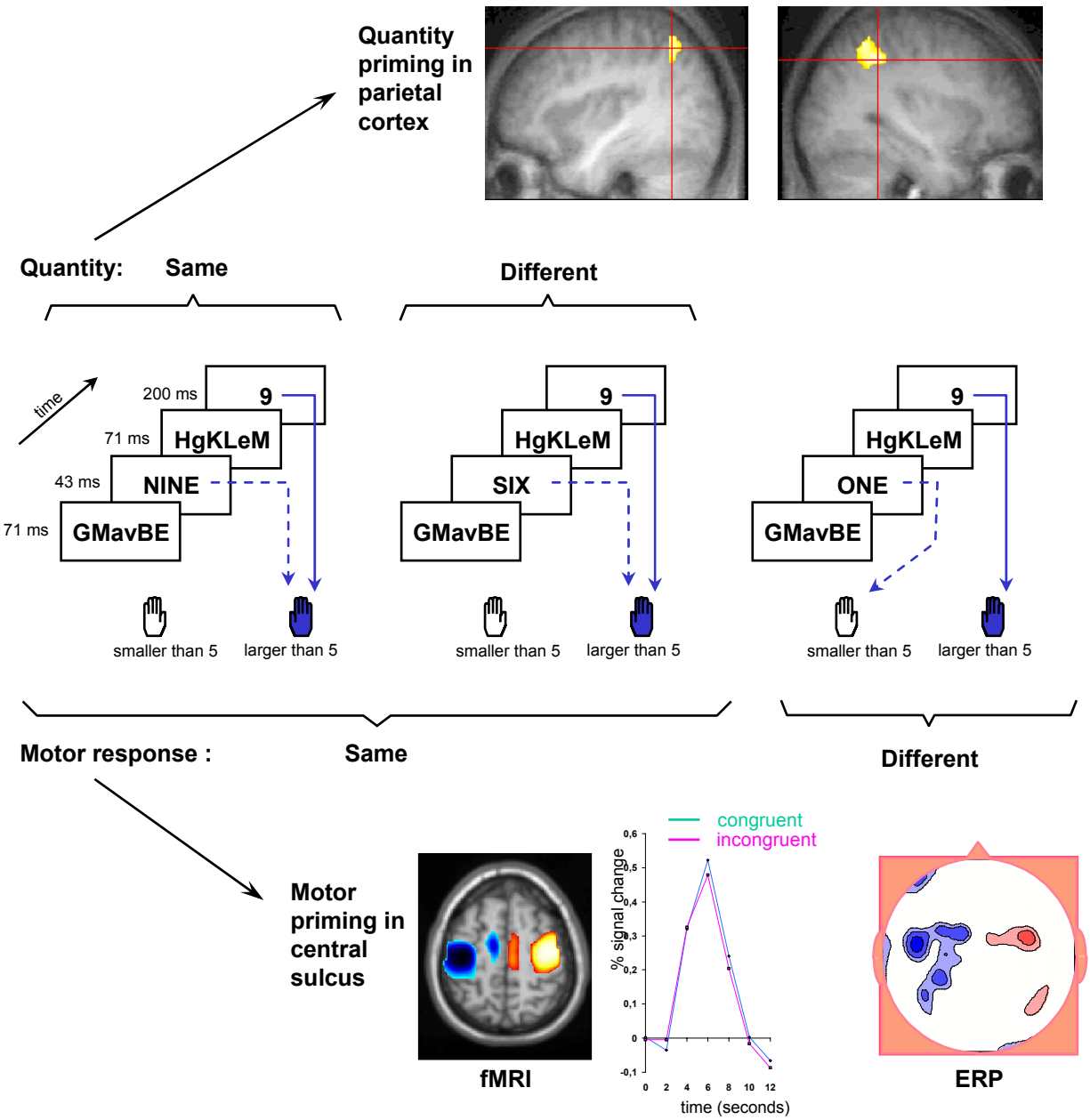
Same-word trial

Prime identification

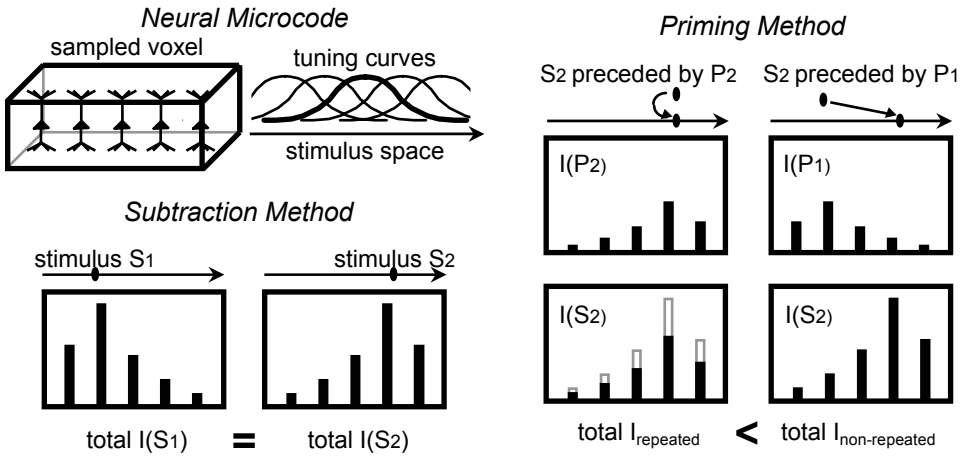
Dehaene A&P 2002, figure 1

Parameter manipulated	Response times	Left fusiform activation (VWFA)
Exp. 1 : Case change **same case:** RADIO-RADIO **diff. Case:** RADIO-Radio		
Exp. 2 : Letter similarity **similar:** COUP-coup **dissimilar:** RAGE-rage		
Exp. 3 : Letter or Word content **same word:** TREFLE-trefle **anagram:** TREFLE-reflet **diff. word:** TREFLE-roquet		

Dehaene A&P 2002, figure 2



Dehaene A&P 2002, figure 3



Dehaene A&P 2002, figure 4