frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 09 January 2012
doi: 10.3389/fpsyg.2011.00394

Mechanisms and representations of language-mediated visual attention

Falk Huettig[1,2]*, Ramesh Kumar Mishra[3] and Christian N. L. Olivers[4]

[1] *Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands*
[2] *Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, Netherlands*
[3] *Centre of Behavioral and Cognitive Sciences, University of Allahabad, Allahabad, India*
[4] *Cognitive Psychology, VU University Amsterdam, Amsterdam, Netherlands*

Edited by:
Andriy Myachykov, University of Glasgow, UK

Reviewed by:
Christoph Scheepers, University of Glasgow, UK
Gerry Altmann, University of York, UK

****Correspondence:***
Falk Huettig, Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, Netherlands.
e-mail: falk.huettig@mpi.nl

The experimental investigation of language-mediated visual attention is a promising way to study the interaction of the cognitive systems involved in language, vision, attention, and memory. Here we highlight four challenges for a mechanistic account of this oculomotor behavior: the levels of representation at which language-derived and vision-derived representations are integrated; attentional mechanisms; types of memory; and the degree of individual and group differences. Central points in our discussion are (a) the possibility that local microcircuitries involving feedforward and feedback loops instantiate a common representational substrate of linguistic and non-linguistic information and attention; and (b) that an explicit working memory may be central to explaining interactions between language and visual attention. We conclude that a synthesis of further experimental evidence from a variety of fields of inquiry and the testing of distinct, non-student, participant populations will prove to be critical.

Keywords: language, attention, vision, memory, eye movements

INTRODUCTION

A hallmark of human cognition is its ability to integrate rapidly perceptual (e.g., visual or auditory) input with stored linguistic and non-linguistic mental representations. This is particularly apparent during language-mediated eye gaze, a behavior almost all of us are engaged in every day. For instance, when a mother asks her child to "*look at the frog*" or, during dinner, we are asked to "*pass the salt*," linguistic and visual systems, attention and memory processes, must all be quickly integrated. Yet we know surprisingly little about the nature of these cognitive interactions and the representations involved.

How higher level representations involved in language and memory interact with visual input during language-mediated eye gaze has most directly been explored in the *visual world* paradigm in psycholinguistics and the *visual search* paradigm in the field of visual attention. In the visual world paradigm, participants hear an utterance while looking at a visual display (e.g., a semi-realistic scene, or four spatially distinct objects, or printed words; Cooper, 1974; Tanenhaus et al., 1995; see Huettig et al., 2011b, for review). Typically, the display includes objects mentioned in the utterance as well as distractor objects that are not mentioned. The spoken utterances can be instructions to the participants ("direct action" tasks, e.g., "Pick up the candy," Allopenna et al., 1998) or descriptions or comments on the display ("look and listen" tasks, e.g., Huettig and Altmann, 2005). In the latter case, the participants are asked to look at the screen and to listen carefully to the sentences. The participants' eye movements are recorded for later analyses. Some visual world studies have examined whether items that are phonologically, semantically, or visually related (so-called competitors) to a critical spoken word attract attention. Other studies have investigated how the listeners' perception of the scene and/or their world knowledge about scenes and events affect their understanding of the spoken utterances (e.g., whether listeners anticipate up-coming words). In the visual search paradigm, participants are presented with a display of multiple objects and their task is to find a pre-specified target (defined by a certain feature) as quickly as possible (see Wolfe, 1998, for a review). In most studies of these studies, it is assumed that participants will set up some sort of "perceptual" template (or "attentional set") of the target (e.g., when told to "*look for the red square*") for the remainder of the task. The goal of most visual search studies is to investigate the interaction between the bottom-up salience of the stimulus and the top-down goals of the observer (e.g., Treisman and Sato, 1990; Humphreys and Müller, 1993; Wolfe, 1994; Cave, 1999; Itti and Koch, 2000; Palmer et al., 2000). An important difference between the two paradigms is that in the visual world paradigm the visual display precedes (or occurs simultaneously) with the spoken instruction (or sentence) whereas in visual search studies the (linguistic or visual) instruction precedes the search display.

In short, the main interest of researchers using the visual world paradigm tends to be on aspects of linguistic processing whereas visual search investigators are primarily interested in what determines the efficiency of the search process, how easily conjunctions of basic features (e.g., color and shape) can be found, and whether search involves serial or parallel processing. These distinct focal points of interest have resulted in a theoretical no-man's land in which the exact nature of the interaction of linguistic and visual processing, and of attention and memory, have been left largely unexplored.

The aim of the present paper is to highlight, (a) theoretical challenges to explaining how language, vision, memory, and attention interact and, (b) empirical challenges in view of recent data with young children and illiterates/low literates in the visual world paradigm. We argue that existing theoretical proposals do not discuss (at all or in sufficient detail) four major underpinnings of this oculomotor behavior: levels of representation involved, attentional mechanisms, the nature of memory, and the degree of individual and group differences.

We will therefore first discuss the levels of representation at which language-derived and vision-derived representations are integrated (see Levels of Representation). An explanation of attention will be central for a mechanistic account about how this oculomotor behavior is instantiated and thus, in Section "Attention," we consider the attentional mechanisms which may underlie language-mediated eye gaze. Language–vision interactions of course also involve temporary and long-term memory storage; we reflect on what types of memory may be involved and their nature (see Memory). In Section "Individual and Group Differences," before concluding, we discuss empirical challenges for the investigation of the mechanisms and representations shared by language, vision, attention, and memory; in particular the need to study distinct, non-student, participant populations.

LEVELS OF REPRESENTATION

To understand how language interacts with vision, it is necessary to establish what knowledge types are retrieved when someone is confronted with both language and visual input, as well as how these linguistic and visual representations interact. Furthermore, such representations are likely to change over time as the linguistic input unfolds and the visual image has been available for some time. An early linguistic–visual linking hypothesis was proposed by Tanenhaus and collaborators (Allopenna et al., 1998) which Huettig and McQueen (2007) termed the *phonological mapping hypothesis*. According to this hypothesis, phonological representations are activated by both spoken words and visual objects (i.e., the names of the objects in the display). A match in phonological representations retrieved from both modalities results in an increased likelihood of a saccade toward the location of the (partially) matching visual source. This is in line with many models of spoken word recognition which assume that at a phonological level different candidate words are considered in parallel (cf. Marslen-Wilson and Welsh, 1978; Marslen-Wilson, 1987). Continuous mapping models of spoken word recognition (e.g., McClelland and Elman, 1986; Gaskell and Marslen-Wilson, 1997) assume that lexical access during spoken word recognition is continuous and thus predict that rhyming words (e.g., beaker/speaker) should also be at least weakly activated.

Consistent with these models, Allopenna et al. (1998) observed that the likelihood of fixations to both a picture of a beaker and a picture of a beetle increased as participants heard the word "*beaker*." As acoustic information from "*beaker*" started to mismatch with the phonological information of "beetle," the likelihood of looks to the beetle decreased as the likelihood of looks to the beaker continued to rise. In addition, looks to a picture of a speaker started to increase as the end of the word "*beaker*" acoustically unfolded. The finding that simulations with the TRACE model (McClelland and Elman, 1986) of speech perception, which includes an explicit phoneme layer, closely fit the eye movement data of Allopenna et al. (1998) provided further support for the phonological mapping hypothesis.

It is however important to note that the many demonstrations of the influence of acoustic–phonetic information in visual world studies (e.g., McMurray et al., 2002; Salverda et al., 2003; Shatzman and McQueen, 2006) are consistent with the phonological mapping hypothesis but do not necessarily provide support for it. This is because there is general agreement that spoken word recognition is a cascaded rather than a strictly serial process (e.g., that information from the acoustic signal cascades to higher levels before processing at lower levels is completed) and that thus activation of word form representations cascades further to, for instance, morphological, semantic, and syntactic representational levels. Thus, the initial phonological representations retrieved on hearing the spoken word "*beaker*" may activate semantic representations of beakers as well as beetles, and the mapping between spoken words and the different competing visual objects may therefore take place at the level of semantic/conceptual rather than phonological (or phonetic) representations. This is the *semantic mapping hypothesis*. One could go even further than that and argue that activation of phonetic and semantic representations automatically spreads to the associated visual shapes and thus the match with the visual input occurs at a perceptual level. We could call this the *visual mapping hypothesis*.

Semantic mapping effects were first reported by Cooper (1974), who observed that participants were more likely to fixate pictures showing a snake, a zebra, or a lion when hearing the semantically related word "Africa" than they were to fixate referents of semantically unrelated control words. However, Cooper did not investigate systematically the nature of the semantic effects he observed (e.g., the words "Africa" and "lion" are not only semantically but also associatively related, as they often co-occur, like "computer" and "mouse"). Huettig and Altmann, 2005, see also Yee and Sedivy, 2001, 2006; Dunabeitia et al., 2009) further pursued Cooper's finding by investigating whether semantic properties of spoken words could direct eye gaze toward objects in the visual field in the absence of any associative relationships. Huettig and Altmann (2005) found that participants directed overt attention toward a depicted object (e.g., a trumpet) when a semantically related but not associatively related target word (e.g., "*piano*") acoustically unfolded, and that the likelihood of fixation was proportional to the degree of conceptual overlap (cf. Cree and McRae, 2003). In a similar study (Huettig et al., 2006; see also Yee et al., 2009) observed that several corpus-based measures of word semantics (latent semantic analysis, Landauer and Dumais, 1997; contextual similarity, McDonald, 2000) each correlated well with fixation behavior. Thus, language-mediated eye movements are a sensitive indicator of the degree of overlap between the semantic information conveyed by speech and the conceptual knowledge retrieved from visual objects. The fact that phonological relationships were avoided between spoken words and visual objects in the semantic studies shows that semantic mapping behavior can occur in the absence of phonological mapping.

Evidence for visual mapping (i.e., increased looks to visually related entities, e.g., matching in color or shape) have also been

observed. For example, participants shifted overt attention to a picture of a cable during the acoustic unfolding of the word "*snake*" (shape being the obvious match here, Huettig and Altmann, 2004, 2007; Dahan and Tanenhaus, 2005). In a related study, Huettig and Altmann (2004) found that participants shifted their eye gaze to a picture of a strawberry when they heard "*lips*" (presumably on the basis of the typical color of these objects). The likelihood of fixating a particular visual object thus reflects the overlap between stored knowledge of visual features of a word's referent, accessed on hearing the spoken word, and visual features extracted from the objects in the visual environment.

It is important to note that there are two possible ways in which visual mapping may occur: between the typical visual form of the referent retrieved on hearing the spoken word (e.g., the typical shape of snakes on hearing "*snake*" or the typical color of lips on hearing "*lips*") and the *perceived* visual form or color of the displayed object (in absence of any stored visual form object knowledge) and/or the stored *knowledge* about the typical visual form or color of the displayed object (as retrieved from viewing the object). The shape of an object, the long and thin form of a cable, or the color of a strawberry, can be perceived but is also known. Eye movements that are contingent upon currently perceived information (which may be temporarily stored in visual working memory) cannot easily be dissociated from eye movements that are contingent upon stored information about object form (see also Yee et al., 2011). To investigate this issue Huettig and Altmann (2011) manipulated the presence of color in a series of experiments. The *conceptual* representation of an object's color (i.e., the stored color knowledge about an object) and the *perceived* but non-diagnostic color of an object (i.e., its surface color) can be dissociated. Participants were presented with spoken target words whose concepts are associated with a typical color (e.g., "*spinach*") while their eye gaze was monitored to (i) objects associated with the same typical color but presented in black and white (e.g., a black and white line drawing of a frog), (ii) objects associated with the same typical color but presented in an appropriate but atypical color (e.g., a color photograph of a yellow frog), and (iii) objects typically not associated with the color but presented in the color associated with the target concept (e.g., a green blouse). No effect of stored object color knowledge was found when black and white line drawings or black and white photos were used. A small effect of stored object color knowledge was found when color photographs were used depicting the target object (e.g., a frog) in an atypical but appropriate color (e.g., a yellow frog). The finding that the effect was marginal and occurred more than 1 s after information from the acoustic target word started to become available suggests that stored object color, if anything, has a minor influence on language-mediated eye movements. In contrast, Huettig and Altmann (2011) found a large bias toward objects displayed in the same surface color (as the prototypical color associated with the spoken word) even though the referent of the picture (e.g., a green blouse) was not itself associated with that color. These experiments suggest that online visual mapping between spoken words and visual objects is mainly contingent upon the perceived visual information (temporarily stored in visual working memory) rather than stored object form or color knowledge accessed on viewing the visual objects. Overall thus, three main hypotheses (visual, phonological, and semantic mapping) about the representational levels at which linguistic and visual input match have been proposed. Some recent research has been directed at evaluating these hypotheses.

To counter criticism that looks to phonological competitors in the visual world paradigm might just be due to strategic covert object naming rather than normal lexical analysis of the spoken words (i.e., that the phonological effects reflect a match between the phonological input of the spoken words with strategically retrieved object names bypassing further lexical analysis of the spoken words), Dahan and Tanenhaus (2005) have recently argued that the visual competition effects are "inconsistent with the hypothesis that eye movements merely reflect a match between the unfolding speech and pre-activated phonological representations associated with object locations" (p. 457). They then go on to claim that mapping occurs at the perceptual level, not the lexical level. This is correct in the sense that visual (and semantic) effects also occur in absence of phonological overlap, ruling out the claim that "word–object matching" in the visual world paradigm is *entirely* due to phonological mapping. The visual effects however do not rule out that phonological and semantic mapping (at least sometimes) occur. Moreover, from word–object mapping at a phonological level of representation does not necessarily follow that there is no further lexical analysis of the spoken words.

There is evidence from other paradigms showing that viewers often access the names of objects, even when they do not intend to name them (e.g., Noizet and Pynte, 1976; Zelinsky and Murphy, 2000; Morsella and Miozzo, 2002; Navarette and Costa, 2005; Meyer and Damian, 2007; Meyer et al., 2007; Mani and Plunkett, 2010). Noizet and Pynte (1976) for instance asked their participants to shift eye gaze to three objects, one after another, and to identify them silently. Participants were told that no response was required and that they would not be tested afterward. Noizet and Pynte (1976) observed that participants gazed about 200 ms longer at objects with multi-syllable names (e.g., hélicoptère) than objects with one-syllable names (e.g., main; see Zelinsky and Murphy, 2000, for a similar result). Morsella and Miozzo (2002) used a picture–picture version of the Stroop task in which speakers were shown pairs of superimposed pictures and were instructed to name one picture and ignore the other. They found that participants were faster at naming pictures with distractors that were phonologically related. Thus, the pictures participants were instructed to ignore exerted a phonological influence on production which suggests that participants retrieved the phonological forms of the names of the distractor pictures. As a final example, Mani and Plunkett (2010) recently showed that even 18-months-olds implicitly name visual objects and that these implicitly generated phonological representations prime the infants' subsequent responses in a paired visual object spoken word recognition task. These results suggest that viewing a display of visual objects does result in lexical analysis of the displayed objects, at least in these tasks and if participants have sufficient time to inspect the scene/display.

Huettig and McQueen (2007) tested the hypothesis that neither the simple phonological or visual or semantic mapping hypotheses are correct and that instead there appears to be a complex three-way tug of war among matches on all three levels of representation. In four experiments, participants listened to spoken

sentences including a critical word. The visual displays contained four spatially distinct visual items (a phonological, a semantic, and a visual competitor of the critical spoken word, and a completely unrelated distractor). When participants were given sufficient time to look at the display (i.e., before the critical spoken word), shifts in eye gaze to the phonological competitor of the critical word preceded shifts in eye gaze to shape and semantic competitors. Importantly, with only 200 ms of preview of the same picture displays prior to onset of the critical word, participants no longer preferred the phonological competitor over unrelated distractors, and prioritized the shape and semantic competitors instead. Thus it appears that when there is plenty of time to view the display picture processing progresses as far as retrieval of the pictures' names. But when there was only 200 ms of preview before the onset of the critical spoken word, picture processing still involved retrieval of visual and semantic features, but there was insufficient time to retrieve the pictures' names.

Yee et al. (2011) have recently suggested that *long-term* knowledge about an object's form becomes available *before* information about its function (cf. Schreuder et al., 1984; but see Moss et al., 1997) based on their finding that eye movements mediated by conceptual shape (i.e., a slice of pizza activating the round shape of a whole pizza) were observed with 1000 ms but not with 2000 ms preview of the visual display. The opposite pattern was observed for looks to semantic competitors (i.e., no effects with 1000 ms but a significant bias with 2000 ms preview). This pattern of results is striking but the semantic results appear to be inconsistent with previous research since strong semantic effects have been observed with as little as 200 ms preview in other visual world studies (see Table 4 of Huettig and McQueen, 2007; see also Dell'Acqua and Grainger, 1999, for evidence that 17 ms exposure to pictures of objects is enough to activate gross semantic category information). Future studies could usefully be directed at investigating the differences underlying these seemingly contradictory results.

There is evidence that the nature of the visual environment induces implicit biases toward particular types of mapping during language-mediated visual search. This is because Huettig and McQueen (2007) found a different pattern of results when the pictures were replaced with printed words (the names of the same objects as before). Under these conditions shifts in eye gaze were directed only to the phonological competitors, both when there was only 200 ms of preview and when the displays appeared at sentence onset. This suggests that eye gaze is co-determined by the type of information in the display (i.e., visual objects or words). Further support for this notion was provided in a subsequent series of experiments (Huettig and McQueen, 2011). The same sentences and printed words as in Huettig and McQueen (2007) were used. When semantic and shape competitors of the targets were displayed along with two unrelated words, significant shifts in eye gaze toward semantic but not shape competitors were observed as targets were heard. The results were the same when, semantic competitors were replaced with unrelated words, and in addition, semantically richer sentences were presented to encourage visual imagery, and moreover, participants rated the shape similarity of the stimuli before doing the eye-tracking experiment. Yet none of the cases resulted in rapid shifts in eye gaze to shape competitors, There was a late shape-competitor bias (more than 2500 ms after target onset) in all experiments, which shows that participants can in principle access shape information from printed words. These data thus show that shape information is not used in online search of printed word displays whereas it is used with picture displays. In other words, the likelihood of mapping between language-derived visual representations and vision-derived visual representations is contingent upon the nature of the visual environment. Finally, at least when printed word displays are used, recent results suggest that language–vision mapping can also occur at an orthographic representational level (Salverda and Tanenhaus, 2010; see also Myachykov et al., 2011, for discussion of mapping processes at the syntactic level in a language production task; and Mishra and Marmolejo-Ramos, 2010, for an embodied cognition account).

In sum, research has shown that with picture displays, fixations can be determined by matches between knowledge retrieved on the basis of information in the linguistic and in the visual input at phonological, semantic, and visual levels of representation. With printed word displays, fixations are determined by online matches at phonological, semantic, and orthographic levels. The exact dynamics of the representational level at which such mapping occurs however is co-determined by the timing of cascaded processing in the spoken word and object/visual word recognition systems, by the temporal unfolding of the spoken language, and by the nature of the visual environment (e.g., which other representational matches are possible).

ATTENTION

The mapping hypotheses outlined so far describe the levels at which language-derived and vision-derived representations match during language-mediated eye gaze. They do not however provide any mechanistic account about how this oculomotor behavior is instantiated. Attention will probably be central to such an explanation, as the eye movements are likely an overt expression of shifts in the attentional landscape (such shifts may of course also occur covertly, e.g., Posner, 1980). Within the field of attention research, objects in the visual field are assumed to compete for representation, with the strongest object being selected for further behavior (e.g., a manual or oculomotor response; Wolfe, 1994; Desimone and Duncan, 1995; Itti and Koch, 2000; Miller and Cohen, 2001). This competition is generally thought to be biased by two types of mechanism: a bottom-up or feedforward mechanism representing stimulus strength, and a top-down or feedback mechanism representing the current goals of the observer (see, e.g., Theeuwes, 2010, for a review). For example, a bright red poppy in a field of grass may automatically capture one's eyes, but it will especially do so if one is looking to compile a nice bouquet of wild flowers.

Note that this attentional framework is not immediately applicable to visual world behavior. For one, in many visual world studies there is no clear task goal that would *a priori* be expected to induce visual biases. The task is often simply to look around and at the same time to just listen to the spoken input. As has been pointed out recently (Huettig et al., 2011b; Salverda et al., 2011) visual world type interactions may well be modulated by different task settings, but so far this has received little systematic investigation. Furthermore, visual world experiments are typically little concerned with the visual stimulus properties. The visual objects are chosen for linguistically relevant characteristics (i.e., their names

or meanings), and not their physical characteristics (though see Huettig and Altmann, 2004, 2007, 2011; Dahan and Tanenhaus, 2005).

We can remedy this by assuming that not only visual features or task goals add to the attentional weight of a visual object, but also its linguistic (e.g., phonological) and semantic properties. Indeed this is what a number of models reported in the visual world literature do. Roy and Mukherjee's (2005) probabilistic rule model integrates sentence-level and visual information, such that each word in an unfolding sentence incrementally influences the distribution of probabilities across the visual scene, based on the fit of the visual context with the current word. The distribution of probabilities are interpreted as attentional distributions, such that processing priority is assumed to be distributed over the visual objects in the scene. According to Altmann and colleagues (Altmann and Kamide, 2007; Altmann and Mirkovic, 2009), attending to a language-matching visual object is an emergent property of spreading activation. The visual and linguistic input overlap at for example the semantic level, where they reinforce each other. This increased activation then spreads back to the specific linguistic and visual representations, including the visual location, which then serves as a saccadic target. In the model of Mayberry et al. (2009), attention is directed to identified visual regions in order to establish a reference for the spoken input. The relationship between language and vision is reciprocal, in that the referent (i.e., attended) object in turn influences the interpretation of the incoming speech. In other words, the language comprehension system makes use of whatever information is available, including visual information. This way, language becomes grounded in a visual environment, in line with for example developmental findings. Likewise, a neural net implementation of the model learns to interpret ambiguous linguistic input by attending to seemingly relevant (i.e., matching) visual input. The net result is the same as for the other models: matching visual input becomes more strongly represented. Finally, in Kukona and Tabor's (2011) recent dynamical systems model of the visual world paradigm, attention is expressed as a landscape of local attractors reflecting the visual objects, a landscape that continuously changes on the basis of the linguistic input.

Whereas psycholinguistics has welcomed attention into their models, very few visual search studies have looked at the role of language. One exception is a study by Wolfe et al., 2004; see also Vickery et al., 2005), who compared visual search under verbal (i.e., written) instructions to that under visual instructions. Observers were asked to search a complex display for a unique (but non-salient) target. The target changed from trial to trial, as was indicated by an instruction. This instruction was either pictorial in nature (i.e., it showed an exact picture of the target), or it was a written description (e.g., it read "blue square"). Furthermore, the SOA between the cue and the search display was varied. The results showed that pictorial cues were very effective: already for SOAs of 200 ms, performance reached asymptote, and search was as fast as in a baseline condition in which the target always remained identical from trial to trial (and thus no instruction was necessary). Performance was considerably worse for the written cues. Search speed was never comparable to the baseline condition, and even after 1600 ms (the greatest SOA measured) it had not reached asymptote yet. This despite the fact that the written cues described very simple visual forms that the observers had seen over and over again during the course of the experiment. This suggests that, in visual search, observers do not necessarily create a visual template from a verbal description, and instead complete the task on the basis of a less precise representation which could be linguistic in nature, but is in any case more abstract than a visual template.

Whatever the precise model, note that for linguistic content to be translated into a spatial attentional landscape, a considerable binding problem needs to be solved, linking the phonological and semantic codes to a specific visual location. Cognition needs what has been referred to as *grounding*, *situating*, or *indexing*. This problem has been recognized by many (e.g., Richardson and Spivey, 2000; Kukona and Tabor, 2011), but so far has not been adequately solved by visual world models. According to Altmann and Mirkovic (2009), the increased activation of the overlapping representations within a supramodal network automatically spreads back to the matching object's location. Such a network is not necessarily a separate supramodal module in itself, but may emerge from the global, linked activity in the range of networks involved in representing the visual and linguistic input. Useful as it is as a general explanatory framework, it begs the question as to how a representation within such a network knows what the (spatial) source is of its activity. If everything is active, how can one piece of information be specifically bound to another? In the typical visual world display, there are multiple objects, and hence multiple active locations, any of which could be the source. Altmann and Mirkovic propose that an object's location as well as its more symbolic properties are part of one and the same "representational substrate," but they left unspecified how this representational substrate would look like.

The problem has been recognized within the attention literature, where the question boils down to how separate visual features such as color and orientation can be tied to a specific object or location (Treisman and Gelade, 1980; Treisman, 1996; Reynolds and Desimone, 1999). One classic solution has been the idea that by locally attending to an object, its features will become activated together. Thus, attention causes binding. Obviously, this solution does not suffice here, since we try to explain exactly the opposite: how the binding of information causes attention. One promising way of creating a representational conglomerate that includes an object's location as well as its identity is through local interactions of feedforward and feedback mechanisms (e.g., Lamme and Roelfsema, 2000; van der Velde and de Kamps, 2001; Hamker, 2004; Vanduffel et al., 2008). The idea is that a visual target object is first represented in low-level perceptual layers, which due to their retinotopic organization and small receptive fields include detailed spatial information. These layers then feed forward into layers that eventually recognize the identity of the object. These higher layers are not retinotopically organized and due to large receptive fields, location information is largely lost. Part of the recognition layers will recognize the target object and become active accordingly. This activity is fed back to the lower layers, but due to the loss of location information this feedback is spatially non-specific. However, the feedback can be made spatially specific by making it interact with the feedforward activity that drove the recognition in the first place. That is, at each layer, the feedback is gated by,

or correlated with the feedforward activity that fed into that layer. Thus, the feedback trickles down the representational ladder and becomes more and more localized, thus tying a recognition unit to a specific visual instantiation. There is no *a priori* reason why layers representing linguistic information about visual objects could not be linked in the same fashion, and thus create the representational substrate proposed by Altmann and Mirkovic (2009).

In sum, little research so far has investigated the exact nature of the attentional mechanisms underlying language-mediated eye gaze. The most concrete proposal to date postulates that language-mediated visual orienting arises because linguistic and non-linguistic information and attention are instantiated in the same common coding substrate. Local microcircuitries involving feedforward and feedback loops may instantiate such a representational substrate.

MEMORY

As with virtually any cognitive process, the interactions between language and eye movements involve memory. The question is what types of memory are involved. There is no doubt that long-term memory plays a crucial role, as it provides the semantic, phonological, and visual knowledge base (or "type" representations) on which these interactions are based. Spreading activation then travels along the associations formed within and between these different types of knowledge networks. Indeed there is growing evidence that both visual and semantic knowledge stored in long-term memory representations automatically affect visual selection. For example, in a visual search task, Olivers (2011) asked participants to search a display for a grayscale version of a known traffic sign. On each trial a distractor sign was presented in a color which was either related or unrelated to the target sign. For example, when looking for a black and white hexagonal STOP sign (which is usually red in Europe) the distractor could be a red triangular warning sign (related) or a blue square parking sign (unrelated). Distractors interfered more with participants' search when the color of the distractor sign was related than when their color was unrelated even though color was completely irrelevant to the task. Apparently, the participants could not help but retrieve the associated color. Similarly, Moores et al. (2003) found interference stemming from a conceptual relationship. For example, when observers were asked to look for a picture of a motorbike, they were more distracted by a picture of a helmet than a picture of a football. Finally, Meyer et al. (2007) reported interference from an overlap in object name, for example when observers were asked to look for a bat (the animal), they were distracted by a picture of a baseball bat. Similarly, Soto and Humphreys (2007) found that after the instruction to remember the word "red," observers were more distracted by red objects in the display. Although some working memory was involved in this study, the link between the word and the visual color representation must obviously rely on LTM knowledge.

However, as argued earlier, the mere spread of activation on the basis of long-term links is insufficient to explain such findings in visual search, as well as visual world behavior. Note that both visual search and visual world displays are often characterized by a substantial degree of arbitrariness in the collection of objects presented and the locations where these objects are put. Unlike real world scenes in which particular objects are often associated with particular locations (for example when opening the fridge, the milk bottle is typically located in the lower door compartment), in visual world displays the target object (e.g., the "trumpet") may be presented in the top left of the screen on one trial, and in the bottom-right on the next. There is no *a priori* long-term memory that links these objects to those locations, yet attention is directed there. Some temporary memory therefore seems necessary, a memory that links the type representations to a "token" representation of the specific instance of an object in a spatiotemporal world (also referred to as object files, indices, or deictic pointers; Kanwisher, 1987; Kahneman et al., 1992; Pylyshyn, 2001; Spivey et al., 2004; Hoover and Richardson, 2008).

The nature of this temporary memory is subject to debate. Some refer to it as being "episodic" (e.g., Altmann, 2004; Altmann and Kamide, 2007), but that obviously says little about its exact nature. The field will need to answer questions such as whether the binding of linguistic types to visual tokens is an implicit process, occurring automatically, without much cognitive control and/or awareness, or an explicit process, relying on the awareness of the stimuli involved, and therefore subject to cognitive control but also capacity limitations. Implicit representations are more likely to last for a longer period, while shorter term explicit memories are more subject to interference. Naturally, both types of memory may contribute to visual–linguistic interactions. An implicit memory is most clearly advocated by Altmann and colleagues (Altmann and Kamide, 2007; Altmann and Mirkovic, 2009), who argue that visual world type interactions are inevitable given the automatic spread of activation within a conglomerate of linguistic and visual representations. As we have argued above, such an account could work if the sprawl of activity can be channeled back to the original source – something that can be achieved through gating the feedback signal with the feedforward signal between layers of representation (van der Velde and de Kamps, 2001). Another argument for an implicit mechanism is that visual world interactions occur even though the visual and spoken input are often irrelevant to the observer (i.e., there is no explicit physical task), suggesting a substantial automatic component.

Others have advocated an important role for an explicit type of memory, most notably working memory (Spivey et al., 2004; Knoeferle and Crocker, 2007; Huettig et al., 2011a). The fact that visual world effects occur despite the absence of a clear task does not preclude such a contribution. After all, participants are at the very least instructed to "just" look at the display and "just" listen to the input, which may facilitate at least a partial entrance into working memory. One reason for assuming this type of memory comes from visual attention studies that suggest that the number of visual tokens or indices that can be simultaneously maintained is limited to four – a limit assumed to be the limit of visual working memory (Cowan, 2001). If visual world interactions depend on such tokens, they would thus also depend on visual working memory. But also on the psycholinguistic side, it has been argued that working memory is a real prerequisite for disambiguating and understanding language (Jackendoff, 2002, see also Marcus, 1998, 2001). It remains to be tested whether visual world effects are also subject to a limit of four visual objects and how they respond to different forms of cognitive load.

One advantage of the explicit memory account is that it allows the cognitive system to flexibly juggle the maintenance of visual memories between the internal and external world. As long as a visual stimulus is present, in principle it suffices to have only a minimal visual memory representation of them. Instead, the indices or pointers can be used to refer to the location of the object, allowing the cognitive system to only retrieve detailed percepts when necessary. This way the world serves as an outside memory, limiting the load on the cognitive system (O'Regan, 1992; O'Regan and Noë, 2001; Spivey et al., 2004). This would mean that the spatial pointers as alluded to when explaining visual world type effects are not just side effects of a memory system that cannot help but bind all sorts of information, but actually have a functional role in establishing the memory in the first place by directly referring to the outside world (a reference that then may be sustained even if the outside scene has been removed). A study by Wolfe et al. (2000) is directly relevant here. In some of their experiments, they presented observers with a visual search display that remained constantly on screen from trial to trial. The specific target changed from trial to trial (through an instruction in the center of the screen). For example, the search display might always consist of a red circle, a green square, a red triangle, and a blue diamond – all continuously present in the same position. On the first trial the target may then be a green square, whereas on the next it may be the red circle, and so on. Remarkably, even though the search display remained constant from trial to trial, search hardly improved. Even after 300 trials there was no notable improvement in search. Wolfe et al. (2000) concluded that no memory of the display was built up, despite countless inspections. They argued that for the lazy cognitive system, learning the display was unnecessary, since the stimulus remained visible and could be used as an outside memory. In contrast, when the search display was taken away after the first presentation, performance rapidly became fast and efficient. Now observers were forced to commit the items to internal memory, making them more rapidly available for selection. This flexibility (as induced by task demands) suggests some form of working memory, but it remains to be seen whether visual world interactions are equally flexible.

That working memory content *can* guide visual attention has been shown in several studies now (Soto et al., 2005; Olivers et al., 2006; Soto and Humphreys, 2007; Olivers, 2009). In these studies, observers are asked to look for a simple visual shape target among distracters, while keeping an unrelated object in working memory. However, one of the search distracters can match the memorized object (e.g., in color), and when it does, search suffers. It appears that an object that matches the contents of working memory captures attention, something which has been confirmed with eye movement measures. Of course, the fact that working memory can affect attentional guidance does not necessarily mean that it also does so in visual world settings. This remains to be investigated (see Huettig et al., 2011a, for a more detailed review).

INDIVIDUAL AND GROUP DIFFERENCES

The vast majority of studies investigating language-mediated eye gaze have been conducted with undergraduate students. This is of course not only the case for studies using the visual world and visuals search paradigms but a pervasive problem in experimental psychology more generally (see Arnett, 2008). It is an open empirical question how much one can generalize from the sophisticated behavior of highly educated university students to draw general inferences about mind and behavior beyond these narrow samples. Indeed it has been argued that the homogeneous Western student participants used in most studies are the "weirdest" (Western Educated Industrialized Rich Democratic) people in the world and the least representative populations one can find to draw general conclusions about human behavior (Henrich et al., 2010, for further discussion). Besides the theoretical challenges discussed above, there are thus some empirical challenges, which research on the interaction of the cognitive systems involved in language, vision, attention, and memory, must address. One promising line of inquiry will be the investigation of individual differences (see McMurray et al., 2010, for an example). Another approach, and one we shall discuss here in more detail, are studies with distinct non-student participant populations. Recent studies investigating language-mediated visual orienting in young children and in individuals with little formal schooling (i.e., low literacy levels) suggest that this approach may prove to be particularly fruitful.

There is the possibility that the mapping between spoken words and visual objects is mediated by stored verbal labels. Consider the color effects reported by Huettig and Altmann (2004, 2011). On hearing target words that are associated with a prototypical color (e.g., "*frog*"), participants tend to look at objects displayed in that color even though the depicted objects (e.g., a green blouse) are not themselves associated with that prototypical color (see Johnson and Huettig, 2011, for a similar results with 36-month-olds). But when listeners hear the word "*frog*," do they access an associated stored color label (GREEN), which makes them more likely to look at green things in their visual surroundings? Or, alternatively, do listeners on hearing "*frog*" access a target template, a sort of veridical perceptual description of the target (including its color) which then leads to a match with items matching this "perceptual" template (as tends to be assumed in visual search studies)? Note that verbal mediation is a genuine possibility; participants in free word association tasks typically produce the answer "green" when asked to write down the first word that comes to mind when thinking about "frog" (Nelson et al., 1998). Davidoff and Mitchell (1993) for instance have argued that "3-year-olds have more difficulty matching object colors with mental templates than they do with color naming" (p. 133) based on the finding that their 3-year-old participants tended to successfully judge that a banana is colored yellow in a verbal task but failed to choose the yellow banana as the correct one from differently colored bananas. Moreover, developmental psychologists have argued that "early in life, sensory, and linguistic color knowledge seem to coexist, but a proper map connecting names and perception is late in developing" (p. 78, Bornstein, 1985).

To examine this issue, Johnson et al. (2011) tested 48 two-year-olds who lacked reliable color term knowledge and found that on hearing the spoken target words they looked significantly more at the objects that were either color-related or semantically related to the named absent targets (e.g., on hearing "frog" they were more likely to look at a green truck and a bird than completely unrelated objects). Interestingly, there was a clear dissociation: words such as "*frog*" resulted in shifts in eye gaze to green things but color

words such as "*green*" did not. Thus, 2-year-olds look to color-matched competitors even if they do not know the label for that color. The Johnson et al. (2011) results do not rule out that adults have both direct and indirect routes linking color knowledge of words. What the Johnson et al. (2011) results suggest, however, is that the direct perceptual route exists before the indirect, lexically mediated route, has had a chance to develop.

Recent research involving adult individuals with little formal schooling also provides new insights with regard to the mechanisms and representations during language-mediated visual orienting. Studies using the blank screen paradigm (Spivey and Geng, 2001; Altmann, 2004), in which participants preview a visual scene and then listen to a spoken sentence while a *blank* screen is shown, have found that people have a tendency to re-fixate the regions on the blank screen that were previously occupied by relevant objects. Strong claims have been made regarding the nature of these "looking at nothing" effects. Altmann (2004, cf. Richardson and Spivey, 2000) has proposed that "the spatial pointers are a component of the episodic trace associated with each item – activating that trace necessarily activates the (experiential) component encoding the location of that item, and it is this component that automatically drives the eyes toward that location" (p. B86). Similarly, Ferreira et al. (2008) claimed that "whether the looks are intentional or are unconsciously triggered, the conclusion is the same: looking at nothing is an entirely expected consequence of human cognitive architecture" (p. 409).

However, Mishra et al. (2011) have found that this is not a universal trait of human cognition. Mishra et al. (2011) studied Indian low literates (2 mean years of formal schooling, but proficient speakers/listeners) and high literates (15 mean years of formal schooling) on the same "look and listen" task as used by Altmann (2004) to test these claims. If "looking at nothing" is an automatic reflex of the cognitive system to refer to previously presented visual objects, then it should be present in all proficient speakers/listeners regardless of their level of formal schooling. High and low literates were presented with a visual display of four objects (a semantic competitor, e.g., "kachuwa," turtle, and three distractors) for 5 s. Then the visual display was replaced with a blank screen and participants listened to simple spoken sentences containing a target word (e.g., "magar," crocodile, a semantic competitor of "kachuwa," turtle). High but not low literates looked at the empty region previously occupied by the semantic competitor as the spoken target word was heard. In a follow-up experiment, the same participants were presented with the identical materials except that the visual display (containing the semantic competitor and the distractors) was present as participants heard the spoken sentences. With such a set up both low literates and high literates did shift their eye gaze toward the semantic competitors immediately as the target word was heard. In another study, Huettig et al. (2011d), found that low literates also made fewer anticipatory eye movements than high literates. Low and high literates (2 and 12 years of schooling) listened to simple spoken sentences containing a target word (e.g., "door") while looking at a visual display of four objects (the target, i.e., the door, and three distractors). The spoken Hindi sentences contained adjectives followed by the (semantically neutral) particle wala/wali and a noun (e.g., "Abhi aap ek uncha wala darwaja dekhnge," Right now you are going to see a high door). Adjective (e.g., uncha/unchi, high) and particle (wala/wali) are gender-marked in Hindi and thus participants could use syntactic information to predict the target. To maximize the likelihood to observe anticipation effects, adjectives which were also semantically and associatively related to the target object were chosen. High literates started to shift their eye gaze to the target object well before target word onset. Low literates' fixations on the targets only started to differ from looks on the unrelated distractors once the spoken target word acoustically unfolded (more than a second later than the high literates).

Further research is currently underway to establish why these populations differ in language-mediated eye movement behavior (see also Huettig et al., 2011c). We know from control tests that they do not depend on IQ. The results are also unlikely to be due to differences in processing 2D information during picture processing. In a recent study we observed very high picture naming accuracy scores in the low literate group. Moreover, in Experiment 2 of Huettig et al. (2011d), low literates were not slower than high literates in their shifts in eye gaze to the target objects *when hearing the target word*, they just did not use contextual information to predict them before the target word was heard. This makes it very unlikely that the observed pattern of results is due to slow information retrieval during picture processing. Instead, we conjecture that literacy is a main factor underlying differences in language-mediated anticipation. To maintain a high reading speed, prediction is helpful if not necessary. Reading and spoken language comprehension, for instance, differ in the amount of information that is processed per time unit (approx. 250 vs. 150 words/min). It has also been observed that readers make use of statistical knowledge in the form of transitional probabilities, i.e., that the occurrence of one word can be predicted from the occurrence of another (McDonald and Shillcock, 2003). Low levels of reading and writing practice greatly decreases the exposure to such word-to-word contingency statistics in low literates. Huettig et al. (2011d) propose that formal literacy may enhance individuals' abilities to generate lexical predictions, abilities that help literates to exploit contextually relevant predictive information in other situations such as when anticipating which object an interlocutor will refer to next in one's visual environment.

In terms of the absence of looks to the semantic competitors by the low literates in the "blank screen" study it is less clear how literacy may have mediated these results. An intriguing possibility is that the well-known "looking at nothing" effects (Spivey and Geng, 2001; Altmann, 2004) reflect merely that participants with high levels of formal education are more familiar with the concept of experimentation and attempt to link "explicitly" the previewed visual display and the unfolding spoken sentence when viewing the blank screen and that low literates are much less likely to do so. A related possibility is that high literates may simply be better in correctly guessing the "purpose" of "blank screen" experiments. Alternatively, it may be that working memory differences underlie the differences between high and low literates' "looking at nothing" behavior. In any case, these results underscore the need to investigate the behavior of non-student participant populations. Ongoing research also examines the attentional basis of these differences between low and high literates. What seems

clear from these data is that the language–vision interaction is modulated by cognitive factors which correlate with formal literacy and/or general schooling and thus accounts which assume that this language-mediated eye movement behavior is automatic or a non-trivial consequence of human cognitive architecture may have to be revised.

FUTURE DIRECTIONS AND CONCLUSION

How will we be most likely to make progress in our understanding of the mechanisms and representations shared by language, vision, attention, and memory during language-mediated eye gaze? Besides a focus on individual and group differences, neuroscientific approaches will undoubtedly prove to be important. For example, activity in different brain areas may reveal at what level linguistic and visual input map onto each other (ranging from occipital to temporal areas), how this is translated into a saccadic signal (ranging from parietal areas to the frontal eye fields, as well as subcortical areas such as the superior colliculus), and to what extent systems are involved that are typically associated with top-down attention and working memory (such as the dorsolateral prefrontal cortex).

Computational modeling will also increasingly play an important role (see Allopenna et al., 1998; Roy and Mukherjee, 2005; Mayberry et al., 2009; Mirman and Magnuson, 2009; Stephen et al., 2009; McMurray et al., 2010; Kukona and Tabor, 2011). An advantage of such models is that theoretical notions and representations underlying language-mediated eye gaze are explicitly exposed. They also allow direct manipulation of representations, processes, and specific factors (e.g., past experience, age of acquisition) which are difficult to control in real participants. In addition, novel predictions about human performance can be derived since models often produce output phenomena which have not been reported previously.

A further fruitful avenue of research is the investigation of brain lesions using single case studies, studies involving groups of patients, or the application of transcranial magnetic stimulation (TMS) on healthy participants. Patients suffering from Balint's syndrome, for instance, have brain damage to the left and right parietal lobes and severe spatial deficits. One particularly interesting symptom is the difficulty that these patients appear to have with the binding of different visual features of an object (e.g., color and shape, cf. Friedman-Hill et al., 1995). One question is whether this type of lesion would also affect the binding of linguistic information to visual locations, as in the visual world paradigm, or whether linguistic information escapes the disintegration that characterizes the visual features.

In sum, we conclude that the investigation of language-mediated eye gaze is a useful approach to study the interaction of linguistic and non-linguistic cognitive processes. The data reviewed suggest that the representational level at which language–vision mapping occurs is co-determined by the timing of cascaded processing in the spoken word and object/visual word recognition systems, by the temporal unfolding of the spoken language, and by the nature of the visual environment (e.g., the characteristics of the visual stimuli, and the possibility of other representational matches). The most concrete proposal regarding attentional mechanisms to date postulates that language-mediated visual orienting arises because linguistic and non-linguistic information and attention are instantiated in the same common coding substrate. We suggest that local microcircuitries involving feedforward and feedback loops may instantiate such a representational substrate. We further conclude that little is currently known about the exact nature of the types of memory involved. Questions that remain to be answered include whether the binding of linguistic types to visual tokens is an implicit or an explicit process, occurs automatically or is subject to cognitive control, whether it is restricted by capacity limitations, and to what extent it suffers from interference and decay. We conjecture that an explicit working memory will be central to explaining interactions between language and visual attention. Though much progress has been made it is clear that a synthesis of further experimental evidence from a variety of fields of inquiry, methods, and distinct participant populations will prove to be crucial for our understanding about how language, vision, attention, and memory interact.

ACKNOWLEDGMENTS

Support was provided by the Max Planck Society (Falk Huettig), a grant from the Department of Science and Technology, India (Ramesh Kumar Mishra), and a VIDI grant from NWO (awarded to Christian N. L. Olivers).

REFERENCES

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *J. Mem. Lang.* 38, 419–439.

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: the "blank screen paradigm." *Cognition* 93, 79–87.

Altmann, G. T. M., and Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: linking anticipatory (and other) eye movements to linguistic processing. *J. Mem. Lang.* 57, 502–518.

Altmann, G. T. M., and Mirkovic, J. (2009). Incrementality and prediction in human sentence processing. *Cogn. Sci.* 33, 583–609.

Arnett, J. (2008). The neglected 95%: why American psychology needs to become less American. *Am. Psychol.* 63, 602–614.

Bornstein, M. H. (1985). On the development of color naming in young children: data and theory. *Brain Lang.* 26, 72–93.

Cave, K. R. (1999). The feature gate model of visual attention. *Psychol. Res.* 62, 182–194.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: a new methodology for the real-time investigation of speech perception, memory, and language processing. *Cogn. Psychol.* 6, 84–107.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* 24, 87–185.

Cree, G. S., and McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *J. Exp. Psychol. Gen.* 132, 163–201.

Dahan, D., and Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: conceptually mediated eye movements during spoken-word recognition. *Psychon. Bull. Rev.* 12, 453–459.

Davidoff, J., and Mitchell, D. (1993). The color cognition of children. *Cognition* 48, 121–137.

Dell'Acqua, R., and Grainger, J. (1999). Unconscious semantic priming form pictures. *Cognition* 73, B1–B15.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222.

Dunabeitia, J. A., Aviles, A., Afonso, O., Scheepers, C., and Carreiras, M. (2009). Qualitative differences in the representation of abstract versus concrete words: evidence from the visual-world paradigm. *Cognition* 110, 284–292.

Ferreira, F., Apel, J., and Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends Cogn. Sci. (Regul. Ed.)* 12, 405–410.

Friedman-Hill, S. R., Robertson, L. C., and Treisman, A. (1995). Parietal contributions to visual feature binding: evidence from a patient with bilateral lesions. *Science* 269, 853–855.

Gaskell, M. G., and Marslen-Wilson, W. D. (1997). Integrating form and meaning: a distributed model of speech perception. *Lang. Cogn. Process.* 12, 613–656.

Hamker, F. H. (2004). A dynamic model of how feature cues guide spatial attention. *Vision Res.* 44, 501–521.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behav. Brain Sci.* 33, 61–135.

Hoover, M. A., and Richardson, D. C. (2008). When facts go down the rabbit hole: contrasting features and object hood as indexes to memory. *Cognition* 108, 533–542.

Huettig, F., and Altmann, G. (2011). Looking at anything that is green when hearing "frog": how object surface colour and stored object colour knowledge influence language-mediated overt attention. *Q. J. Exp. Psychol.* 64, 122–145.

Huettig, F., and Altmann, G. T. M. (2004). "The online processing of ambiguous and unambiguous words in context: evidence from head-mounted eye-tracking," in *The Online Study of Sentence Comprehension: Eyetracking, ERP and Beyond*, eds M. Carreiras and C. Clifton (New York, NY: Psychology Press), 187–207.

Huettig, F., and Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. *Cognition* 96, B23–B32.

Huettig, F., and Altmann, G. T. M. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Vis. Cogn.* 15, 985–1018.

Huettig, F., and McQueen, J. M. (2007). The tug of war between phonological, semantic, and shape information in language-mediated visual search. *J. Mem. Lang.* 54, 460–482.

Huettig, F., and McQueen, J. M. (2011). The nature of the visual environment induces implicit biases during language-mediated visual search. *Mem. Cognit.* 39, 1068–1084.

Huettig, F., Olivers, C. N. L., and Hartsuiker, R. J. (2011a). Looking, language, and memory: bridging research from the visual world and visual search paradigms. *Acta Psychol. (Amst.)* 137, 138–150.

Huettig, F., Rommers, J., and Meyer, A. S. (2011b). Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychol. (Amst.)* 137, 151–171.

Huettig, F., Singh, N., and Mishra, R. K. (2011c). Language-mediated visual orienting behavior in low and high literates. *Front. Psychol.* 2:285. doi:10.3389/fpsyg.2011.00285

Huettig, F., Singh, N., Singh, S., and Mishra, R. K. (2011d). "Language-mediated prediction is related to reading ability and formal literacy," in *Paper Presented at the AMLaP 2011 Conference in Paris*, Paris.

Huettig, F., Quinlan, P. T., McDonald, S. A., and Altmann, G. T. M. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychol. (Amst.)* 121, 65–80.

Humphreys, G. W., and Müller, H. J. (1993). SEarch via recursive rejection (SERR): a connectionist model of visual search. *Cogn. Psychol.* 25, 43–110.

Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* 40, 1489–1506.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution.* New York: Oxford University Press.

Johnson, E., McQueen, J. M., and Huettig, F. (2011). Toddlers' language-mediated visual search: they need not have the words for it. *Q. J. Exp. Psychol.* 64, 1672–1682.

Johnson, E. K., and Huettig, F. (2011). Eye movements during language-mediated visual search reveal a strong link between overt visual attention and lexical processing in 36-month-olds. *Psychol. Res.* 75, 35–42.

Kahneman, D., Treisman, A., and Gibbs, B. (1992). The reviewing of object files: object-specific integration of information. *Cogn. Psychol.* 24, 175–219.

Kanwisher, N. (1987). Repetition blindness: type recognition without token individuation. *Cognition* 27, 117–143.

Knoeferle, P., and Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: evidence from eye-movements. *J. Mem. Lang.* 57, 519–543.

Konkle, T., Brady, T. F., Alvarez, G. A., and Oliva, A. (2010a). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *J. Exp. Psychol. Gen.* 139, 558–578.

Konkle, T., Brady, T. F., Alvarez, G. A., and Oliva, A. (2010b). Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psychol. Sci.* 21, 1551–1556.

Kukona, A., and Tabor, W. (2011). Impulse processing: a dynamical systems model of the visual world paradigm. *Cogn. Sci.* 35, 1009–1051.

Lamme, V. A. F., and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579.

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychol. Rev.* 104, 211–240.

Mani, N., and Plunkett, K. (2010). In the infant's mind's ear: evidence for implicit naming in 18-month-olds. *Psychol. Sci.* 21, 908–913.

Marcus, G. (1998). Rethinking eliminative connectionism. *Cogn. Psychol.* 37, 243–282.

Marcus, G. (2001). *The Algebraic Mind.* Cambridge, MA: MIT Press.

Marslen-Wilson, W., and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cogn. Psychol.* 10, 29–63.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition* 25, 71–102.

Mayberry, M., Crocker, M. W., and Knoeferle, P. (2009). Learning to attend: a connectionist model of the coordinated interplay of utterance, visual context, and world knowledge. *Cogn. Sci.* 33, 449–496.

McClelland, J. L., and Elman, J. L. (1986). The TRACE model of speech perception. *Cogn. Psychol.* 18, 1–86.

McDonald, S. A. (2000). *Environmental Determinants of Lexical Processing Effort.* Unpublished doctoral dissertation, University of Edinburgh, Scotland. Available at: http://www.inf.ed.ac.uk/publications/thesis/online/IP000007.pdf [accessed December 10, 2004].

McDonald, S. A., and Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities. *Psychol. Sci.* 14, 648–652.

McMurray, B., Samelson, V. M., Lee, S. H., and Tomblin, J. B. (2010). Individual differences in online spoken word recognition: implications for SLI. *Cogn. Psychol.* 60, 1–39.

McMurray, B., Tanenhaus, M., and Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition* 86, B33–B42.

Meyer, A. S., Belke, E., Telling, A. L., and Humphreys, G. W. (2007). Early activation of object names in visual search. *Psychon. Bull. Rev.* 14, 710–716.

Meyer, A. S., and Damian, M. F. (2007). Activation of distractor names in the picture-picture interference paradigm. *Mem. Cognit.* 35, 494–503.

Miller, E. K., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202.

Mirman, D., and Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Mem. Cognit.* 37, 1026–1039.

Mishra, R. K., and Marmolejo-Ramos, F. (2010). On the mental representations originating during the interaction between language and vision. *Cogn. Process.* 11, 295–305.

Mishra, R. K., Singh, N., and Huettig, F. (2011). "'Looking at nothing' is neither automatic nor an inevitable consequence of human cognitive architecture," in *Paper Presented at the AMLaP 2011 Conference in Paris,* Paris.

Moores, E., Laiti, L., and Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. *Nat. Neurosci.* 6, 182–189.

Morsella, E., and Miozzo, M. (2002). Evidence for a cascade model of lexical access in speech production. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 555–563.

Moss, H. E., McCormick, S. F., and Tyler, L. K. (1997). The time-course of semantic activation during spoken word recognition. *Lang. Cogn. Process.* 12, 695–731.

Myachykov, A., Thompson, D., Scheepers, C., and Garrod, S. (2011). Visual attention and structural choice in

sentence production across languages. *Lang. Linguist. Compass* 5, 95–107.

Navarette, E., and Costa, A. (2005). Phonological activation of ignored pictures: further evidence for a cascade model of lexical access. *J. Mem. Lang.* 53, 359–377.

Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (1998). *The University of South Florida Word Association, Rhyme, and Word Fragment Norms.* Available at: http://www.usf.edu/FreeAssociation/

Noizet, G., and Pynte, J. (1976). Implicit labeling and readiness for pronunciation during the perceptual process. *Perception* 5, 217–223.

Olivers, C. N. L. (2009). What drives memory-driven attentional capture? The effects of memory type, display type, and search type. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1275–1291.

Olivers, C. N. L. (2011). Long-term visual associations affect attentional guidance. *Acta Psychol. (Amst.)* 137, 243–247.

Olivers, C. N. L., Meijer, F., and Theeuwes, J. (2006). Feature-based memory-driven attentional capture: visual working memory content affects visual attention. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 1243–1265.

O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: the world as an outside memory. *Can. J. Psychol.* 46, 461–488.

O'Regan, J. K., and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behav. Brain Sci.* 24, 939–1011.

Palmer, J., Verghese, P., and Pavel, M. (2000). The psychophysics of visual search. *Vision Res.* 40, 1227–1268.

Posner, M. I. (1980). Orienting of attention, the VIIth Sir Frederic Bartlett Lecture. *Q. J. Exp. Psychol.* 32, 3–25.

Pylyshyn, Z. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition* 80, 127–158.

Reynolds, J. H., and Desimone, R. (1999). The role of neural mechanisms of attention in solving the binding problem. *Neuron* 24, 19–29.

Richardson, D. C., and Spivey, M. J. (2000). Representation, space and hollywood squares: looking at things that aren't there anymore. *Cognition* 76, 269–295.

Roy, D., and Mukherjee, N. (2005). Towards situated speech understanding: visual context priming of language models. *Comput. Speech Lang.* 19, 227–248.

Salverda, A. P., Brown, M., and Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual-world studies. *Acta Psychol. (Amst.)* 137, 172–180.

Salverda, A. P., Dahan, D., and McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition* 90, 51–89.

Salverda, A. P., and Tanenhaus, M. K. (2010). Tracking the time course of orthographic information in spoken-word recognition. *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 1108–1117.

Schreuder, R., Flores D'Arcais, G. B., and Glazenborg, G. (1984). Effects of perceptual and conceptual similarity in semantic priming. *Psychol. Res.* 45, 339–354.

Shatzman, K. B., and McQueen, J. M. (2006). Prosodic knowledge affects the recognition of newly acquired words. *Psychol. Sci.* 17, 372–377.

Soto, D., Heinke, D., Humphreys, G. W., and Blanco, M. J. (2005). Early, involuntary top-down guidance of attention from working memory. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 248–261.

Soto, D., and Humphreys, G. W. (2007). Automatic guidance of visual attention from verbal working memory. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 730–757.

Spivey, M., and Geng, J. (2001). Oculomotor mechanisms activated by imagery and memory: eye movements to absent objects. *Psychol. Res.* 65, 235–241.

Spivey, M. J., Richardson, D. C., and Fitneva, S. A. (2004). "Memory outside of the brain: oculomotor indexes to visual and linguistic information," in *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*, eds J. Henderson and F. Ferreira (New York: Psychology Press), 161–189.

Stephen, D. G., Mirman, D., Magnuson, J. S., and Dixon, J. A. (2009). Lévy-like diffusion in eye movements during spoken-language comprehension. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 79, 056114.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634.

Theeuwes, J. (2010), Top-down and bottom-up control of visual selection. *Acta Psychol. (Amst.)* 135, 77–99.

Treisman, A. (1996). The binding problem. *Curr. Opin. Neurobiol.* 6, 171–178.

Treisman, A., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136.

Treisman, A., and Sato, S. (1990). Conjunction search revisited. *J. Exp. Psychol. Hum. Percept. Perform.*16, 459-478.

van der Velde, F., and de Kamps, M. (2001). From knowing what to knowing where: modeling object-based attention with feedback disinhibition of activation. *J. Cogn. Neurosci.* 13, 479–491.

Vanduffel, W., Ekstrom, L. B., Roelfsema, P. R., Arsenault, J. T., and Bonmassar, G. (2008). Bottom-up dependent gating of frontal signals in early visual cortex. *Science* 321, 414–417.

Vickery, T. J., King, L.-W., and Jiang, Y. (2005). Setting up the target template in visual search. *J. Vis.* 5, 81–92.

Wolfe, J. M. (1994). Guided Search 2.0. A revised model of visual search. *Psychon. Bull. Rev.* 1, 202–238.

Wolfe, J. M. (1998). "Visual search," in *Attention,* ed. H. Pashler (Hove: Psychological Press), 14–73.

Wolfe, J. M., Horowitz, T. S., Kenner, N., Hyle, M., and Vasan, N. (2004). How fast can you change your mind? *Vision Res.* 44, 1411–1426.

Wolfe, J. M., Klempen, N., and Dahlen, K. (2000). Post attentive vision. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 693–716.

Yee, E., Huffstetler, S., and Thompson-Schill, S. L. (2011). Function follows form: activation of shape and function features during object identification. *J. Exp. Psychol. Gen.* 140, 348–363.

Yee, E., Overton, E., and Thompson-Schill, S. L. (2009). Looking for meaning: eye movements are sensitive to overlapping semantic features, not association. *Psychon. Bull. Rev.* 16, 869–874.

Yee, E., and Sedivy, J. C. (2001). "Using eye movements to track the spread of semantic activation during spoken word recognition," in *Paper Presented at the 13th Annual CUNY Sentence Processing Conference*, Philadelphia.

Yee, E., and Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 1–14.

Zelinsky, G. J., and Murphy, G. L. (2000). Synchronizing visual and language processing: an effect of object name length on eye movements. *Psychol. Sci.* 11, 125–131.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 August 2011; paper pending published: 13 September 2011; accepted: 20 December 2011; published online: 09 January 2012.

Citation: Huettig F, Mishra RK and Olivers CNL (2012) Mechanisms and representations of language-mediated visual attention. Front. Psychology ***2****:394. doi: 10.3389/fpsyg.2011.00394*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.