

How speech processing affects our attention to visually similar objects: Shape competitor effects and the visual world paradigm

Falk Huettig (f.huettig@psych.york.ac.uk)
M. Gareth Gaskell (g.gaskell@psych.york.ac.uk)
Philip T. Quinlan (p.quinlan@psych.york.ac.uk)
Department of Psychology, University of York
York, YO10 5DD, United Kingdom

Abstract

It was investigated how spoken language is mapped onto the mental representations of objects in the visual field. Specifically, the visual world paradigm was used to test the hypothesis that during 'passive' listening tasks attention is directed more towards objects in the visual field that match the physical shape of the concept of the word concurrently heard than towards objects that do not match on physical shape. Participants listened to sentences containing certain critical target words of concepts with a typical shape (e.g. 'snake') while concurrently viewing a visual display of four objects. We found that participants tended to fixate conceptually unrelated objects with a similar physical shape (e.g. cable) as soon as information from the target word (e.g. 'snake') started to acoustically unfold. The results indicate that (contrary to some priming studies, e.g. Moss et al., 1997) shape information is accessed long before the offset of the spoken word. We discuss the findings with respect to the applicability of the visual world paradigm for the investigation of the access of lexical representations and theories of active vision.

Humans 'translate' between spoken language and concurrent visual input in such a natural way that we are hardly ever consciously aware of the processes involved. Surprisingly, there has been little research that has attempted to explore explicitly the interaction of these processes. A notable exception has been the research within the 'visual world paradigm' (the measuring of eye movements around a visual scene or display of objects in response to concurrent speech: Cooper, 1974; Tanenhaus et al. 1995) using both linguistic and visual contexts. Cooper (1974) established that when participants were presented simultaneously with spoken language and a visual field containing referents of the spoken words, participants tended to spontaneously fixate the visual referents of the words currently being heard. For example, participants were more likely to fixate the picture of a snake when hearing part or the entire word 'snake' than to fixate pictures of unrelated control words. Cooper (1974) also found that participants were more likely to fixate pictures showing a lion, a zebra, or a snake when hearing the semantically related word 'Africa' than to fixate semantically unrelated control words. Cooper's (1974) early study thus established two main findings: first, during the acoustic duration of a spoken word participants show a strong tendency to fixate objects that

the word refers to. Second, his study highlighted the influence of semantic relationships on language-mediated fixation behavior: participants are more likely to fixate a visual referent that has some semantic relationship with the word heard than a semantically unrelated visual referent (see Huettig & Altmann, 2004; Yee & Sedivy, 2001; for follow-up studies). The primary goal of research in the visual world paradigm following Cooper's (1974) pioneering study has been to use eye movements as a tool to shed light on linguistic processing. For example, Allopenna, et al. (1998) asked participants to 'Pick up the candy. Now put it ...' in the context of a visual display of objects including (among other things) a candy and a candle. They found evidence for a phonological competitor effect: eye movements to both the candy and the candle increased as the word 'candy' acoustically unfolded but that soon after its acoustic offset, looks to the candle decreased while looks to the candy continued to rise. The Allopenna et al. (1998) study provided evidence for a standard competitor effect as predicted by theories of auditory word recognition such as TRACE (e.g. McClelland & Elman, 1986). Importantly the study demonstrated this effect in real-time as the speech stream was unfolding acoustically (see also Dahan et al., 2001; for more evidence that the visual world paradigm provides fine-grained measures of lexical processing).

However, far less attention has focused on examining the interaction of spoken language with directed attention and the *visual* properties of the presented objects. In this regard, Cooper (1974) also found that participants tended to fixate a picture of a snake when hearing the word 'wormed' (in the context 'just as I had wormed my way on my stomach'). This finding (although not discussed by Cooper) suggests that there may also be a strong link between lexical processing and the visual properties of an object such as an object's shape (although it cannot be ruled out that in Cooper's experiment participants mistook the snake for a worm and therefore directed their attention to the picture of the snake when hearing 'wormed').

The visual world paradigm and the access of lexical representations Importantly, Cooper's study (1974) is indicative that similar processes to semantic priming are taking place when people map spoken words onto related visual objects. The semantic priming paradigm (Meyer & Schvaneveldt, 1971) has proven to be particularly useful for

the investigation of lexical representations. The currently dominant view is that a word's representation is composed of smaller units (or 'features') of different kinds that are accessed during spoken word recognition. Recent distributed models of spoken word recognition (e.g. Gaskell & Marslen-Wilson, 1997) assume that some aspects of a word's meaning may be activated more rapidly than others resulting in a dynamic pattern of changes in the semantic properties throughout the duration of the spoken word. Evidence supporting this notion comes, for example, from priming studies by Moss et al. (1997) that found a significant priming effect for functional properties of words early during the duration of the word but priming for perceptual targets (e.g. the shape overlap between *hook* and *curve*) only at the offset of the prime word. In order to investigate these time-course issues the visual world paradigm may be particularly useful because of the closely time-locked, fine-grained measures the method provides.

In the current study we explored how the physical shape of objects in a visual display interacts with language-directed attention. Participants' eye movements were measured, during the acoustic duration of certain target words (concepts with a typical shape, e.g. 'snake'), to conceptually unrelated visual objects that have a similar shape (e.g. the image of a cable). If participants shift their attention to conceptually unrelated objects with a similar shape (e.g. cable) when the word 'snake' unfolds during online speech, then the inspection of the time-course of fixation probabilities should shed light on the issue whether (lexical) shape information is accessed only at word offset or before. In other words, if (lexical) 'perceptual' information such as shape is not accessed before the offset of the spoken word then the prediction is that there should be no increased attention to shape competitors (e.g. cable) *before* the offset of the acoustic target words (e.g. 'snake').

The visual world paradigm and 'active vision' Our study explored the effect of 'shape overlap' between spoken words and visual objects on *overt* attention. Relevant in this regard is that most vision research has focused on what Findlay & Gilchrist (2003) term 'passive vision': the assumption that image interpretation is largely passive and that parallel processing occurs across the visual image with algorithms charting the "progress from a grey-scale retinal input to an internal representation in the head". Findlay & Gilchrist (2003) reject this view in favor of 'active vision': the notion that *overt* gaze orienting is an essential and crucial feature of vision. This approach emphasizes the importance of re-directing attention overtly (rather than covertly) by moving the gaze in order for the attended location to obtain the instant benefit of high-resolution foveal vision. These proposals are similar to the notion that the perceptual system offloads information by leaving it in the environment rather than just passively passing information on to the cognitive system for propositional representations to be created. According to this view, perceptual information in the environment is accessed when needed, with the visual world functioning as a kind of external memory (e.g.

O'Regan, 1992). Objects in this situated memory are represented in a spatial data structure which contains 'pointers' to the real-world location of the object. Thus, the system need not store internally detailed information about the object, but can instead locate that information, when it has to, by directing attention back to that object in the environment. Essentially, the focus in active vision research is on understanding why and when gaze is re-directed. In other words active vision places vision in a context. And one important variable that impacts on (and guides) active vision is spoken language.

Importantly, there was one second preview of the visual display in our study and the target words unfolded approx. five seconds after the onset of the visual display. This means that all four objects were fixated (usually several times) before the onset of the target word. Therefore the prediction is that on 'passive vision' accounts, arguably, there is no need for an *immediate* shift in overt attention towards a conceptually *unrelated* object (e.g. cable) when the word 'snake' unfolds. In other words on 'passive vision' accounts all relevant information has already been encoded and is available to the system for further cognitive processing. On active or situated vision accounts, however, the prediction is that on hearing the target word, overt attention will be re-directed to the shape competitor to retrieve more information about that object to establish its fit with the specification provided by the target word.

Methods

Participants 48 participants from the University of York student community took part in this study. All were native speakers of British English and had either uncorrected vision or wore soft contact lenses or glasses.

Materials The experiment made use of three conditions. 21 items were created consisting of two types of spoken sentences containing a target word (e.g. 'snake') for two sets of visual stimuli. In the 'target set' the visual stimulus was a picture depicting a fully matching referent (e.g. a snake) for the acoustic target word plus three distractors depicting objects from different conceptual categories. The stimuli in the 'shape competitor set' consisted of the same four pictures, in identical positions, except that the target picture (e.g. the snake) was replaced by a picture depicting an object with a similar shape as the target word (e.g. a cable, Figure 1).

The sentential stimuli were constructed for three conditions: neutral sentences with the pictures of the 'target set' (the neutral condition), sentences biasing the visual target referent (biasing snake) with the pictures of the 'target set' (the biasing condition), and the same sentences as in the biasing condition but where the picture of the target object (e.g. snake) was replaced by a picture of a shape competitor (the competitor condition, e.g. the picture of a cable). The rationale for presenting the shape competitor in a biasing context was simply to make it relatively unlikely that participants would anticipate, prior

to the target word, that the shape competitor would be the object of attention (even though it was not going to be referred to directly). The neutral sentences were included in order to establish a baseline against which the efficacy of the biasing context could be determined.

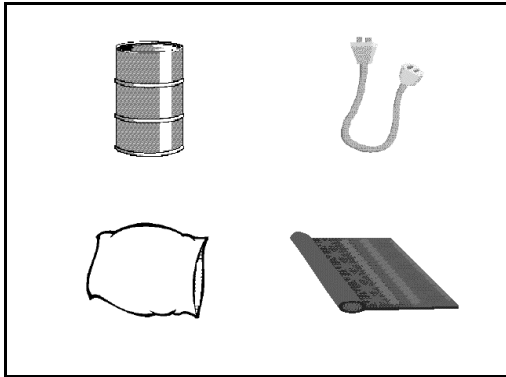


Figure 1. Visual stimulus in the competitor condition (depicting: physical shape competitor of the acoustic target word ‘snake’: cable, 3 distractors)

In sum, for the neutral condition, the sentence did not provide any contextual bias up until the target word that would favor any of the pictures depicted in the visual scene (*‘In the beginning, the man watched closely, but then he looked at the snake and realized that it was harmless’*). In the biasing condition, the sentence was constructed to contextually bias towards the depicted target object (e.g. the snake): *‘In the beginning, the zookeeper worried greatly, but then he looked at the snake and realized that it was harmless’*. In the competitor condition the sentence was identical to the biasing condition. However, the picture of the shape competitor (e.g. cable) was semantically unrelated to the target word (e.g. ‘snake’) and thus the sentential context did not provide any contextual bias towards the picture of the shape competitor. The target-competitor pairs were: anchor/arrow, apple/moon, banana/sword, bell/hat, button/coin, candle/tube, cigar/carrot, chimney/rocket, dice/ice cube, football/planet, globe/orange, horseshoe/magnet, lighthouse/flask, microphone/cone, mirror/frame, pencil/column, plate/wheel, racket/saucepan, scissors/chopsticks, snake/cable, wheelbarrow/sledge.

Norming study In order to determine the relative similarity in physical shape of the target concept activated by the acoustic target word with the depicted objects a norming study was conducted. Twelve participants provided normative data. Participants were presented with the written target word (e.g. *snake*) and the actual visual items. Participants were asked to judge how similar the typical physical shape of the target concept (*snake*) was with the physical shape of the depicted objects on a scale from 0 to 10 (zero representing: ‘absolutely no similarity in physical shape’, 10 representing: ‘identical in physical

shape’). The mean similarity for the shape competitors was 7.1 (SD = 1.8) and for the distractors 1.4 (SD = 0.7). These differences in the shape similarity judgments between the shape competitors and the visually dissimilar distractors were highly significant ($F_1(1, 11) = 268.89$, $MSE = 0.07$, $p < 0.01$; $F_2(1, 20) = 200.35$, $MSE = 0.17$, $p < 0.001$).

Procedure and Design There were 21 experimental items (counterbalanced across the three conditions). For 14 of the experimental items per participant the visual stimulus included a visual referent matching the full ‘target specification’ of the target word (e.g. the acoustic word ‘snake’ and the picture of a snake in the ‘neutral condition’ and the ‘biasing condition’). For 7 of the experimental items there was no picture matching the full ‘target specification’. For these items there was only a ‘physical shape match’ between the acoustic target word and the shape competitor picture (e.g. the acoustic target *snake* and the picture of a cable in the ‘shape competitor condition’). 15 additional filler items were added, which all included a fully matching visual referent of an acoustic target word. There were four practice trials. Thus 82% of the 40 trials included a *fully* matching target visual referent (e.g. hearing ‘snake’ and seeing a snake). This design made it very unlikely that the participants were able to note the physical shape relationship and adopt a conscious strategy accordingly. In addition, participants consistently stated in self-report that they neither moved their eyes according to some kind of explicit strategy nor noticed the ‘shape manipulation’.

Participants were seated at a comfortable distance in front of a 17” display and wore an SMI EyeLink head-mounted eye-tracker, sampling at 250Hz from the right eye (viewing was binocular). They were told that they should listen to the sentences carefully. They were also told that they could look at whatever they wanted but were asked not to take their eyes off the screen throughout the experiment. The onset of the visual stimulus was one second before the onset of the spoken stimulus. The onset of the acoustic target word was on average 4 seconds after the onset of the spoken sentence and thus the acoustic target word started to unfold on average 5 seconds after the onset of the visual stimulus. The entire experiment lasted approximately twenty minutes. Participants’ eye movements were recorded as they listened to the sentences.

Results

Fixation probabilities: $p(\text{fix})$ The probability of fixating a type of picture at a defined moment in time, $p(\text{fix})$, will be reported. The visual display consisted of four quadrants, each with one object. Gaze positions were categorized by the quadrant in which an object was depicted. The *a priori* probability of fixating one of the four pictures in absence of any bias was thus 0.25.

Table 1: $P(\text{fix})$ at the acoustic onsets and offsets of the target words (e.g. 'snake') and difference scores per condition

condition	neutral		biasing		competitor	
	target (snake)	distractor	target (snake)	distractor	competitor (cable)	distractor
$p(\text{fix})$ at onset	.27	.24	.39	.19	.25	.24
difference score at onset	.04		0.21		.01	
$p(\text{fix})$ at offset	.50	.15	.52	.14	.33	.22
difference score at offset	.35		.38		.11	

Table 1 shows $p(\text{fix})$ for the type of picture at the time points of critical interest: the acoustic onset of the target word (e.g. the acoustic onset of 'snake'), and the offset of the target word (e.g. the acoustic offset of 'snake'). The probability to fixate the three distractors was averaged to obtain one distractor value. Note that we did not add any time to account for the time it takes to program a saccade. All measures and analyses were based on the *real* acoustic time points. Figure 2 shows the time-course of $p(\text{fix})$ in the three conditions from the acoustic onset of the target word for 1000 ms.

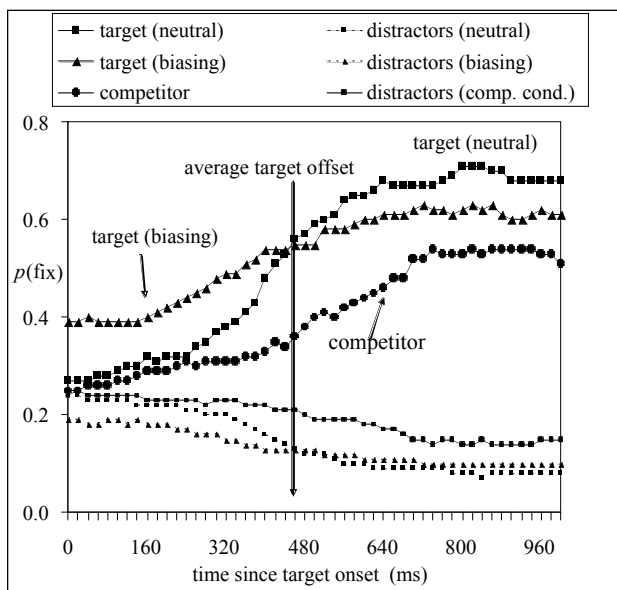


Figure 2. Time-course of $p(\text{fix})$ to the target in the neutral condition and the biasing condition, and to the shape competitor in the competitor condition (and averaged distractors of each condition).

Fixations of the different types of pictures at the acoustic onset of the target word is of interest in order to assess whether there were any biases in attention before information from the critical target word (e.g. 'snake') became available. It was predicted that at this point there would be no such bias in the neutral condition and the competitor condition if the context had been neutral with respect to directed attention to any of the pictures. However, it was predicted that at the acoustic onset of the

target word there would be a bias towards the target picture (e.g. the snake) in the biasing condition because of the biasing sentential context. These predictions are apparently born out by the data. Table 1 shows that $p(\text{fix})$ at the onset of the target word was around 0.25 in the neutral and competitor conditions but that there was a strong bias towards the target object in the biasing condition. The acoustic offset of the target word reflects the point when the entire spoken target word has been heard by the participants. Fixations at this point are of interest in order to assess whether the acoustic unfolding of the target word resulted in changes in overt attention. Table 1 and Figure 2 show that the target and competitor fixations had increased in the neutral and the competitor conditions, whereas in the biasing condition the probability to fixate the target increased further. Nonetheless, Table 1 and Figure 2 suggest that $p(\text{fix})$ of the targets in the neutral and the biasing conditions was higher than $p(\text{fix})$ of the shape competitors in the competitor competition. In other words as acoustic information from the target words became available the probability to fixate the target picture *and* the shape competitor picture increased. However, $p(\text{fix})$ of the target pictures increased much more than $p(\text{fix})$ of the shape competitors.

Statistical analyses In order not to violate statistical assumptions (in particular that pertaining to the independence of observations), difference scores obtained in each condition are compared. For instance to assess a bias to look at a critical picture (target or competitor vs. a distractor referent) the differences in fixation probabilities to these stimuli are considered. Such difference scores reveal both the magnitude and direction of the effects. In the current study $p(\text{fix distractor})$ was subtracted from $p(\text{fix target})$ and $p(\text{fix competitor})$. Any positive difference reveals a bias of looks towards the critical picture, a negative difference reveals a bias of looks towards the distractors, and difference scores close to zero reveals neither bias. The use of error bars in the form of 95% confidence intervals plotted around the sample means provides a quantitative visual representation of the faith that should be placed in the pattern of sample means as an estimate of corresponding patterns of population means. Figure 3 shows the mean of the difference scores for participants and items at the acoustic onset of the

target words (e.g. 'snake') in the three conditions. Error bars represent the 95% confidence intervals of the means computed individually for each difference score.

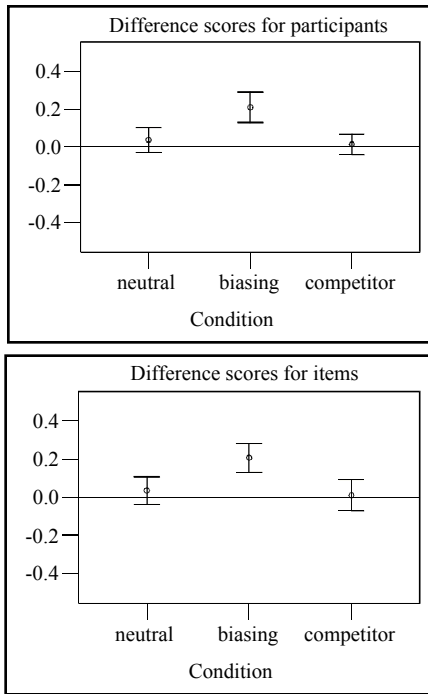


Figure 3. Means of the difference scores (participants and items) at the acoustic onset of the target words (Error bars represent the 95% confidence intervals)

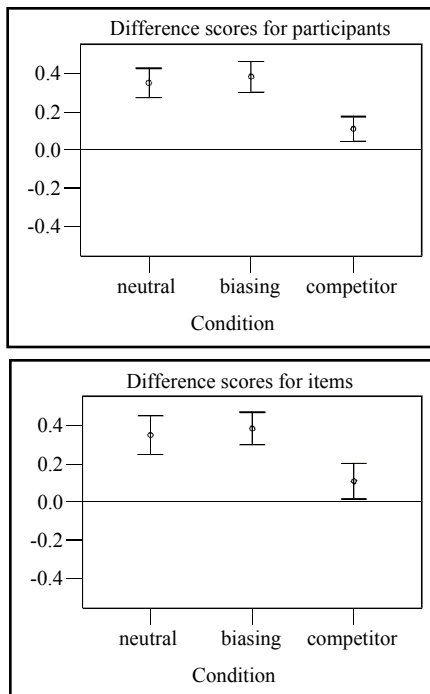


Figure 4. Means of the difference scores (participants and items) at the acoustic offset of the target words (Error bars represent the 95% confidence intervals)

Figure 3 shows that at target onset there were no reliable differences between looks to the critical pictures and the distractors in the neutral and the competitor conditions. In other words there were no differences in directed attention for type of picture at the acoustic onset of the target. However, there was a reliable bias in directed attention towards the target object in the biasing condition. Thus the contextual manipulation had been successful.

Figure 4 shows the mean of the difference scores for participants and items at the acoustic offset of the critical words. There was a reliable bias in directed attention to the critical pictures in all three conditions. Importantly, there was a statistically robust higher probability to fixate the shape competitor (e.g. the cable) than the distractors at the acoustic offset of the target word (e.g. 'snake').

General Discussion

Our findings directly link online conceptual processing during lexical access in speech to attentional behavior in the visual world. They extend Cooper's (1974) work by showing that during 'passive' listening tasks attention is also directed significantly more towards objects that match the shape of the word concurrently heard than towards objects that do not match on shape.

Importantly, the competitor effect was significant at the offset of the acoustic target words. This means that the shape competitor effect started to occur as soon as information from the target word acoustically unfolded given that the average duration of the target words was 447ms and that the *minimum* latency to program and initiate a saccade is 150 to 200 ms (e.g. Saslow, 1967). This result is contrary to priming studies (Moss et al., 1997) that found activation of perceptual (including shape) information only at the offset of the prime word. The current study suggests that the visual world paradigm is particularly sensitive for capturing the access of conceptual and perceptual information during lexical processing. Note that Moss et al. (1997) included five different types of 'perceptual' properties in their study. A reason for the discrepancy to our results thus may be that Moss et al. (1997) did not distinguish between different 'perceptual' properties such as color and shape. In other words they may have found delayed priming for perceptual targets because of the differential properties of the items they selected for their perceptual condition. An alternative explanation is that the information (from the spoken words) provided for the attentional system to visual objects involves such a tight 'loop' that other means of observing the access of (lexical) perceptual representations (e.g. the lexical decision task) can only do so at some delay. Similarly the activation in our study may have partly originated through spreading activation from shape information portrayed within the visual display. The shape information may have activated concepts sharing those features resulting in an earlier access of shape information during spoken word recognition. Our data do not show to what extent this activation may have originated from the visual display.

Notably, the present results are in line with the predictions derived from active or situated vision accounts that on hearing the target word, *overt* attention may be re-directed *immediately* towards only partly matching objects in order to retrieve more information and to establish their fit with the target specification as provided by the target word. Arguably, our findings coupled with the fact that there was one second preview of the visual display and that the target words unfolded approximately five seconds after the onset of the visual display, cannot be as *easily* incorporated in 'passive vision' accounts.

Pickering, McElree, & Garrod (submitted) have recently proposed that participants may engage in a covert naming strategy in visual world experiments. Pickering et al. state that "many effects in this paradigm may be partly the result of participants' regularly naming the objects covertly... further research is needed to determine the extent to which visual world results depend upon linguistic recoding (covert naming) of the objects". The current data do not rule out that our participants on occasion named an object covertly. However, the immediate and robust shift in directed attention to a conceptually unrelated and clearly identifiable shape competitor with a different name (e.g. cable) on hearing the target word (e.g. 'snake') does *not* fit comfortably with their suggestion. The current study thus casts doubt on the claim that participants *regularly* name objects in visual world studies.

The shape competitor effects are unlikely to be limited to the passive listening task we employed. Dahan & Tanenhaus (2002) recently presented evidence that similar visual form competitor effects also occur when participants are required to engage in an explicit physical task (moving the objects mentioned in spoken sentences using a computer mouse). Our procedure of a 'passive' listening task is strong evidence that these perceptual competitor effects are not limited to certain specific 'goal-directed' task demands.

In sum, the findings are best compatible with the notion of a rich mapping process between spoken language and concurrent visual input. On a methodological note, the visual world paradigm promises to be a valuable research tool for investigations into the access of (lexical) perceptual and conceptual representations as well as into issues in visual perception.

Acknowledgments

FH would like to thank Gerry Altmann and Graham Hitch for discussion of this research.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38(4), 419-439.

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. *Cognitive Psychology*, 6, 813-839.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time-course of frequency effects in spoken word recognition: Evidence from eye movements. *Cognitive Psychology*, 42, 317-367.

Dahan, D. & Tanenhaus, M. K. (2002). Activation of conceptual representations during spoken-word recognition. Poster presented at the 43rd Annual Meeting of the Psychonomics Society, Kansas, USA.

Findlay J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. Oxford University Press.

Gaskell, M.G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, 12, 613-656.

Huettig, F. & Altmann, G.T.M. (2004). The online processing of ambiguous and unambiguous words in context: Evidence from head-mounted eye-tracking. In M. Carreiras & C. Clifton (Eds.). *The On-line Study of Sentence Comprehension: Eyetracking, ERP and Beyond*. New York, NY: Psychology Press.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.

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

Moss, H. E., McCormick, S. F., & Tyler, L. K. (1997). The time course of activation of semantic information during spoken word recognition. *Language and Cognitive Processes*, 12, 695-731.

O'Regan, J. K. (1992). Solving the 'real' mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology*, 46, 461-488.

Pickering, M. J., McElree, B., & Garrod, S. (submitted). Interactions of language and vision restrict "visual world" interpretations. <http://ila.psych.nyu.edu/users/bd/m/Dept/index.html>

Saslow, M. G. (1967). Latency for saccadic eye movement. *Journal of the Optical Society of America*, 57, 1030-1033.

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

Yee, E., & Sedivy, J. (2001) Using Eye Movements to Track the Spread of Semantic activation during spoken word recognition. Paper presented to the 13th Annual CUNY Conference, Philadelphia, USA.