Spoken Language Comprehension Improves the Efficiency of Visual Search

Melinda J. Tyler (mjt15@cornell.edu)
Department of Psychology, Cornell University
Ithaca, NY 14853 USA

Michael J. Spivey (spivey@cornell.edu)
Department of Psychology, Cornell University
Ithaca, NY 14853 USA

Abstract

Much recent eye-tracking research has demonstrated that visual perception plays an integral part in on-line spoken language comprehension, in environments that closely mimic our normal interaction with our physical environment and other humans. To test for the inverse, an influence of language on visual processing, we modified the standard visual search task by introducing spoken linguistic input. In classic visual search tasks, targets defined by only one feature appear to "pop-out" regardless of the number of distractors, suggesting a parallel search process. In contrast, when the target is defined by a conjunction of features, the number of distractors in the display causes a highly linear increase in search time, suggesting a more serial search process. However, we found that when a conjunction target was identified by a spoken instruction presented concurrently with the visual display, the effect of set size on search time was dramatically reduced. These results suggest that the incremental linguistic processing of the two spoken target features allows the visual search process to, essentially, conduct two nested single-feature parallel searches instead of one serial conjunction search.

Introduction

For a psycholinguist studying spoken language comprehension, the visual environment would be considered "context". However, for a vision researcher, the visual environment is the primary target of study, and auditory/linguistic information would be considered the "context". Clearly, this variable use of the label "context" is due to differences in perspective, not due to any objective differences between language and vision. In everyday perceptual/communicative circumstances, humans must integrate visual and linguistic information extremely rapidly for even the simplest of exercises. Consider the real-time dance of linguistic, visual, and even gestural events that takes place during a conversation about the weather. This continuous coreferencing between visual and linguistic signals may render the very idea of labeling something as "context" arbitrary at best, and perhaps even misleading.

The problem of "context" has traditionally been dealt with in a rather drastic fashion: researchers forcibly ignore it. If context does not influence the primary functions of the process of interest (be it in language, vision, memory, reasoning, or action), then that process can be thought of as an encapsulated module which will permit dissection via a nicely limited set of theoretical and methodological tools. For example, prominent theories of visual perception and attention posit that the visual system is functionally independent of other cognitive processes (Pylyshyn, 1999; Zeki, 1993). This kind of modularity thesis has been applied to accounts of language processing as well (Chomsky, 1965; Fodor, 1983). As a result, a great deal of progress has been made toward developing first approximations of how vision may function and how language may function.

However, recent eye-tracking studies have shown evidence that visual perception constrains real-time spoken language comprehension. For example, temporary ambiguities in word recognition and in syntactic parsing are quickly resolved by information in the visual context (Allopenna, Magnuson, & Tanenhaus, 1998; Spivey & Marian, 1998; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Findings like these are difficult for modular theories of language to accommodate.

The present experiment demonstrates the converse: that language processing can constrain visual perception. In a standard visual search task, a target object is typically defined by a conjunction of features, and reaction time increases linearly with the number of distractors, often in the range of 15-25 milliseconds per item (Duncan & Humphreys, 1989; Treisman & Gelade, 1980; Wolfe, 1994). However, when we presented the visual display first, and then provided the spoken target features incrementally, we found that reaction time was considerably less sensitive to the number of distractors.

With conjunction search displays, increased reaction times as a linear function of set size were originally interpreted as evidence for serial processing of the objects in the display, and contrasted with the near-flat function of reaction time by set size observed with

feature search displays -- where a single feature is sufficient to identify the target object. It was argued that the early stages of the visual system process individual features independently and in parallel (Livingstone & Hubel, 1988), allowing the target object to "pop out" in the display if it is discriminable by a single feature, but requiring application of an attentional window to the individual objects, one at a time, if the target object is discriminable only by a conjunction of features (Treisman & Gelade, 1980). This categorical distinction between parallel search of single feature displays and serial search of conjunction displays has been supported by PET scan evidence for a region in the superior parietal cortex that is active during conjunction search for motion and color, but not during single feature search for motion or for color (Corbetta, Shulman, Miezin, & Petersen, 1995).

However, several studies have discovered particular conjunctions of features that do not produce steeply sloped reaction-time functions by set size (e.g., McLeod, Driver & Crisp, 1988; Nakayama & Silverman, 1986). Additionally, it is possible to observe the phenomenology of 'pop-out' while still obtaining a significant (albeit, small) effect of set size on reaction time (Bridgeman & Aiken, 1994). Moreover, it has been argued that steeply sloped reaction-time functions may not reflect serial processing of objects in the display, but rather noise in the human visual system (Eckstein, 1998; Palmer, Verghese, & Pavel, 2000). Overall, a wide range of studies have suggested that the distinction between putatively "serial" and "parallel" search functions is continuous rather than discrete, and should be considered extremes on a continuum of search difficulty (Duncan & Humphreys, 1989; Nakayama & Joseph, 1998; Olds, Cowan, Jolicoeur, 2000; Wolfe, 1994, 1998).

In a recent study, Spivey, Tyler, Eberhard, and Tanenhaus (in press b) demonstrated that the incremental processing of linguistic information could, essentially, convert a difficult conjunction search into a pair of easier searches. When target identity was provided via recorded speech presented concurrently with the visual display, displays that typically produced search slopes of 19 ms per item produced search slopes of 8 ms per item. It was argued that if a spoken noun phrase such as "the red vertical" is processed incrementally (cf. Altmann, & Kamide, 1999; Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Marslen-Wilson, 1973, 1975), and there is extremely rapid integration between partial linguistic and visual representations, then one might predict that the listener should be able to search items with the first-mentioned feature before even hearing the second one. If the observer can immediately attend to the subset of objects sharing that first-mentioned feature, such as the target color (Egeth, Virzi, & Garbart, 1984; Friedman-Hill & Wolfe, 1995; Motter & Holsapple, 2000), and subsequently search for the target object in that subset upon hearing the second-mentioned feature, then this initial immediate group selection should reduce the effective set size to only those objects in the display that share the first-mentioned feature – effectively cutting the search slope in half.

At least two concerns remain before this basic finding can be extended and tested in the many different variations of visual search displays. First, since a slope of 8 ms per item is clearly in the range of what has traditionally been considered “parallel search”, it is somewhat unclear whether the result is in fact a *halving* of the effective set size or a near *elimination* of the effect of set size. Essentially, the question is whether the first feature extraction is a genuine “pop-out” effect and the second is a genuine serial search of those “popped out” objects (half of the set size), or are both searches “practically parallel”. A replication of the study may provide some insight into this question. Second, the experiments reported by Spivey et al. (in press b) ran participants in separate blocks of control trials and trials with concurrent auditory/visual input. It is in principle possible that practice was somehow more effective in the auditory/visual concurrent condition, or that subjects developed some unusual strategy in that condition that they didn’t use in the control condition. To be confident in the result, it is necessary to replicate it with a mixed (instead of blocked) design, where the control trials and the A/V concurrent trials are randomly interspersed.

Experiment

Method

Participants Eighteen Cornell undergraduate students were recruited from various Psychology classes. Participants were reimbursed 1 point of course extra credit for participating in the study.

Procedure The experiment was composed of two types of trials presented in random mixed order within one continuous block of 192 trials. Participants were instructed to take breaks between trials when they felt it was necessary. In one type of trial, the participant was auditorily informed of the target identity *before* presentation of the visual display (‘Auditory First’ control condition). In the other type of trial, the participant was auditorily informed of the two defining feature words of the target *concurrently with* the onset of the visual display (‘A/V Concurrent’ condition) (see Figure 1) Of the 192 trials, 96 were ‘Auditory First’, and 96 were ‘A/V concurrent.’

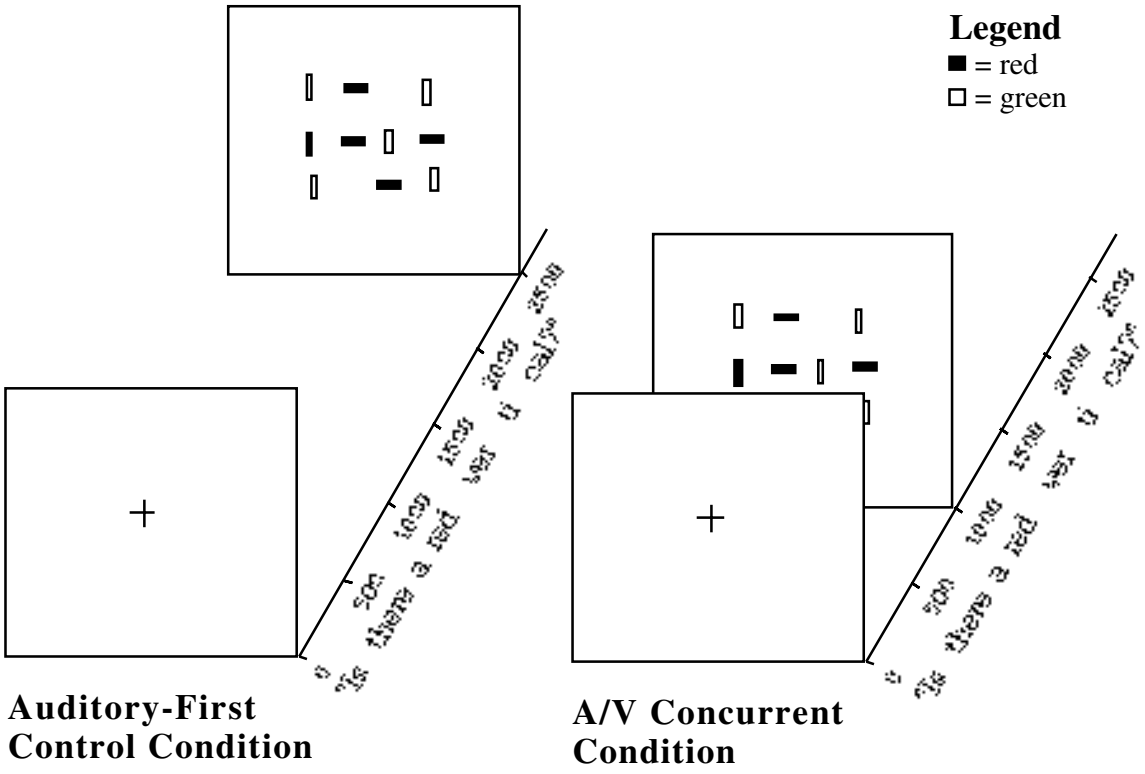

Figure 1. Schematic diagram of the two conditions. In the Auditory-First condition, the search display is presented after the entire spoken query is heard, whereas in the A/V Concurrent condition, the search display is presented immediately before the two target features are heard. Reaction time is measured from the point of display onset.

Trials began with a question delivered in the format of a speech file. The same female speaker recorded all speech files with the same preamble recording, "Is there a…" being spliced onto the beginning of each of the four types of target query types ("…red vertical?", "…red horizontal?", "…green vertical?", and "…green horizontal?"). Each of the four types of speech files were edited to be almost identical in length, and with almost identical auditory spacing of defining feature words. Participants were instructed to press a 'yes' key on a computer keyboard if the queried object was present in the display, and the 'no' key if it was absent. It was stressed to participants that they should do this as quickly and accurately as possible. An initial fixation cross preceded the onset of the visual display in order to direct participants' gaze to the central region of the display. Each stimulus bar subtended 2.8 degrees X 0.4 degrees of visual angle, and neighboring bars were separated from one another by an average of 2 degrees of visual angle. Trials with red vertical bars as targets and trials with green vertical bars as targets, as well as red and green horizontal bars as targets, were equally and randomly distributed throughout the session. All participants had normal or corrected-to-normal vision, and all had normal color perception. The objects comprising the visual display appeared in a grid-like arrangement positioned centrally in the screen (see Figure 1). Set sizes of objects comprising the visual displays were 5, 10, 15, and 20.

Results

Mean accuracy was 95% and did not differ across conditions. Figure 2 shows the reaction time by set size functions for target-present trials (filled symbols) and target-absent trials (open symbols) in the A/V Concurrent condition and the Auditory-First condition. The best-fit linear equations are accompanied by their r^2 values indicating the percentage of variance accounted for by the linear regression.

Overall mean reaction time was slower in the A/V Concurrent condition as a result of the complete auditory notification of target identity being delayed by approximately 1.5 seconds relative to the Auditory-First control condition. However, since spoken word recognition is incremental, participants were able to begin processing before both target feature words had been presented, and overall reaction time was only delayed by about 600 milliseconds.

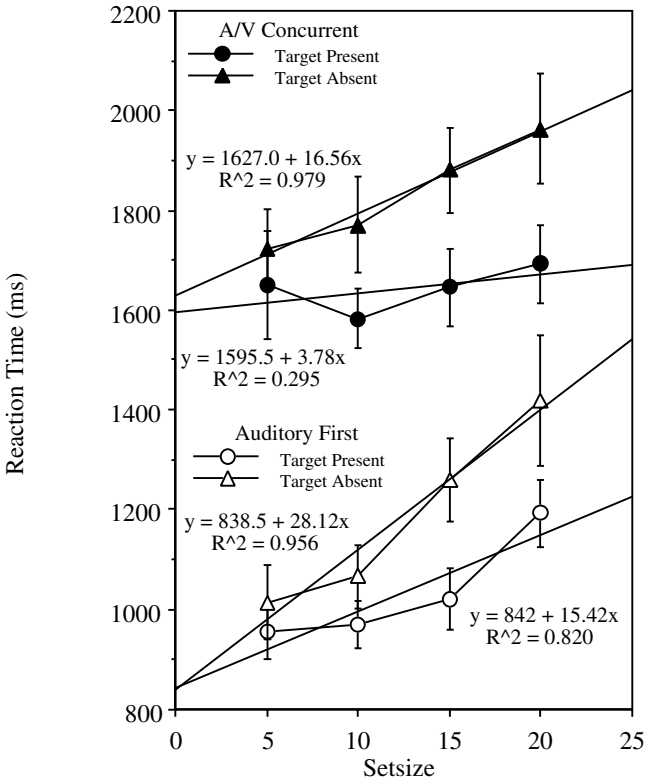

Figure 2: Reaction time as a function of set size.

Repeated-measures analysis of variance revealed significant main effects of Condition [F(1, 16)=230.27, p<.01], Target Presence/Absence [F(1,16)=27.97, p<.01], and Set Size [F(3, 48)=22.83, p<.01]. The effect of Condition was simply that overall reaction times were slower when the delivery of target identity was delayed in the A/V Concurrent condition. The effect of Target Presence/Absence was the common finding that it takes longer to determine that the target is absent than to determine that it is present. Essentially, the target-present case involves something akin to a self-terminating search, and the target-absent case requires something like an exhaustive search. The main effect of Set Size simply showed that, when Condition and Target Presence/Absence are ignored, having more distractors increased reaction time. There was also an interaction between Set Size and Target Presence/Absence, showing that the effect of Set Size was more pronounced in the target-absent trials than in the target-present trials [F(3,48)=4.36, p<.01].

The important result for the purposes of our inquiry was the significant interaction between Condition and Set Size, indicating that the effect of Set Size was more pronounced in the Auditory-First control condition than in the A/V Concurrent condition [F(3, 48)=4.92, p<.01]. Despite the fact that the visual displays were identical, results indicated shallower slopes for reaction-time functions in the A/V Concurrent condition compared to the Auditory-First control condition (Figure 2).

To specifically test whether the mean slope was significantly shallower in the A/V Concurrent condition, participants' individual set size slopes from the two conditions were compared via paired t-tests, and revealed significantly shallower slopes for the A/V Concurrent condition in target-present trials [3.78 vs. 15.42 ms per item; t(16)=2.09, P<.05] and in target-absent trials [16.56 vs. 28.12 ms per item; t(16)=3.39, P<.01]. These results indicate that adjusting the timing of the spoken query (e.g., "Is there a red vertical?"), such that the two target feature words were presented while the visual display was visible, allowed participants to find the target object in a manner that was substantially less affected by the total number of distractors. In fact, the mean slope of 3.78 ms per item in the target-present trials of the A/V Concurrent condition was not significantly greater than zero [t(16)=1.12, p>.25], whereas the mean slope of 15.42 ms per item in the target-present trials of the Auditory-First control condition was robustly greater than zero [t(16)=4.47, p<.001].

Hence it appears that, in the Auditory-First condition, the search process may employ a conjunction template to find the target, thus forcing a serial-like process akin to sequentially comparing each object to the target template. However, in the A/V concurrent condition, it appears that the incremental nature of the speech input allows the search process to begin when only a single feature of the target identity has been heard. This initial single-feature search proceeds in a more parallel fashion (with the second-mentioned target feature being used to find the target amidst the attended subset), thus dramatically improving the efficiency of search.

Discussion

The results suggest that, due to the incremental nature of spoken language comprehension (Allopenna et al., 1998; Altmann & Kamide, 1999; Cooper, 1974; Eberhard et al., 1995; Marslen-Wilson, 1973, 1975; Tanenhaus et al., 1995) the listener/observer can selectively attend to the subset of objects that exhibit the target feature which is mentioned first in the speech stream. Upon hearing even just a portion of the second-mentioned target feature a few hundred milliseconds later, the observer can then locate the conjunction target object amidst this attended (spatially noncontiguous) subset. These results highlight the incremental processing of spoken language comprehension, and demonstrate the human brain's ability to seamlessly cross-index partial linguistic representations (of a noun phrase, for example) with partial visual representations (of a cluttered visual display).

A more detailed question remains, concerning whether the improved efficiency in visual search is due to the first-mentioned target feature initiating an instantaneous parallel search and the second-mentioned target feature initiating a serial search among the attended items (thus cutting the search slope in half) or to both spoken target features initiating parallel searches (causing the search slope to look like that of a single-feature search). In Spivey et al. (in press b, Experiments 1 and 2), the target-present search slopes of 7.7 and 8.9 ms per item in the A/V Concurrent conditions were roughly consistent with both of those interpretations. When parallel and serial search were conceived of as discrete categories, any set-size function of less than 10 ms per item was generally interpreted as "parallel search." However, Spivey et al.'s target-present Auditory First conditions produced search slopes of 19.8 and 18.6 ms per item -- approximately twice the A/V Concurrent slopes.

The present results, with a target-present search slope of 3.78 ms per item in the A/V Concurrent condition, appear to support the "two parallel searches" alternative. However, in all likelihood, the two alternatives outlined above rely too much on the discrete distinction between "parallel" and "serial" search. If there is indeed a continuum of search efficiency (Duncan & Humphreys, 1989; Nakayama & Joseph, 1998; Olds, Cowan, Jolicoeur, 2000; Wolfe, 1994, 1998), and conjunction search is not quite accurately described as an object-by-object serial process (Eckstein, 1998; Palmer, Verghese, & Pavel, 2000), then it might be safest to conclude that each spoken target feature initiates a "relatively efficient" search that is not quite parallel and not quite serial. Importantly, the second search appears to be able to work selectively on the output of the first one -- compelling evidence for the continuous incrementality with which linguistic and visual processing can coordinate.

Until now, there has been little or no evidence for real-time visual perception being influenced by linguistic context. However, there is considerable work reporting demonstrations of real-time language processing being influenced by visual context (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; McGurk & MacDonald, 1976; Spivey & Marian, 1998; Tanenhaus et al., 1995). Recent eyetracking research has shown in a number of circumstances that the resolution of temporary ambiguities is immediately biased by information in the visual array. For example, when participants were instructed to "pick up the candy", they often looked first at a *candle* before then fixating the candy (Tanenhaus et al., 1995). A precise timing analysis of the eye movements suggested that, when the candle and candy are in the visual display, participants had mental representations for both 'candle' and 'candy' equally partially active during the first couple hundred milliseconds of the word (Allopenna et al., 1998). When only the candy was in the display, the eye-movement data suggested that word recognition was faster, involving less competition from partially active alternatives.

Similar findings were reported for syntactic ambiguity resolution. When the visual context contained a referential ambiguity (e.g., two apples for the instruction "Put the apple…") that was best resolved by pursuing the correct parse of a syntactic ambiguity ("Put the apple on the towel in the box."), participants eye-movement patterns suggested a fast and correct interpretation of the instruction. When there was no such referential ambiguity (e.g., only one apple), participants produced eye-movement patterns indicating a mis-parse of the instruction (Spivey, Tanenhaus, Eberhard, & Sedivy, in press a; Tanenhaus et al., 1995).

Those effects of visual context immediately influencing language comprehension were initially met with considerable resistance from a substantial portion of psycholinguists. However, when we would discuss those findings with vision researchers, they were often appreciative but not terribly impressed. To them, it made perfect sense that the visual system was important and powerful enough to occasionally tell the language system what to do very quickly. We are curious to see the reaction of the vision research community to the present results.

Returning to our discussion of the notion of "context", which began this paper, it seems that the rapidity with which the visual system and the language system can coordinate their representations suggests that any attempt to label some signal as "context" is doomed to be an arbitrary choice – a choice that risks marginalizing important information sources as well as opaquely lumping discriminable information sources. Essentially, the less we assume encapsulated modular processes in language and in vision, the less use we have for the notion of "context" in language and in vision. Instead of visual processing and linguistic processing, perhaps a considerable portion of our daily mental life is made up of visuolinguistic processing.

Since the human brain is neither a psycholinguist nor a vision researcher (indeed, it is much more than even the two of them combined), it is not susceptible to the biased perspectives they exhibit. As far as the brain is concerned, no one incoming signal is the "primary signal" with the others being "context". Each time slice of perceptual experience is a rich tapestry of multi-modal environmental inputs, all of which the brain integrates and incorporates simultaneously. Our results suggest that, across the domains of language and vision, it is surprisingly good at doing that job immediately and continuously.

Acknowledgments

We thank Michael Tanenhaus, Kathy Eberhard, and Julie Sedivy for helpful discussions, and Quinn Hamilton for assistance with data collection. Supported by a Sloan Fellowship in Neuroscience (M.J.S.).

References

Allopenna, P. D., Magnuson, J., & 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, 419-439.

Altmann, G. T. M. & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73, 247-264.

Bridgeman, B. & Aiken, W. (1994). Attentional "popout" and parallel search are separate phenomena. *Investigative Ophthalmology & Visual Science,* **35**, 1623.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Corbetta, M., Shulman, G., Miezin, F., & Petersen, S. (1995). Superior parietal cortex activation during spatial attention shifts and visual feature conjunction. *Science,* 270, 802-805.

Duncan, J. & Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review*, 96, 433-458.

Eberhard, K. M., Spivey-Knowlton, M., Sedivy, J. & Tanenhaus, M. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research,* 24, 409.

Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science* 9, 111.

Egeth, H. E., Virzi, R., & Garbart, H. (1984). Searching for conjunctively defined targets. *Journal of Experimental Psychology: Human Perception and Performance,* 10, 32.

Fodor, J. A. (1983). *Modularity of Mind*. Cambridge, MA: MIT Press.

Friedman-Hill, S. & Wolfe, J. (1995). Second-order parallel processing: Visual search for the odd item in a subset. *Journal of Experimental Psychology: Human Perception and Performance,* 21, 531.

Livingstone, M. & Hubel, D. (1988). Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science,* 240, 740-749.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, 244, 522-523.

Marslen-Wilson, W. (1975). Sentence perception as an interactive parallel process. *Science*, 189, 226-228.

McGurk, & MacDonald (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

McLeod, P., Driver, J., & Crisp, J. (1988). Visual search for conjunctions of movement in visual search. *Nature,* 332, 154-155.

Motter, B. C. & Holsapple, J. W. (2000). Cortical image density determines the probability of target discovery during active search. *Vision Research,* 40, 1311 1322.

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *Journal of Neurophysiology,* 70, 909-919.

Nakayama, K. & Silverman, G. H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature,* 320, 264 265.

Nakayama, K. & Joseph, J. (1998). Attention, pattern recognition, and pop-out in visual search. In R. Parasuraman (Ed.), *The Attentive Brain*. Cambridge, MA: MIT Press. pp. 279-298.

Olds, E. S., Cowan, W. & Jolicoeur, P. (2000). The time-course of pop-out search. *Vision Research,* 40, 891-912.

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research,* 40, 1227-1268.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case of impenetrability of visual perception. *Behavioral and Brain Sciences,* 22, 341-423.

Spivey, M. J. & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science,* 10, 281-284.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (in press a). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*.

Spivey, M. J., Tyler, M. J., Eberhard, K. M., & Tanenhaus, M. K. (in press b). Linguistically mediated visual search. *Psychological Science*.

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

Treisman, A. & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, 12, 97-136.

Wolfe, J. M. (1994). Guided Search 2.0: A revised mode of visual search. *Psychonomic Bulletin & Review,* 1, 202-238.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science,* 9, 33-39.

Zeki, S. (1993). *A Vision of the Brain*. Oxford: Blackwell Scientific (1993).