Published in final edited form as:
Acta Psychol (Amst). 2011 June ; 137(2): 172–180. doi:10.1016/j.actpsy.2010.09.010.

A goal-based perspective on eye movements in visual world studies

Anne Pier Salverda, **Meredith Brown**, and **Michael K. Tanenhaus**
Department of Brain and Cognitive Sciences, University of Rochester

Abstract

There is an emerging literature on visual search in natural tasks suggesting that task-relevant goals account for a remarkably high proportion of saccades, including anticipatory eye-movements. Moreover, factors such as “visual saliency” that otherwise affect fixations become less important when they are bound to objects that are not relevant to the task at hand. We briefly review this literature and discuss the implications for task-based variants of the visual world paradigm. We argue that the results and their likely interpretation may profoundly affect the “linking hypothesis” between language processing and the location and timing of fixations in task-based visual world studies. We outline a goal-based linking hypothesis and discuss some of the implications for how we conduct visual world studies, including how we interpret and analyze the data. Finally, we outline some avenues of research, including examples of some classes of experiments that might prove fruitful for evaluating the effects of goals in visual world experiments and the viability of a goal-based linking hypothesis.

In 1974, Roger Cooper published a remarkable article titled “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”. Cooper tracked the eye movements of participants as they listened to stories while looking at a display of pictures. He found that participants initiated saccades to pictures that were named in the stories, as well as pictures associated with those words. Moreover, fixations were often generated before a spoken word was heard entirely, suggesting a tight coupling of visual and linguistic processing.

More than twenty years later, Tanenhaus, Spivey-Knowlton, Eberhard and Sedivy (1995), who were unaware of Cooper’s experiments at the time, published a report that described experiments in which participants followed spoken instructions to manipulate real objects in a visual workspace. They reported that reference resolution was closely time-locked to the point in the speech stream when the combination of the objects in the workspace and the unfolding language was, in principle, sufficient to uniquely identify a referent. These results demonstrated nearly immediate integration of visual and linguistic information in word recognition, reference resolution and syntactic processing. Tanenhaus et al. dubbed their paradigm the “visual world” paradigm.[i]

Address for correspondence: Anne Pier Salverda, University of Rochester, Department of Brain & Cognitive Sciences, Meliora Hall, Box 270267, Rochester, NY 14627, Tel: +1 585 275 1844, Fax: +1 585 442 9216, asalverda@bcs.rochester.edu.

[i]The term "visual world" was used in presentations by MKT beginning in 1994. To the best of our knowledge it first appeared in print in Allopenna, Magnuson and Tanenhaus (1998).

During the last decade or so, the visual world paradigm has become a methodology of choice for studying a wide range of topics in spoken language processing, ranging from speech perception to interactive conversation, in infants, children, adults, older adults and special populations, making the title of Cooper's article prescient indeed (for recent reviews see Tanenhaus & Trueswell, 2006; Tanenhaus, 2007a,b). The recent surge of studies using eye movements to study spoken language processing can be traced to the Tanenhaus et al. (1995) report (for discussion, see Henderson & Ferreira, 2004; van Gompel, Fischer, Murray, & Hill, 2007; Ferreira & Tanenhaus, 2007). However, the perspective on the relationship between spoken language and eye movements that is implicit in much current work using the visual world paradigm is arguably closer to that reflected in the title of Cooper's (1974) article, that is, *the control of eye fixation by (the meaning of) spoken language*, than to the perspective that guided the work of the Rochester group.

Tanenhaus and colleagues chose to embed their studies in an explicit task that required visually guided reaching, because they were influenced by important early work by Ballard, Hayhoe and colleagues at Rochester using eye movements to investigate vision in natural tasks. That research was motivated by a theoretical approach that viewed understanding the task of the visual system (e.g., visually guided reaching, navigation, etc.), as central to understanding not only the role of eye movements in behavior, but, more generally, the nature of vision itself. According to this approach, sometimes called *animate vision* in the artificial intelligence literature, eyes give an organism the ability to acquire real-time information from the environment in order to initiate or guide goal-directed action. From this perspective, then, it is no surprise that whereas eyes have evolved multiple times in multiple ways in animate organisms (Land & Fernald, 1992), they have not, to the best of our knowledge, evolved in non-animate living organisms that are nonetheless exquisitely sensitive to light, for example plants.

We believe that the emerging literature on vision in natural tasks provides important insights about attention and visual search that may have important consequences for our understanding of the interface between vision and language processing, especially in task-based studies. Most generally, this literature shows that task-relevant goals account for a remarkably high proportion of saccades and fixations, including anticipatory eye movements. Moreover, they may also play a central role in determining what aspects of objects are attended to and which of those are represented in visual working memory. Our goal in this article is to consider the implications of results from the literature on vision in natural tasks for research in psycholinguistics using the visual world paradigm.

Visual search in natural tasks

Traditional approaches to visual search, such as feature integration theory (Treisman & Gelade, 1980), have focused extensively on the role of low-level visual features (e.g., color, orientation, shape) in pre-attentive visual processing and in the subsequent allocation of visual attention. These approaches emerged from and promoted an experimental tradition involving the measurement of behavioral responses to visual displays containing simple objects varying along various proposed basic feature dimensions (e.g., identifying a rotated T within a set of upright Ts on a monochromatic background). The assumption underlying this body of work is that the elementary perceptual features and the visual processes studied using these simplified static displays are the fundamental units of visual search in more complex, real-life scenes. Given this assumption, it follows that basic stimulus features should be key predictors of the deployment of visual attention and search behaviors in more complex real-world environments.

However, investigations of active visual search in dynamic scenes and in the context of specific tasks have called into question the extent to which invariant, static properties of a scene are behaviorally relevant in real-world contexts. On the one hand, stimulus features are, of course, involved in the initial parsing and representation of the contents of a scene, and global estimates of visual salience derived by integrating multiple feature values at each position within a scene (i.e., saliency maps; see Koch & Ullman, 1985) correlate with fixation probabilities during viewing of a scene when the observer is given no particular task (Parkhurst, Law, & Niebur, 2002). On the other hand, however, feature-based salience turns out to be a poor predictor of gaze patterns when a participant is engaged in a well-defined task and therefore needs to derive certain information from the visual input to successfully complete the task (e.g., Buswell, 1935; Yarbus, 1967; Turano, Geruschat, & Baker, 2003; Henderson, Malcolm, & Schandl, 2009; Jovancevic-Misic & Hayhoe, 2009). Under these circumstances, task-based predictors account for much more of the variance in fixation patterns than otherwise highly salient featural attributes of the environment. The utility of feature-based salience measures in explaining real-world search behaviors therefore appears to be more limited than predicted by bottom-up theories of visual search.

In task-based approaches to visual search, the allocation of attention within a scene (and the resulting pattern of saccades and fixations) is primarily governed by the observer's cognitive goals, which are strongly influenced by the task that they are performing. These goals influence where, when and for how long people look at aspects of the visual world. Studies of every-day but complex visuomotor behaviors, such as preparing tea, making sandwiches, and driving, indicate that eye movements are tightly linked to immediate task demands (e.g., Ballard, Hayhoe, & Pelz, 1995; Land, Mennie, & Rusted, 1999; Hayhoe, Shrivastava, Mruczek, & Pelz, 2003; Land & Lee, 1994). Participants exhibit a strong tendency to fixate objects immediately before they become relevant to the execution of a task subgoal (e.g. fixating an object immediately prior to reaching for it). Moreover, participants direct their fixations to those parts of an object that are behaviorally most relevant (e.g., the spout of a tea kettle during the pouring of hot water). In most of these studies, the vast majority of fixations can clearly be attributed to task-based goals. These findings suggest that the spatial and temporal distribution of fixations when carrying out a well-defined task is related to the acquisition of specific, task-related information in service of behavioral goals.

In addition to influencing the location and timing of fixations, cognitive goals play a key role in determining the information encoded during fixations. Evidence from change blindness and transsaccadic memory studies indicates that the visual information that is represented or available for comparison between sequential views of a scene is generally limited (e.g., Simons & Levin, 1997; Irwin, 1991; O'Regan, 1992). Moreover, studies of visual behavior during the execution of real-world tasks suggest that task demands influence the amount of resources available to encode information in short-term memory, an observation that is compatible with emerging theories of visual short-term memory based on information load rather than number of object files (cf. Alvarez & Cavanagh, 2004).

For example, in a now classic experiment, the Hayhoe/Ballard group used a block-copying task, illustrated in Figure 1 (e.g., Ballard et al., 1995). The participant, who was wearing a head-mounted eye-tracker, copied a block pattern from a model area into a target area, using blocks from a resource area. Participants nearly always looked at the block they intended to pick up immediately before and while reaching for it, replicating prior findings on visually-guided reaching. More strikingly, participants typically made two fixations to each block in the model area during the execution of the task: one fixation prior to reaching for a block in the resource area and one fixation prior to moving that block to the target area. Ballard, Hayhoe and their collaborators (cf. Ballard, Hayhoe, Pook, & Rao, 1997) proposed that participants engaged in this behavior to minimize memory demands. During the first fixation

to the model area, they encoded only the color of the relevant block. After selecting a matching block from the resource area, the participants then returned to the model area to check and encode the location of the block. This memory strategy was, however, flexible. If the model area was moved far enough away from the resource area to require a head movement for inspection, participants typically made only a single fixation to the model area (Ballard et al., 1995). Therefore, the time and effort required to access information from a scene via a new fixation influences both the pattern of fixations and the information retrieved during a fixation that is stored in memory. Importantly, aspects of the task that a participant performs, including aspects that change dynamically during the performance of the task, can strongly influence the time and resources available for accessing information, and thus the information that is encoded during a fixation. For instance, as task complexity increases in a block-sorting task, participants begin to rely less on working memory and more on the external environment for the successful completion of the task (Droll and Hayhoe, 2007). Taken together, these findings support the notion that the information encoded by a fixation is largely determined by cognitive goals.

Linking hypotheses between language processing and the visual world

Interactions between spoken language and visual scene

In most visual world experiments, the relevant visual environment consists of a collection of clipart drawings on a computer screen. The visual display corresponds to a small array of objects (e.g. four objects in a grid in a setup developed by Allopenna, Magnuson, & Tanenhaus, 1998, which has been widely used to study spoken-word recognition), or a static, schematic scene consisting of a collection of objects, including potential agents (e.g. Altmann & Kamide, 1999, who developed this setup to study sentence-level processing). Shortly after the visual display appears, a spoken word or sentence is presented, which relates to one or more objects in the scene. Of interest is the location and timing of fixations to the objects in the visual display in response to hearing the spoken stimulus—and in particular how this pattern of fixations is affected by particular properties of the linguistic stimulus, which is carefully controlled and varies across experimental conditions.

In order to make inferences about language processing on the basis of the pattern of fixations observed in a visual world experiment, one needs to specify a linking hypothesis that describes the relationship between eye movements and linguistic processing. A standard assumption in the visual-world literature is that a shift in visual spatial attention is typically, though not necessarily, followed by a saccadic eye movement to the attended region. As a consequence, the locus of fixation provides insights into the current interpretation of the speech input. A well-specified linking hypothesis should also provide an answer to the basic question of why, when and how listeners shift their visual attention to visual information that is related to an unfolding linguistic expression. Furthermore, it should specify how eye movements reflect the mutual effects of linguistic processing and the visual processing of the contents of the visual display or scene.

In the sections that follow, we contrast a comprehension-based linking hypothesis with a goal-based linking hypothesis. These linking hypotheses differ primarily in the degree of emphasis on the importance of goal structures and the degree to which they view situated language comprehension as a separable component that can be embedded within more specific task-based goals. We begin by outlining the modal, though often implicit, linking hypothesis that we believe is assumed by most researchers who use the visual world paradigm. According to this view the goal of the participant is to understand the utterance, the scene, and how they relate to one another. We then introduce a more task-based perspective, inspired by the work on vision in natural tasks that we briefly reviewed earlier.

We then consider some of the possible consequences of a goal-based linking hypothesis for the interpretation of fixation patterns in visual world experiments.

A comprehension-based linking hypothesis

Numerous studies have now demonstrated rich interactions between spoken language and the visual world. Perhaps the simplest effects occur when a listener fixates an object upon hearing its name (see Salverda & Altmann, in revision, for a study suggesting that such effects are automatic, i.e., at least partially independent of task). In addition, linguistic material that is relevant for reference resolution but that is not inherently referential, such as verbs, prepositions, and scalar adjectives, has immediate effects on fixations to potential upcoming referents (e.g., Altmann & Kamide, 1999; Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002; Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995). Furthermore, fixations to a blank screen as people listen to descriptions of events with spatial descriptions can systematically reflect the construction of an internal model (Spivey & Geng, 2001) and can reflect expectations for an upcoming noun that refers to an object no longer present (e.g. looking at the location where a cake had just been presented upon hearing the verb "eat" in the sentence "The boy will eat the cake"; Altmann, 2004; Altmann & Kamide, 2009). The earliest moments of reference resolution and syntactic-ambiguity resolution (as indexed by eye movements) are influenced by the presence of referential alternatives (Altmann & Kamide, 2007, 2009; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Tanenhaus et al., 1995; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Trueswell, Sekerina, Hill, & Logrip, 1999; Novick, Thompson-Schill, & Trueswell, 2008; Kaiser & Trueswell, 2008).

A comprehension-based linking hypothesis can account for the aforementioned language-vision interactions while preserving some degree of representational autonomy between linguistic and visual representations. Fixations can be modeled by assuming that language and vision are independent or partially independent systems, and that visual-linguistic interactions arise due to areas of representational overlap (cf. Jackendoff, 2002). The underlying assumption that guides this approach, from the perspective of language processing, is that language comprehension is itself is a well-defined task-independent process that can be decomposed into building representations at traditional levels of linguistic representation, such as recognizing words, building syntactic structures, resolving reference, and building event representations (i.e., representations concerning the roles that entities play in the events or activities denoted by the predicates in the sentence). Most crucially, language comprehension is assumed to be fundamentally different from and independent of other task goals.

If we merge the comprehension-based view of language processing with a bottom-up approach to scene representation, then the appearance of a stimulus display in a visual world study would result in the bottom-up construction of a visual scene representation. This process would proceed independently of any expectations about the language that would subsequently be presented. We will assume that the visual scene is represented as an event-based conceptual representation.

When a spoken sentence is presented, incremental analysis of the spoken input results in the construction of linguistic representations. Because the linguistic input unfolds over time, and because speech is processed incrementally, listeners encounter temporary ambiguities at multiple levels of representation (e.g. lexical ambiguity, referential ambiguity, or syntactic ambiguity). At each levels of representation, listeners attempt to resolve these temporary ambiguities by making provisional commitments to likely alternatives, taking into account “correlated constraints” (cf. MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell, Tanenhaus, & Garnsey, 1994), and generate expectations about upcoming information (e.g., Altmann & Kamide, 1999, 2007; Levy, 2008). The interpretation of the speech input thus

changes dynamically, over time, as more information in the speech signal becomes available. At each moment in time, the linguistic representations that are (most) activated given the currently available speech input may interact with static visual representations derived from processing the visual scene, which results in fixations to objects and entities in the scene that reflect the current interpretation of the unfolding utterance.

This view, which has been developed in a series of papers by Altmann and colleagues (Altmann & Kamide, 2007; also see Knoeferle & Crocker, 2006; Mayberry, Knoeferle, & Crocker, 2009 for a related proposal and implemented models) allows for the possibility that some common representational format drives eye movements to aspects of a display in a visual world experiment. According to this perspective, eye movements in visual world studies reflect interactions between linguistic and visual representations that occur at the level of conceptual representations. The linguistic representations that are constructed during the incremental processing of the spoken input include conceptual representations pertaining to the meaning of individual words and linguistic expressions. These conceptual representations include visually-based information, such as the shape of an object (see Dahan & Tanenhaus, 2005, and Huettig & Altmann, 2007, for evidence that visually-based information is rapidly available and used to search for the referent of a spoken noun in a visual world study). According to the linking hypothesis put forward by Altmann and Kamide (2007), featural overlap between conceptual representations activated by the spoken language and conceptual representations activated by the visual world result in a boost in activation of those representations. This boost in activation then results in increased attention to, and therefore, an increased likelihood of fixating, those objects in the visual world whose conceptual representations match the conceptual representations activated by the concurrently presented spoken language. Thus, the overlap in linguistic and visual representations naturally and automatically leads to the identification of the referents of referring expressions (e.g., fixating a piece of candy in a visual display upon hearing "The boy looked at the candy" or "Click on the candy").

The comprehension-based view also accommodates richer interactions between visual and linguistic systems, that is, fixations to relevant objects upon hearing linguistic information that is not inherently referential (e.g., looking at a cake upon hearing "The boy will eat the …"). Relevant information in the representation of the scene can also be used as a probabilistic constraint to resolve ambiguities in the linguistic input. For example, referential context could influence parsing of ambiguous input by providing information that is more consistent with one potential analysis. Likewise, the language might resolve ambiguities in the scene, for example, the role that real or depicted entities might play in an event suggested by the scene, or the type of event represented in the scene. However, interactions between language and vision are claimed to originate from, and be constrained to, areas of representational overlap between visual and linguistic representations. According to the comprehension-based view, then, the resulting interactions should in principle be fully predictable given the precise content of visual representations on the one hand, and linguistic representations on the other hand. Thus, patterns of fixations are not predicted to vary under different task conditions.

The linking hypothesis we have described thus far does not account for some clear effects of task on fixations. For example, Chambers et al. (2002) showed that the size of a potential container influenced whether or not it was considered as a potential goal for an instruction, such as “Pick up the cube. Now put the cube into the can.” Perhaps more strikingly, though, Chambers, Magnuson & Tanenhaus (2004) found that non-linguistic task-based affordances affect the earliest moments of syntactic ambiguity resolution. An example display in Experiment 2 in Chambers et al. (2004) had the following four objects: two whistles, one of which was on a clipboard and had a looped string attached to it; another clipboard; and a

box. Participants heard an instruction to manipulate objects in the display ("Put the whistle on the clipboard into the box"). Because the display contained two whistles, when listeners heard "Put the whistle …", they were expecting a phrase that would disambiguate the intended referent, that is, information that would clarify which whistle the speaker intended to refer to (see Altmann & Steedman, 1988). As a result, they correctly parsed the immediately following prepositional phrase "on the clipboard" as an NP modifier, and therefore fixated the whistle on the clipboard. But, when participants used a hook to carry out the spoken instruction, they incorrectly parsed "on the clipboard" as the goal even though the potential instrument (i.e., the hook) was not introduced linguistically and most of the instructions thoughout the experiment required the participant to use the hook in a non-instrument role (e.g., "Put the hook in the box"). In this condition, the referential domain was constrained by the action to be performed in service of the experimental task.

Task-based effects like those observed in the Chambers et al. studies are difficult to explain using the form of the comprehension-based linking hypothesis we have just outlined. We could, however, maintain the comprehension-based linking hypothesis and augment it by adding another layer of representation to represent task goals. This approach would necessitate having a principled way of determining when task goals can interact with comprehension. Recent incremental parsers in which probabilistic representations are computed in parallel, and real world knowledge can be used to adjust the probabilities of alternatives when new information arises, are one promising direction. If we adopt this approach then real world knowledge, including task constraints, could come into play at points of potential ambiguity.

A goal-based linking hypothesis

Although the comprehension-based linking hypothesis we sketched above remains viable, the vision in natural tasks literature suggests to us that it might be fruitful to explore a more radical "top-down" oriented linking hypothesis that assigns a more central role to task goals and does not assign a privileged role to comprehension *per se*. According to this view, different types of representations may be involved in the mapping of spoken language onto a visual scene, depending on the actual task that the participant is performing. Most generally, as in the literature on vision in natural tasks, we believe that a richer understanding of both language processing and how it is coupled with visual search will require using paradigms where the goal structure can be modeled.

In taking preliminary steps toward a goal-based model of the integration of visual and linguistic information in the visual world paradigm, we consider a model developed by Sprague, Ballard, and Robinson (2007) which implements some of the insights of researchers studying visual search in natural tasks. This model assumes that the basic elements of visually guided behavior are visuomotor routines. These routines are sequences of elemental operations that are involved in the selection of task-specific information from the visual world. In the case of visual search, the elemental operations include shifting the focus of processing and mentally marking specific locations for further processing or for future reference (cf. Ullman, 1984). A routine comprises a coordinated sequence of these basic operations that is assembled and selected to complete a particular visual task, such as monitoring the rising water level when filling a cup (Hayhoe, 2000). When confronted with a new task, new task-specific routines can be compiled and stored in memory for future use. These routines can in turn be combined to form a vast array of more complex representations and behaviors.

The Sprague et al. (2007) model simulates the scheduling of eye movements in a virtual walking environment with three tasks: litter collection, sidewalk navigation, and collision avoidance. The model learns to associate the states of each of these tasks with a discrete set

of visuomotor routines, which involve the extraction of only the information from the visual environment that is relevant for the successful execution of a particular motor sequence. Uncertainty about task-relevant properties of the environment (e.g., the exact location of a piece of litter) reduces the probability of successful task completion. Thus, eye movements are scheduled and directed in such a way as to minimize the total cost associated with the set of tasks performed at a given point in time, weighted by the reward associated with each task. To capture the effects of reward on task performance, Sprague et al. incorporated a reinforcement learning component into their model. This learning component captures key findings on incremental changes in the timing of saccades in the acquisition of complex skills (e.g., Land & McLeod, 2000; Sailer, Flanagan, & Johannson, 2005). Crucially, the model predicts that the pattern of fixations and frequency of fixations to different objects changes as a participant learns.

The idea of visuomotor routines as the primitive units of visually guided behavior can be applied straightforwardly to the integration of visual and linguistic information in visual world experiments. Routines relevant within the visual world domain would include some general-purpose visual routines, such as shifting the focus of processing and mentally marking certain locations for future processing (Ullman, 1984). Others would be more specific to visual-linguistic processing. One important routine would correspond to mapping the perceptually-based representations that become available as a spoken word unfolds onto objects in the visual world with matching affordances, resulting in fixations that are in service of identifying a potential referent of a spoken noun (for discussion see Dahan & Tanenhaus, 2005; Salverda & Tanenhaus, 2010).

In task-based visual world studies, in which a participant carries out spoken instructions, additional routines are necessary to enable the participant to execute the motor behaviors that constitute the task (e.g. clicking on or moving an object in a visual display on a computer screen). Thus, the eye movement record in task-based visual world studies usually consists of a succession of different patterns of eye movements, which each consistently reflect the execution of clearly distinguishable task-related routines.

Previous research has shown that goal representations in working memory include only immediately task-relevant components of the visual display (Patsenko & Altmann, 2010); only local subtasks or routines are predicted to significantly affect patterns of fixations. This prediction explains why potential referents that do not match the affordances of a local goal attract relatively few fixations, as in the Chambers et al. (2002, 2004) studies. It further predicts that even basic word recognition phenomena, such as looks to a cohort competitor, would be sharply reduced (even when the referent is in foveal vision) when that referent is incompatible with a local goal. The goal-based view also predicts that looks to the referent of an anaphoric expression, such as a pronoun, would be reduced when the referent is highly salient in memory or is not relevant to a local task. Thus, looks to a referent – or the absence of looks to a referent – do not provide direct evidence about when a referential expression has been interpreted.

The mapping from a goal to a set of routines is straightforward in studies in which the participant has an explicit task (e.g., Allopenna et al., 1998, and variants of this paradigm). When the goal is less articulated, for instance in “look-and-listen” studies in which participants are not given an explicit task, we can assume that these routines might also be activated and thus compete for control of saccades, much as high saliency areas might attract saccades in viewing a scene without a explicit task. Thus, the goal-based linking hypothesis can accommodate evidence for basic effects of visual-linguistic integration that we might consider to be “automatic” (see Salverda & Altmann, in revision, for evidence that task-irrelevant named objects can capture attention). However, according to the goal-based view,

these automatic effects arise from routines that constitute the bottom level of a hierarchically organized goal structure. The magnitude of these effects should therefore be influenced by the subgoals dominating the routines in this hierarchical structure. In look-and-listen studies, then, we might expect to see a smaller proportion of saccades to potential referents than we would see in goal-based tasks, where the visual routines that control fixations are guided by and reinforced by components of the task. Most importantly, however, the goal-based view predicts more fixations to objects that are task-relevant, and fewer fixations to objects that provide an equally good match to the linguistic input but that are not task-relevant.

Comparisons of linking hypotheses

Although few studies have directly compared goal-based linking hypotheses with comprehension-based hypotheses, there are some suggestive results in the literature that support the goal-based approach. Studies using the visual world paradigm typically use either the task-based or the look-and-listen version of the paradigm. This makes it hard to establish if, and to what degree, eye movements in visual world experiments are strongly influenced by the presence or absence of an explicitly defined task. However, the classic paper that reintroduced the look-and-listen version of the visual world paradigm (Altmann & Kamide, 1999) actually showed clear effects of experimental task. Participants viewed visual scenes that depicted a person situated amongst several objects. For instance, a scene might show a boy sitting on the floor, surrounded by a ball, a cake, a toy train, and a toy truck. In an initial study, participants performed a sentence verification task while their eye movements were recorded. They heard a spoken sentence accompanying the visual scene, and assessed whether or not the sentence described something that could happen in the scene. To this end, participants pressed a “yes” or “no” key on a button box. On trials where the sentence plausibly described something that could happen in the scene, the target noun in the sentence was preceded by a verb with selectional restrictions that were either compatible with any of the objects in the scene (e.g., “The boy will move the cake”) or compatible with one particular object in the scene (e.g., “The boy will eat the cake”, when the cake was the only edible object in the scene). In both experimental conditions, hearing the target noun was more likely to trigger a fixation to the target object than to any other objects in the display, a result that was expected on the basis of Cooper’s (1974) results. Importantly, if the verb preceding the noun provided semantic constraints that were consistent with only one object in the display, participants were more likely to initiate a fixation to the target object prior to hearing the target noun. Thus, when listeners process spoken language, they can use verb-based constraints to generate expectations that restrict the domain of subsequent reference to particular entities in the visual world.

Altmann and Kamide next considered the role that the sentence verification task might have played in driving the anticipatory eye movements observed during the presentation of the verb. They hypothesized that “the requirement to make such judgments may have induced processing strategies which do not reflect normal processing” (p. 255). In order to address this concern, the initial study was replicated using the same stimuli. This time, the participants did not receive a task and instructions. Instead, they were only told the sequence of events that constituted a trial. The results of this look-and-listen experiment were qualitatively similar to those of the initial experiment. However, the anticipatory eye movements occurred slightly later during the processing of the verb, i.e., with a delay of about 350 ms compared to when the effect was observed in the initial task-based experiment. Moreover, participants were almost twice as likely to fixate relevant objects in the task-based version of the experiment than in the look-and-listen version of the experiment. This aspect of Altmann and Kamide’s results, which has not received much attention in the visual world literature, is a clear demonstration that the participant’s task can

have strong quantitative as well as qualitative effects on eye movements in visual world studies.

Effects of task-relevance on referential domains

The goal-based perspective predicts that the immediate relevance of visual information to goal execution is a key predictor of fixation patterns. A corollary of this claim is that factors influencing the difficulty of executing a task (e.g., the difficulty of grasping an object, the complexity of a sequence of movements or task-induced constraints on attentional or memory resources) are likely to affect the proportion of fixations to objects in a visual world study.

Support for these predictions comes from an experiment reported by Eberhard et al. (1995). In this experiment, participants viewed a display with playing cards and followed instructions such as “Put the five of hearts that is below the eight of clubs above the three of diamonds”. On trials of interest there were two fives of hearts, one of which was the target card. The displays for three “point of disambiguation” conditions are displayed in Figure 2. In all conditions, the target card (i.e., one of the fives of hearts) was directly below an eight of clubs. In the so-called early disambiguation condition only the target five was directly below another card. In the mid-disambiguation condition, the other five of hearts was below a card of a different denomination (e.g., a ten of clubs). In the late disambiguation condition, both fives were below cards with eights but the distractor card was below an eight of a different suit (e.g., an eight of spades). Importantly, there were relatively few looks to the eight of clubs in the early disambiguation condition. This finding can be interpreted in terms of the relative difficulty of evaluating information in the context of the task across conditions. Because the presence or absence of a card is an easier visual discrimination than identifying the specific features that distinguish two cards, such as the number or suit, attention was not needed to identify the eight of clubs.

Likewise, the nature of the task can influence the set of potential referents in a display that a listener is likely to look at. Examples of this type of effect of task come from a series of studies examining how listeners circumscribe referential domains, including the Chambers et al. (2002, 2004) studies that we mentioned earlier.

Brown-Schmidt and Tanenhaus (2008) found clear effects of task-relevance in a collaborative task. Pairs of participants worked together to arrange a set of Duplo™ blocks in a matching pattern. Partners were separated by a curtain and seated in front of a board with stickers and a resource area with blocks. Boards were divided into five distinct sub-areas, with 57 stickers representing the blocks. Stickers were divided between the boards; where one partner had a sticker, the other had an empty spot. Thirty-six blocks were assorted colored squares and rectangles. The participants’ task was to replace each sticker with a matching block and instruct their partner to place a block in the same location to make their boards match. The positions of the stickers were determined by the experimenter, allowing for experimental control over the layout of the board. The shapes of the sub-areas and the initial placement of the stickers were designed to create conditions where the proximity of the blocks and the constraints of the task were likely to influence the strategies adopted by the participants.

Brown-Schmidt and Tanenhaus found that for point-of-disambiguation manipulations only blocks that were both proximal and task-relevant were taken into account by the speaker when generating referring expressions and by the listener when interpreting them. For example, speakers would use instructions such as “put the green block above the red block” in situations where there were two red blocks in a sub-area, but one of the red blocks could not be the referent because it didn’t have an empty space above it. Crucially, there was no

evidence that listeners were confused by the presence of two red blocks, and no evidence that the red block with no space above it attracted more fixations than any of the unrelated blocks. Similarly, Brown-Schmidt and Tanenhaus found that a cohort competitor (e.g., a block with a picture of a cloud) did not attract fixations when the speaker use the word "clown" to refer to a nearby block with a picture of a clown, during the interactive task, presumably because the cloud was outside of the relevant local domain. These results suggest that visual search did not take into account task-irrelevant blocks. This finding bears a close resemblance to the types of findings we reviewed from visual search in natural tasks, where salient but task-irrelevant objects do not attract fixations.

Developing goal-based linking hypotheses: future challenges

If the goal-based hypothesis is correct, then accounting for patterns of fixations will require explicitly modeling the participant's goal structure. Under relatively simple circumstances, e.g., a simple instruction and a simple display depicting a single event, comprehension-based models might approximate the goal structure. However, models that account for a high proportion of fixations are likely to require explicit tasks and explicit models of the goal structure.

We want to emphasize, however, that merely using a task is insufficient to meet the challenges associated with the goal-based approach. The goal-based view assumes that the instruction in a task-based visual world experiment triggers a basic language-vision routine, namely, e.g., mapping a word onto a potential referent for purposes of visually-guided reaching. When nearly all fixations begin on a fixation cross and the instructions are simple, as in Allopenna et al. (1998), it should in principle be possible to account for most of the variance in the timing and location of fixations with a model that focuses on how the participant's execution of the task mediates the mapping between spoken word recognition and visually identified referential candidates. Indeed, this was true for the models developed by Allopenna et al. (1998), Dahan, Magnuson, Tanenhaus, and Hogan (2001), Dahan, Magnuson, and Tanenhaus (2001), and Magnuson, McMurray, Tanenhaus, and Aslin (2003).

Once we move toward tasks involving slightly more complex instructions, however, developing a task model is more challenging and without such a model we lose a clear link between fixations and the unfolding language. Thus, we may encounter a weak form of an obstacle that plagued scientists who attempted to build upon Yarbus's classic work: Whereas it is easy to demonstrate that task affects fixation patterns, interpreting the processes underlying a pattern of fixations can be extremely difficult (for a valuable discussion, see Vivianni, 1990). For example, in the visual world experiments reported in Tanenhaus et al. (1995) and Spivey et al. (2002), participants followed instructions such as "Put the apple on the towel into the box". The crucial conditions involved three types of displays, each of which included an apple on a towel, another towel, and a box, as illustrated in Figure 3. The data of most interest were the proportion of fixations to the empty towel. Examination of the proportion of fixations over time (Figures 4–6 in Spivey et al., 2002) demonstrates tight time-locking of fixations to the initial referent or potential referents (a singleton apple) and the goal, with the proportion of looks reaching asymptotes of over .8. In contrast, participants made many fewer fixations to the "incorrect" or "garden-path" goal, the empty towel. Even in the condition that is interpreted as showing the largest garden-path effect, (Fig. 3a) the proportion of looks to the empty towel never rises above .2. This same pattern is observed in other studies using similar structures (e.g., Chambers et al., 2004; Trueswell et al., 1999; Novick et al., 2008).

Both of the objects which received a high proportion of fixations at a particular point in time were involved in visually guided reaching: the object to be grasped and moved (the apple)

and the location to which this object was to be moved (the box). It is less clear, however, to what we can attribute the fixations to the incorrect goal. These fixations could be influenced by a number of distinct subtasks engaged by the participant. For example, the stage of execution of the motor plan is likely to influence fixations to the empty towel. For example, participants on some proportion of trials may verify the expected goal with an eye movement before reaching for the apple. On other trials, they may instead have already begun reaching for the apple, and may not be ready to make an eye movement to the goal until after they have grasped the apple. This would decrease the probability of fixating on the incorrect goal. Moreover, the subgoal structures used to accomplish a task may differ across participants, and these differences may give rise to systematic variation in fixation patterns. For example, we have informally observed that participants who have expertise in video games or in a signed language make far fewer fixations prior to initiating an action. These results are consistent with a classic line of research initiated by Green and Bavelier (2003), demonstrating that experienced “gamers” process information more efficiently in the periphery.

Fixation patterns are also likely to change as the specific allocation of resources to immediately task-relevant information changes throughout an experiment. In most standard experiments in cognitive psychology, and in particular visual world experiments, factors such as learning, potential effects of contingencies, and individual differences in how efficiently information is accessed from a visual display are controlled for by standard practices such as counterbalancing items across conditions and averaging data from individual participants and items by condition, thereby reducing the effects of order of presentation on observed outcomes. There are, however, some potential problems associated with this approach. In particular, we are likely to observe effects of factors that have strong priors, but not important variables that might be subject to rapid perceptual learning. If, however, our goal is to model fixations under a goal-based hypothesis, then we need to consider how performance on a task typically changes as a function of experience and learning.

Learning effects can have a considerable impact on an individual’s fixation patterns over the course of an experiment. For example, studies examining the effects of fine-grained acoustic-phonetic variables, such as voice onset time (VOT), often involve repeating pictures many times throughout an experiment (e.g., McMurray, Tanenhaus, & Aslin, 2002; Clayards, Tanenhaus, Aslin, & Jacobs, 2008). In these studies, we frequently observe that the proportion of non-target-related fixations decreases and the proportion of trials in which a mouse-click is not preceded by a look to the target increases.

Within a task-based framework that takes into account effects of learning, it may be possible to speculate further about effects that may be particular to task-based visual world experiments. In such experiments, participants must perform a specific action or set of actions to progress through the experiment. The execution of the visuomotor routines necessary to complete the experimentally defined task is therefore associated with some reward structure and is actively reinforced throughout the experiment. In contrast, look-and-listen experiments do not require participants to extract visual information from the environment to progress through the experiment. Thus, task-based and look-and-listen visual world experiments differ with respect to their reward structure, in that task-based experiments reinforce the use of visuomotor routines above and beyond the reinforcement associated with understanding the language with respect to the visual display. It is possible that this difference in reward structure causes eye movements associated with linguistic processing to be more frequent and more rapidly deployed in task-based experiments than in passive listening experiments. It is also possible that the disparities between two experiments differing only in the presence or absence of a task (e.g. in Altmann & Kamide,

1999, which to our knowledge is the only study that reports such a comparison) would increase in magnitude over the course of the experiment, because effects of implicit reward structure may incrementally change participants' behaviors over time.

Implications and conclusions

What are the implications of the goal-based hypothesis for how we understand the link between language and visual search?

The first implication is that the conclusions we can draw from fixation patterns are limited by our understanding of the task, which includes the participant's goals and the representation of the visual world, which itself will be affected by the task. These predictions contrast with the assumption of a comprehension-based hypothesis, according to which the initial deployment of attention and the construction of visual and linguistic representations involve processes that, to a first approximation, occur automatically and independently of the observer's goals, expectations, and experience with the experimental task.

From the goal-based perspective, participants in a visual world experiment are always engaging in a task, regardless of whether the task is an explicit component of the experimental design. We cannot "control out" task variables by using "passive" listening paradigms. The task in look-and-listen studies is implicit and, according to a goal-based linking hypothesis, exerts primary control over the pattern and timing of fixations. Given the assumption that look-and-listen studies involve an implicit task, the goal-based approach predicts some commonalities and some differences between task-based and look-and-listen versions of visual world studies. For example, the common goal of language comprehension should have similar effects on patterns of fixations in task-based and look-and-listen experiments. A listener trying to make sense of possible links between the visual display and the language must process the visual world, which acts to constrain the interpretation of the linguistic input or to enrich the representation of the meaning of the language. On the other hand, because the participant's task in look-and-listen experiments is not defined explicitly and does not require an overt behavioral response, participants may be more likely to adopt idiosyncratic goals and task structures during the course of the study, or change the task structure throughout the experiment. This may result in higher variability in behavior among participants—for instance, in the degree to which they pay attention to aspects of the visual scene and the spoken language. In addition, the absence of clear reference points that can be used for the interpretation of participants' eye movements (e.g., behavioral responses or eye movement behaviors that can be related to task components) may complicate the understanding of fixation patterns obtained throughout a trial. Several studies in the literature, for example, those that have focused on the processing of fine-grained within-category information in speech processing, draw conclusions that are dependent upon separating fixations based on the participant's response in a task (e.g., McMurray, Tanenhaus & Aslin, 2002, 2009; also see Runner, Sussman & Tanenhaus, 2003, 2006, for extensions of response-contingent analyses to issues in higher-level language processing).

If we adopt the goal-based linking hypothesis, then the exciting but daunting challenge will be to construct task models that integrate language, vision, and action. We believe that these models will require some key components. First, they will need to specify hierarchically organized goal structures whose elementary units are behavioral routines. The linking hypothesis relating these goal structures and fixation patterns will need to take into account effects of learning and the relevance of visual information for participants' immediate goals. This will require treating objects in a display as representing potential states, where the variables of interest affect the probability of making a saccade to transition from one state to

another state at any point in time depending on the participant's immediate goals. These models will therefore also require analysis techniques that take into account state and state history.

There are advantages and disadvantages to pursuing the goal-based approach. The disadvantages are likely to occur to most scientists in psycholinguistics and visual cognition. Introducing complexity complicates data analyses, reduces the control of the experimenter and makes it more difficult to isolate effects of single variables. However, one of the lessons from the literature on vision in natural tasks is that in the absence of using a rich situation, we may be more likely to overemphasize the importance of the variables we are focusing on, potentially confusing statistical significance with overall importance. Most importantly, however, we may only see the underlying simplicity of the system when we observe and model its behavior in complex situations.

Acknowledgments

This research was partially supported by NIH grants HD27206 and DC0005071 to MKT and a Javits fellowship and NSF pre-doctoral fellowship to MB. We would like to thank Mary Hayhoe for invaluable input and comments and Gerry Altmann for helpful discussions.

References

Allopenna PD, Magnuson JS, Tanenhaus MK. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of Memory and Language. 1998; 38(4):419–439.

Altmann GTM. Language-mediated eye movements in the absence of a visual world: The 'blank screen paradigm'. Cognition. 2004; 93:B79–B87. [PubMed: 15147941]

Altmann GTM, Kamide Y. Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition. 1999; 73(3):247–264. [PubMed: 10585516]

Altmann GTM, Kamide Y. The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. Journal of Memory and Language. 2007; 57(4):502–518.

Altmann GTM, Kamide Y. Discourse mediation of the mapping between language and the visual world: Eye movements and mental representation. Cognition. 2009; 111(1):55–71. [PubMed: 19193366]

Altmann G, Steedman M. Interaction with context during human sentence processing. Cognition. 1988; 30(3):191–238. [PubMed: 3215002]

Alvarez GA, Cavanagh P. The capacity of visual short-term memory is set both by visual information load and by number of objects. Psychological Science. 2004; 15(2):106–111. [PubMed: 14738517]

Ballard DH, Hayhoe MM, Pelz JB. Memory representations in natural tasks. Journal of Cognitive Neuroscience. 1995; 7(1):66–80.

Ballard DH, Hayhoe MM, Pook PK, Rao RPN. Deictic codes for the embodiment of cognition. Behavioral and Brain Sciences. 1997; 20(4):723–767. [PubMed: 10097009]

Brown-Schmidt S, Tanenhaus MK. Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. Cognitive Science. 2008; 32(4):643–684. [PubMed: 19890480]

Buswell, GT. How people look at pictures. Chicago: University of Chicago Press; 1935.

Chambers CG, Tanenhaus MK, Eberhard KM, Filip H, Carlson GN. Circumscribing referential domains during real-time language comprehension. Journal of Memory and Language. 2002; 47(1):30–49.

Chambers CG, Tanenhaus MK, Magnuson JS. Actions and affordances in syntactic ambiguity resolution. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2004; 30(3): 687–696.

Clayards M, Tanenhaus MK, Aslin RN, Jacobs RA. Perception of speech reflects optimal use of probabilistic speech cues. Cognition. 2008; 108(3):804–809. [PubMed: 18582855]

Cooper RM. 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. 1974; 6(1):84–107.

Dahan D, Magnuson JS, Tanenhaus MK. Time course of frequency effects in spoken word recognition: Evidence from eye movements. Cognitive Psychology. 2001; 42(4):317–367. [PubMed: 11368527]

Dahan D, Magnuson JS, Tanenhaus MK, Hogan EM. Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. Language and Cognitive Processes. 2001; 16(5–6):507–534.

Dahan D, Tanenhaus MK. Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. Psychological Bulletin & Review. 2005; 12(3):455–459.

Droll JA, Hayhoe MM. Trade-offs between gaze and working memory use. Journal of Experimental Psychology: Human Perception and Performance. 2007; 33(6):1352–1365. [PubMed: 18085948]

Eberhard KM, Spivey-Knowlton MJ, Sedivy JC, Tanenhaus MK. Eye-movements as a window into real-time spoken language comprehension in natural contexts. Journal of Psycholinguistic Research. l995; 24(6):409–436. [PubMed: 8531168]

Ferreira F, Tanenhaus MK. Introduction to special issue on language-vision interactions. Journal of Memory and Language. 2007; 57(4):455–459.

Green CS, Bavelier D. Action video game modifies visual selective attention. Nature. 2003; 423(6939):534–537. [PubMed: 12774121]

Hayhoe M. Vision using routines: A functional account of vision. Visual Cognition. 2000; 7(1–3):43–64.

Hayhoe MM, Shrivastava A, Mruczek R, Pelz JB. Visual memory and motor planning in a natural task. Journal of Vision. 2003; 3(1):49–63. [PubMed: 12678625]

Henderson, JM.; Ferreira, F. The interface of language, vision, and action: Eye movements and the visual world. New York: Psychology Press; 2004.

Henderson JM, Malcolm GL, Schandl C. Searching in the dark: Cognitive relevance drives attention in real-world scenes. Psychonomic Bulletin & Review. 2009; 16(5):850–856. [PubMed: 19815788]

Huettig F, Altmann GTM. Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. Visual Cognition. 2007; 15(8): 985–1018.

Irwin DE. Information integration across saccadic eye-movements. Cognitive Psychology. 1991; 23(3):420–456. [PubMed: 1884598]

Jackendoff, R. Foundations of language: Brain, meaning, grammar, evolution. Oxford: Oxford University Press; 2002.

Jovancevic-Misic J, Hayhoe M. Adaptive gaze control in natural environments. The Journal of Neuroscience. 2009; 29(19):6234–6238. [PubMed: 19439601]

Kaiser E, Trueswell JC. Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution. Language and Cognitive Processes. 2008; 23(5):709–748.

Knoeferle P, Crocker MW. The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking. Cognitive Science. 2006; 30(3):481–529.

Koch C, Ullman S. Shift in selective visual attention: Towards the underlying neural circuitry. Human Neurobiology. 1985; 4(4):219–227. [PubMed: 3836989]

Land MF, Fernald RD. The evolution of eyes. Annual Review of Neuroscience. 1992; 15:1–29.

Land MF, Lee DN. Where we look when we steer. Nature. 1994; 369(6483):742–744. [PubMed: 8008066]

Land MF, McLeod P. From eye movements to actions: How batsmen hit the ball. Nature Neuroscience. 2000; 3(12):1340–1345.

Land M, Mennie N, Rusted J. The roles of vision and eye movements in the control of activities of daily living. Perception. 1999; 28(11):1311–1328. [PubMed: 10755142]

Levy R. Expectation-based syntactic comprehension. Cognition. 2008; 106(3):1126–1177. [PubMed: 17662975]

MacDonald MC, Pearlmutter NJ, Seidenberg MS. The lexical nature of syntactic ambiguity resolution. Psychological Review. 1994; 101(4):676–703. [PubMed: 7984711]

Magnuson JS, Tanenhaus MK, Aslin RN, Dahan D. The time course of spoken word learning and recognition: Studies with artificial lexicons. Journal of Experimental Psychology: General. 2003; 132(2):202–227. [PubMed: 12825637]

Mayberry MR, Crocker MW, Knoeferle P. Learning to attend: A connectionist model of situated language comprehension. Cognitive Science. 2009; 33(3):449–496. [PubMed: 21585477]

McMurray B, Tanenhaus MK, Aslin RN. Gradient effects of within category phonetic variation on lexical access. Cognition. 2002; 86(2):B33–B42. [PubMed: 12435537]

McMurray B, Tanenhaus MK, Aslin RN. Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. Journal of Memory and Language. 2009; 60(1):65–91. [PubMed: 20046217]

Novick JM, Thompson-Schill SL, Trueswell JC. Putting lexical constraints in context into the visual-world paradigm. Cognition. 2008; 107(3):850–903. [PubMed: 18279848]

O'Regan JK. Solving the "real" mysteries of visual perception: The world as an outside memory. Canadian Journal of Psychology. 1992; 46(3):461–488. [PubMed: 1486554]

Parkhurst D, Law K, Niebur E. Modeling the role of salience in the allocation of overt visual attention. Vision Research. 2002; 42(1):107–123. [PubMed: 11804636]

Patsenko EG, Altmann EM. How planful is routine behavior? A selective-attention model of performance in the Tower of Hanoi. Journal of Experimental Psychology: General. 2010; 139(1): 95–116. [PubMed: 20121314]

Runner JT, Sussman RS, Tanenhaus MK. Assignment of reference to reflexives and pronouns in picture noun phrases: Evidence from eye movements. Cognition. 2003; 81:B1–B13. [PubMed: 12893125]

Runner JT, Sussman RS, Tanenhaus MK. Processing reflexives and pronouns in picture noun phrases. Cognitive Science. 2006; 30(2):193–241.

Sailer U, Flanagan JR, Johannson RS. Eye-hand coordination during learning of a novel visuomotor task. Journal of Neuroscience. 2005; 25(39):8833–8842. [PubMed: 16192373]

Salverda AP, Altmann GTM. Attentional capture of objects referred to by spoken language. Journal of Experimental Psychology: Human Perception and Performance. (in press).

Salverda AP, Tanenhaus MK. Tracking the time course of orthographic information in spoken-word recognition. Journal of Experimental Psychology: Learning, Memory, & Cognition. 2010; 36(5): 1108–1117.

Sedivy JC, Tanenhaus MK, Chambers CG, Carlson GN. Achieving incremental semantic interpretation through contextual representation. Cognition. 1999; 71(2):109–147. [PubMed: 10444906]

Simons DJ, Levin DT. Change blindness. Trends in Cognitive Sciences. 1997; 1(7):216–267.

Spivey MJ, Geng JJ. Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. Psychological Research. 2001; 65(4):235–241. [PubMed: 11789427]

Spivey MJ, Tanenhaus MK, Eberhard KM, Sedivy JC. Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. Cognitive Psychology. 2002; 45(4):447–481. [PubMed: 12480476]

Sprague N, Ballard D, Robinson A. Modeling embodied visual behaviors. ACM Transactions on applied perception. 2007; 4(2)

Tanenhaus, MK. Eye movements and spoken language processing. In: van Gompel, RPG.; Fischer, MH.; Murray, WS.; Hill, RL., editors. Eye movements: A window on mind and brain. Oxford: Elsevier; 2007a. p. 443-469.

Tanenhaus, MK. Spoken language comprehension: Insights from eye movements. In: Gaskell, G., editor. Oxford Handbook of Psycholinguistics. Oxford: Oxford University Press; 2007b. p. 309-326.

Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC. Integration of visual and linguistic information in spoken language comprehension. Science. 1995; 268(5217):1632–1634. [PubMed: 7777863]

Tanenhaus, MK.; Trueswell, JC. Eye movements and spoken language comprehension. In: Traxler, MJ.; Gernsbacher, MA., editors. Handbook of Psycholinguistics. (Second edition). New York: Elsevier Academic Press; 2006. p. 863-900.

Treisman AM, Gelade G. Feature-integration theory of attention. Cognitive Psychology. 1980; 12(1): 97–136. [PubMed: 7351125]

Trueswell JC, Sekerina I, Hill NM, Logrip ML. The kindergarten-path effect: Studying on-line sentence processing in young children. Cognition. 1999; 73(2):89–134. [PubMed: 10580160]

Trueswell JC, Tanenhaus MK, Garnsey SM. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. Journal of Memory and Language. 1994; 33(3): 285–318.

Turano KA, Geruschat DR, Baker FH. Oculomotor strategies for the direction of gaze tested with a real-world activity. Vision Research. 2003; 43(3):333–346. [PubMed: 12535991]

Ullman S. Visual routines. Cognition. 1984; 18(1–3):97–157. [PubMed: 6543165]

van Gompel, RPG.; Fischer, MH.; Murray, WS.; Hill, RL. Eye movements research: An overview of current and past developments. In: van Gompel, RPG.; Fischer, MH.; Murray, WS.; Hill, RL., editors. Eye movements: A window on mind and brain. Oxford: Elsevier; 2007. p. 1-28.

Viviani, P. Eye movements in visual search: Cognitive, perceptual, and motor control aspects. In: Kowler, E., editor. Eye movements and their role in visual and cognitive processes: Reviews of oculomotor research. Amsterdam: Elsevier; 1990. p. 353-393.

Yarbus, AL. Eye movements and vision. New York: Plenum Press; 1967.

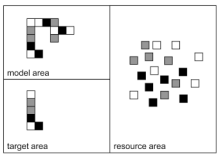

Figure 1.
Experimental setup used in Ballard et al.'s (1995) block-copying task. The participant's task is to copy the block pattern in the model area using blocks from the resource area.

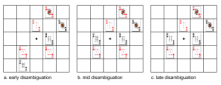

Figure 2.
Three point of disambiguation conditions in the Eberhard et al. (1995) experiment. Participants heard the spoken instruction “Put the five of hearts that is below the eight of clubs above the three of diamonds”.

Figure 3.
Three experimental conditions in the Tanenhaus et al. (1995) and Spivey et al. (2002) experiments. Participants heard the spoken instruction “Put the apple on the towel into the box”.