Chapter 22
Eye Movements and Spoken Language Comprehension

Michael K. Tanenhaus and John C. Trueswell

ABSTRACT

This chapter provides an overview of recent research that uses eye movements to investigate spoken language comprehension. We outline the logic of what is now commonly referred to as the "visual world" paradigm and review some of the foundational studies. We then use some sample experiments to review methodological issues, including issues of data analysis, linking hypotheses, and issues that arise when combining language, vision, and action. We conclude with a brief review of some domains within psycholinguistics in which the visual world paradigm is beginning to play a prominent role.

1. INTRODUCTION

Many everyday tasks require people to rapidly interrogate their visual surroundings. Reading a magazine, looking for a friend at a party, and making breakfast, all require people to frequently shift their attention to task-relevant regions of the visual world. These shifts of attention are accompanied by shifts in gaze, accomplished by ballistic eye movements known as *saccades*, which bring the attended region into the central area of the fovea, where visual acuity is greatest. The pattern and timing of saccades, and the resulting fixations, are one of the most widely used response measures in the brain and cognitive sciences, providing important insights into the functional and neural mechanisms underlying attention, perception, and memory (for reviews, see Liversedge & Findlay, 2000; Rayner, 1998). Eye movements are now an important measure in the perception–action literature, especially for studies examining allocation of attention in natural everyday tasks (Hayhoe & Ballard, 2005; Land, 2004). Within psycholinguistics, eye movements have been one of the most widely used response measures in studies of written word recognition and sentence reading for more than two decades, initiated by the classic work of McConkie and Rayner (1976), Frazier and Rayner (1982), and Just and Carpenter (1980). For reviews, see Rayner (1998, this volume).

Handbook of Psycholinguistics: 2nd Edition
ISBN: 0-12-369374-8

More recently, eye movements have become a widely used response measure for studying spoken language processing in both adults and children, in situations where participants comprehend and generate utterances that are about a circumscribed "visual world." Researchers are now using this method to address issues that run the gamut of current topics in language processing. Eye movements are a response measure of choice for studies addressing many classical questions in psycholinguistics, e.g., is the processing of stop consonants categorical (McMurray, Tanenhaus, & Aslin, 2002); does context influence the earliest moments of temporary lexical and syntactic ambiguity resolution (Dahan & Tanenhaus, 2004; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002); what is the locus of frequency effects in spoken word recognition (Dahan, Magnuson, & Tanenhaus, 2001a); what factors influence the time course with which anaphoric expressions, such as pronouns, are resolved (Arnold, Eisenbad, Brown-Schmidt, & Trueswell, 2000; Järvikivi, van Gompel, Hyönä, & Bertram, 2005) and, for bilingual speakers, does a word spoken in one language activate the lexical representations of similar sounding words in the other language (Spivey & Marian, 1999; Ju & Luce, 2004).

The visual world paradigm has also opened up relatively uncharted territory in language comprehension, including real-time sentence processing in children (Trueswell, Sekerina, Hill, & Logrip, 1999); the role of common ground in on-line processing (Keysar, Barr, Balin, & Brauner, 2000; Hanna, Tanenhaus, & Trueswell, 2003); how listeners make use of disfluencies in real-time language processing (Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Bailey & Ferreira, 2005; Ferreira & Bailey, in press); and how participants in a conversation coordinate their referential domains (Brown-Schmidt, Campana, & Tanenhaus, 2005; Tanenhaus & Brown-Schmidt, in press). Finally, the visual world approach has spawned a new family of studies investigating the interface between action and language and between vision and language (Chambers, Tanenhaus & Magnuson, 2004; Spivey et al., 2002; Kamide, Altmann, & Haywood, 2003; Knoeferle, Crocker, Scheepers, & Pickering, 2005).

Why has the visual world paradigm gained traction so rapidly? First, in contrast to reading, time-locked, relatively natural measures of spoken language processing have been hard to come by. Many of the most widely used tasks for studying spoken language comprehension present only a snapshot of processing at a single point in time, require meta-linguistic judgments, and interrupt the flow of the speech input. In contrast, eye movements provide a sensitive, implicit measure of spoken language processing in which the response is closely time-locked to the input without interrupting the speech stream. Second, the eye-movement paradigm can be used with natural tasks that do not require meta-linguistic judgments. This makes it well suited for studies with young children (Trueswell et al., 1999) and with special populations (Yee, Blumstein, & Sedivy, 2000). Third, the coupling of a visual world with language makes it possible to ask questions about real-time interpretation, especially questions about reference that would be difficult to address, and perhaps would be intractable, if one were limited to measures of processing complexity (e.g., Sedivy, Tanenhaus, Chambers, & Carlson, 1999). It also makes it possible to examine questions at the interface between language, perception, and action (see the chapters by Henderson & Ferreira, 2004 and Trueswell & Tanenhaus, 2005). Fourth, eye movements can be used to study issues about the relationship between real-time message planning and utterance planning (Bock, Irwin, & Davidson, 2004; Griffin, 2004; Brown-Schmidt & Tanenhaus, in press). Finally, the paradigm allows one to study real-time production and comprehension simultaneously in natural tasks involving conversational interaction. This makes it possible to bridge the two dominant traditions in language-processing research: the "language-as-action" tradition, which has focused on natural interactive conversation while generally ignoring questions about the time course of real-time language processing and the "language-as-product" tradition, which has focused on the time course of processing while being primarily limited to "de-contextualized language" (Clark, 1992; Tanenhaus & Trueswell, 2005).

Our goal in this chapter is to provide an introduction and overview to the rapidly growing literature on eye movements and spoken language processing, focusing on applications to spoken language comprehension. Section 2 focuses on methodological issues. As with any new paradigm, excitement about novel findings and new arenas of investigation must be tempered with concerns about the nature of the paradigm itself, including task-specific strategies, and the assumptions that link the behavioral measure to the hypothesized underlying mechanisms. Major topics include the logic linking eye movements to spoken language processing, how eye-movement data are collected and analyzed, sample applications illustrating some of the paradigms, including comparisons to eye-movement reading studies, and associated experimental logics, and finally, concerns and limitations that arise in examining language in a circumscribed visual world. In addressing these issues, we review results from a number of visual world studies. Section 3 presents a selective review of some of the major lines of research that this method has opened up, focusing on topics in language comprehension, including spoken word recognition, use of referential constraints in parsing, issues that arise in interactive conversation and the development of language processing abilities in children. Before turning our attention to these two major sections, we briefly review some of the foundational studies in the eye-movement literature on spoken language processing.

1.1. Some Foundational Studies

1.1.1. Comprehension

The use of eye movements as a tool for studying spoken language comprehension was pioneered by Roger Cooper (1974) in a remarkable article, presciently 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 participant's eye movements 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 to words in the story. Moreover, fixations were often generated before the end of the word.

Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy (1995) initiated the recent wave of visual world studies, taking advantage of the advent of accurate lightweight head-mounted eye-trackers. Tanenhaus et al. examined eye movements as participants followed instructions to perform simple tasks with objects in a workspace. They found that varying the number of potential referents for a temporarily ambiguous prepositional phrase (e.g., *Put the apple on the towel...*) determined whether the phrase was initially parsed as a goal argument (where to put the apple) or as a modifier (the location of the apple to be moved), as predicted by Altmann and Steedman (1988). (A more complete report of the Tanenhaus et al. study is presented in Spivey, et al., 2002.)

Trueswell, Skerina, Hill, and Logrip (1999) replicated the Tanenhaus et al. (1995) study with adults, and more importantly extended it to five-and eight-year-old children. They found important developmental differences in how children weight lexical and referential constraints on sentence parsing, laying the foundation for the rapidly expanding field of online sentence processing in preliterate children.

Eberhard, Spivey- knowlton, Sedivy, and Tanenhaus (1995) demonstrated that fixations to entities referred to in an instruction are remarkably time-locked to the unfolding utterance. Fixations to a target referent among a display of competitors occurred as soon as continuous integration of constraints provided by both the unfolding speech and the visual display could, in principle, distinguish the referent from its competitors. These results obtained both for simple instructions (*touch the starred red square*) and complex instructions (*Put the five of hearts that's below the eight of clubs above the three of diamonds*). This "point-of disambiguation" logic is now widely used in studies of reference resolution.

Sedivy initiated an influential line of research demonstrating that pre-nominal scalar adjectives, such as *tall*, affect the point of disambiguation of potential referents in referential expressions, such as *the tall glass*. Speakers use, and listeners interpret, scalar adjectives contrastively, that is, to distinguish between two or more objects of the same type (Sedivy, et al., 1999; Sedivy, 2003). For example, in a display with a tall glass, a speaker will typically not use the adjective *tall,* unless the display contains, as a potential contrast, another, smaller glass (Sedivy, 2003). As they hear *tall*, eye movements show that listeners immediately interpret *the tall glass* as referring to the taller of two glasses, even when another taller object, e.g., a *pitcher* is present, whereas in the absence of a potential contrast, fixations to the glass do not begin until after the listener hears *glass* (Sedivy et al., 1999). In addition to being interesting in their own right, the processing of pre-nominal adjectives has become an important methodological tool for addressing a range of issues in language processing.

Building on initial results by Spivey-Knowlton (1996), Allopenna, Magnuson, and Tanenhaus (1998) demonstrated that the timing of fixations to a pictured referent, and competitors with different types of phonological overlap, was sufficiently time-locked to the input so as to trace the time course of lexical access. Allopenna et al. also showed that a simple linking hypothesis could be used to map fixations onto computational models of lexical activation, thus laying the foundation for the growing body of work that uses the visual world paradigm to study spoken word recognition.

Altmann and Kamide (1999) made a seminal contribution to the visual world paradigm by demonstrating linguistically mediated, anticipatory eye movements using a task like Cooper's in which participants listened to a description of an upcoming event involving entities depicted in a display. As participants heard sentences such as, *the boy will eat the cake*, they made anticipatory eye movements to a picture of cake before the offset of *eat*, when the other depicted objects were not edible. Anticipatory eye movements are now widely used as a dependent measure, typically with this so-called, passive listening (non-action-based), variant of the visual world paradigm.

In an ingenious experiment by Keysar and colleagues (Keysar, Barr, Balin, & Brauner, 2000), eye movements were used to evaluate when in the time course of comprehension listeners take into account common ground information, i.e., information that is shared with an interlocutor. A confederate speaker, the director, instructed a naive participant, the matcher, to move objects in a box with cubbyholes. Most objects could be seen by both the speaker and the matcher, and thus were in common ground by virtue of physical co-presence (Clark, & Marshall, 1981). However, some objects were blocked from the speaker's view by an opaque barrier, and were therefore only in the matcher's privileged ground. Nonetheless, the matcher looked at these objects when they, along with an object in common ground, were consistent with the speaker's referential description. This (controversial) study has laid the groundwork for investigations of how interlocutors make use of each other's likely knowledge and intentions in real-time language comprehension.

1.1.2. Production

Two studies laid the foundation for using eye movements to study language production. Meyer, Sleiderink, and Levelt (1998) had participant's name sequences of objects. Eye gaze was tightly coordinated with the speech. Participants fixated a to-be-named object about 1 s prior to the onset of naming. This eye-voice lag is similar to the time it takes to initiate naming an object in isolation (Rossion & Pourtois, 2004; Snodgrass & Yuditsky, 1996), suggesting that the eye-voice delay reflects word preparation.

Griffin and Bock (2000) presented participants with a simple event rendered as a line drawing that could be described with either an active or passive sentence, such as a woman shooting a man. The sequence of eye movements reflected the order of constituents in the utterance. Speakers looked at pictured objects about 800 ms to 1 s before naming them. Once speaking began, the sequence and timing of fixations was controlled by the utterance, rather than perceptual properties of the input, suggesting that the speaker had completed message planning prior to beginning to speak (also see Bock, Irwin, Davidson, & Levelt, 2003).

2. METHODOLOGICAL ISSUES

These early studies have raised numerous methodological questions, many of which were highlighted by the authors themselves. We now review what we see as the most important of these issues.

2.1. Data Analysis and Linking Assumptions

We will use Experiment 1 from Allopenna et al. (1998) to briefly describe how eye-movement data are analyzed. This experiment will also prove useful later for discussing some of the methodological concerns that arise in visual world studies in language comprehension. Allopenna et al. (1998) evaluated the time course of activation for lexical competitors that were cohorts, that is, they shared initial phonemes with the target word (e.g., *beaker* and *beetle*) or that rhymed with the target word (e.g., *beaker* and *speaker*). Participants were instructed to fixate a central cross and then followed a spoken instruction to move one of four objects displayed on a computer screen with the computer mouse (e.g., *Look at the cross. Pick up the beaker. Now put it above the square*).

2.1.1. Data analysis

A schematic of a sample display of pictures is presented in Figure 1 (Panel A). The pictures include the target (the beaker), a cohort (the beetle), a rhyme (speaker), and an unrelated picture (the carriage). The particular pictures displayed are used to exemplify types of conditions and are not repeated across trials. For current purposes, we restrict our attention to the target, cohort, and unrelated pictures. Panel B shows five hypothetical trials. The 0 ms point indicates the onset of the spoken word *beaker*. The dotted line begins at about 200 ms–the earliest point where we would expect to see signal-driven fixations, give the 150–200 ms required to program and launch a saccade (Matin, Shao, & Boff, 1993). On the first trial, the hypothetical participant initiated a fixation to the target about 200 ms after the onset of the word, and continued to fixate on it (typically until the hand brings the mouse onto the target). On the second trial, the fixation to the target begins a bit later. On the third trial, the first fixation is to the cohort, followed by a fixation to the target. On the fourth trial, the first fixation is to the unrelated picture. The fifth trial shows another trial where the initial fixation is to the cohort. Panel C illustrates the proportion of fixations over time for the target, cohort, and unrelated pictures, averaged across trials and participants. These fixation proportions are obtained by determining the proportion of looks to the alternative pictures at a given time slice and they show how the pattern of fixations change as the utterance unfolds. The fixations do not sum to 1.0 as the word is initially unfolding because participants are often still looking at the fixation cross.

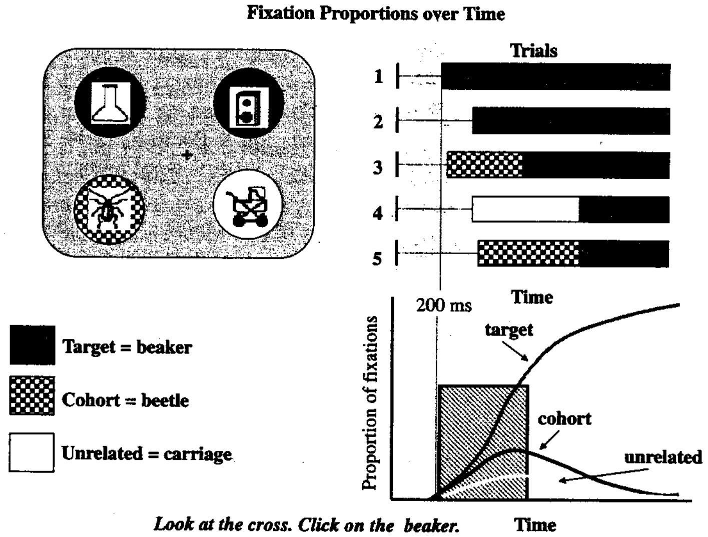

Figure 1. Sample data illustrating display, hypothetical data, proportion of fixation curves and regions of interest, modeled after Allopenna et al. (1998).

Although proportion of fixation curves might seem to imply that eye movements provide a continuous measure it is more accurate to say that eye movements can provide an approximation to a continuous measure. The assumption linking fixations to continuous word recognition processes is that as the instruction unfolds the probability that the listener's attention will shift to a potential referent of a referring expression increases with the activation (evidence for) of its lexical representation, with a saccadic eye movement typically following a shift in visual attention to the region in space where attention has moved. Because saccades are rapid, low-cost, low-threshold responses, a small proportion of saccades will be generated by even small increases in activation, with the likelihood of a saccade increasing as activation increases. Thus, while each saccade is a discrete event, the probabilistic nature of saccades ensures that with sufficient numbers of observations, the results will begin to approximate a continuous measure (see Spivey, Grosjean, & Knoblich, 2005; Magnuson, 2005).

A window of interest is often defined, as illustrated by the rectangle in Panel C. For example, one might want to focus on the fixations to the target and cohort in the region from 200 ms after the onset of the spoken word to the point in the speech stream where disambiguating phonetic information first arrives. The proportion of fixations to pictures or objects and the time spent fixating on the alternative pictures (essentially the area under the curve, which is a simple transformation of proportion of fixations) can then be analyzed. Because each fixation is likely to be 150–250 ms, the proportion of fixations in different time windows is not independent. One way of increasing the independence is to restrict the analysis to the proportion of new saccades generated to pictures within a region of interest. In the future, it will be important for psycholinguists to explore more sophisticated statistical methods for dealing with the temporal dependencies associated with how the linguistic input at time *t* effects location of fixations at subsequent temporal intervals.

Figure 2 (Panel A) shows the data from the Allopenna et al. (1998) experiment. The figure plots the proportion of fixations to the target, cohort, rhyme and unrelated picture. Until 200 ms, nearly all of the fixations are on the fixation cross. These fixations are not shown. The first fixations to pictures begin at about 200 ms after the onset of the target word. These fixations are equally distributed between the target and the cohort. These fixations are remarkably time-locked to the utterance: input-driven fixations occurring

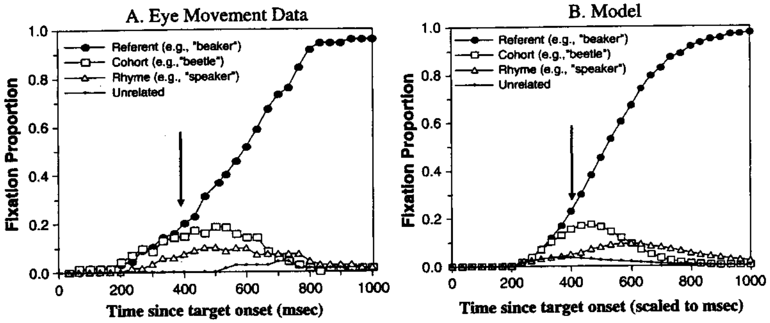

Figure 2. Data from Allopenna et al. (1998) and simulations generated by their linking hypothesis mapping activation in the TRACE model onto predicted proportion of fixations over time.

200–250 ms after the onset of the word are most likely programmed in response to information from the first 50 to 75 ms of the speech signal. At about 400 ms after the onset of the spoken word, the proportion of fixations to the target began to diverge from the proportion of fixations to the cohort. Subsequent research has established that cohorts and targets diverge ~200 ms after the first phonetic input that provides probabilistic evidence favoring the target, including coarticulatory information in vowels (Dahan, Magnuson, Tanenhaus, & Hogan, 2001b, Dahan & Tanenhaus, 2004).

Shortly after fixations to the target and cohort begin to rise, fixations to rhymes begin to increase relative to the proportion of fixations to the unrelated picture. This result supports continuous mapping models, such as TRACE (McClelland & Elman, 1986), which predict competition from similar words that mismatch at onset (e.g., rhymes), but is inconsistent with the cohort model of spoken word recognition and its descendents (e.g., Marslen-Wilson, 1987, 1990, 1993), which assume that any featural mismatch at the onset of a word is sufficient to strongly inhibit a lexical candidate.

2.1.2. *Formalizing a linking hypothesis*

The assumption providing the link between word recognition and eye movements is that the activation of the name of a picture determines the probability that a subject will shift attention to that picture and thus make a saccadic eye movement to fixate it. Allopenna et al. formalized this linking hypothesis by converting activations generated by a TRACE simulation into response strength, following the procedures outlined in Luce (1959). The Luce choice rule is then used to convert the response strengths into response probabilities.

The Luce choice rule assumes that each response is equally probable when there is no information. Thus when the initial instruction is "look at the cross" or "look at picture X," the response probabilities are scaled to be proportional to the amount of activation at each time step. Thus the predicted fixation probability is determined both by the amount of evidence for an alternative and the amount of evidence for that alternative compared to the other possible alternatives. Finally, a 200 ms delay is introduced because programming an eye movement takes ~200 ms (Matin et al., 1993). In experiments without explicit instructions to fixate on a particular picture, initial fixations are randomly distributed among the pictures. Under these conditions, the simple form of the choice rule can be used (see Dahan et al., 2001a, 2001b). Note that the Allopenna et al. formalization is only an approximation to what would be a more accurate formalization of the linking hypothesis which would predict the probability that a saccade would be generated at a particular point in time, contingent upon (a) the location of the previous fixation (and perhaps the several preceding fixations; (b) time from the onset of the last fixation and (c) the current goal state of the listener's task–which can be ignored in a simple "click" task like the Allopenna et al. paradigm.

When the linking hypothesis is applied to TRACE simulations of activations for the stimuli used by Allopenna et al., it generates the predicted fixations over time shown in Figure 2 (Panel B). The predictions for the target, the cohort competitor, and a rhyme competitor closely match the behavioral data.

2.1.3. *Action-contingent analyses*

One useful feature of combining eye movements with an action is that the behavioral responses reveal the participant's interpretation. This allows for *interpretation-contingent* analyses in which fixations are analyzed separately for trials on which participants choose a particular interpretation. Two recent applications, illustrate how interpretation-contingent analyses can be used to distinguish among competing hypotheses.

McMurray et al. (2002) used a variation on the Allopenna et al. task to investigate the hypothesis that lexical processing is sensitive to small-within category differences in Voice-Onset Time (VOT). The stimuli were synthesized minimal pairs that differed only in voicing, such as *bomb/palm* and *peach/beach*. VOT varied in 5 ms step sizes from 0 to 40 ms. McMurray et al. found gradient increases in looks to the cross-category competitor as the VOT moved closer to the category boundary. While these results are consistent with the hypothesis that lexical processing is sensitive to within category variation, the results could also be accounted for without abandoning the traditional assumption that within-category variation is quickly discarded by making the following plausible assumption that there is noise in the system. For example, assume a category boundary of ~18 ms. For trials with a VOT of 20 ms, given some noise, perhaps 20% of the stimuli might be perceived as having a VOT of <18 ms. With a VOT of 25 ms, the percentage might drop to 12%, compared to 8% for trials with a VOT of 30 ms and 4% for a VOT of 35 ms, etc. Thus, the proportion of looks to the cross-category competitor might increase as VOT approaches the category boundary because the data will include more trials where the target word was misheard as the cross-category competitor and not because the underlying system responds in a gradient manner.

McMurray et al. were able to rule out this alternative explanation by filtering any trials where the participant clicked on the cross-category picture. For example, if the VOT was 25 ms, and the participant clicked on the picture of the bomb, rather than the palm, then the eye-movement data from that trial would be excluded from the analyses. McMurray et al. found that looks to the cross-category competitor increased as VOT approached the category boundary, even when all "incorrect" responses were excluded from the analyses, thus providing strong evidence that the system is indeed gradient.

A second illustration comes from recent studies by Runner and his colleagues (e.g., Runner, Sussman, & Tanenhaus, 2003, in press) investigating the interpretation of reflexives and pronouns in so-called picture noun phrases with possessors, e.g., *Harry admired Ken's picture of him/himself.* Participants were seated in front of a display containing three male dolls, Ken, Joe, and Harry, each with distinct facial features. Digitized pictures of the doll's faces were mounted in a column on a board directly above each individual doll. The participant was told that each doll "owned" the set of pictures directly above him; that is, the three pictures in the column above Joe were Joe's pictures, the pictures in the column above Ken were Ken's pictures, etc.

Binding theory predicts that the reflexive, *himself*, will be interpreted as referring to Ken's picture of Ken in instructions such as *Pick up Harry. Now have Harry touch Ken's picture of himself.* Runner et al. found that looks to both the binding-appropriate and inappropriate referents began to increase compared to an unrelated picture in the same row, beginning about ~200 ms after the onset of the reflexive. This result suggests that both binding-appropriate and inappropriate referents are initially considered as potential referents for a reflexive. However, participant's choices showed frequent violations of classic binding for reflexives: on ~20% of trials with reflexives, participants had Harry touch Ken's picture of Harry. Thus, one might argue that the early looks to binding-inappropriate referents came from just those trials on which the participant arrived at the "incorrect" interpretation. Runner et al. were able to rule out this interpretation by analyzing just those trials where the participant made the binding-appropriate response, finding that there was still an increase in looks to the inappropriate referent compared to controls.

2.2. Task Variables

As the eye-movement literature on spoken language comprehension has developed, researchers have begun to vary the sorts of tasks given to their participants. The effects of these variations is important to evaluate and track from experiment to experiment since as discussed in the opening of this chapter, eye movement patterns are heavily task and goal-dependent (i.e., we shift our attention to *task-relevant* regions of the world). It would be a mistake for instance, to assume that the "task" involved in the studies discussed in this chapter can be monolithically described as "spoken language comprehension" or worse still "use of language." Very similar issues of task variation arise in reading eye-movement studies; eye-movement patterns over identical sequences of text will differ substantially depending on whether readers are *skimming*, *understanding*, *memorizing*, or *proofing*. Much greater opportunity for task variability appears to be possible in visual world studies because of the wide range of ways that participants can be asked to interact with the world. However, it is precisely this variability that provides experimenters with the leverage to make the visual world paradigm useful for such a wide range of questions.

One important task dimension is whether or not the linguistic stimuli used in the study involve instructions to act on the world. This variable is likely to be crucial because eye fixation plays an important role in visually guided reaching (see Hayhoe & Ballard, 2005). At one extreme, imperative sentences are commonly used, such that participants are required to manipulate the objects (e.g., *Pick up the ball. Put it inside the cup.*) At the other extreme, participants listen to declarative sentences, while looking at visually co-present referents. Here, the reference is intended to be non-deictic. (*The boy picked up the ball. Then he put it inside the cup.*)

Action-based studies offer several advantages in that participants are required in a highly natural way to remain engaged with their referent world; planning to execute a response requires calculating the spatial location of referents and presumably increases the time-locked nature of the relationship between linguistic interpretation and eye fixation. One clear limitation of the action-based paradigm however is that the linguistic stimuli must be embedded in instructions, which can limit the experimenter's degrees of freedom. The non-action-based listening procedure places far fewer constraints on both the experimenter and the participant. Decoupling fixations from action planning may also increase the proportion of anticipatory eye movements, which are extremely useful for inferring expectations generated by the listener.

Indeed, many of the most important applications of non-action-based listening have explored and documented referential expectations, starting with research initiated by Altmann and colleagues who showed that listeners can anticipate upcoming reference based on the semantic requirements of verbs and/or whole predicates (e.g., Altmann & Kamide, 1999; Kamide et al., 2003). Studies building upon this on this work include Boland (2005), who compared verb-based expectations for adjuncts and arguments, and Knoeferle and Crocker (in press) who studied the effects of visually based information on expectation about thematic role assignment.

We should note that this non-action paradigm is sometimes referred to as "passive" listening, and some investigators (e.g., Boland, 2005) have proposed that differences between fixations in action and passive listening tasks might be used to separate fixations that are controlled by language from those that are controlled by action. We are skeptical for several reasons. First, it is becoming increasingly clear that perception and action are inextricably intertwined in most perceptual domains, and we expect that this is also likely to be case for language. Second, interpreting sequences of fixations in the absence of an explicit task are likely to prove problematic for reasons eloquently articulated by Viviani (1990). We note however that many non-action task studies provide listeners with a well-defined task, typically so as to increase engagement with the scene and decrease the variability. For instance, Kaiser and Trueswell (2004) and Arnold et al. (2000) asked listeners to judge whether the depicted image on a trial matched the spoken description/story.

More generally it is important to keep in mind the following considerations. First, all saccadic eye movements involve some attentional overhead (Kowler, 1995). Second, the concept of passive listening leaves the underlying goals of the listener up to the listener.

Thus, each listener may adopt different goals, or worse, all listeners might adopt a pragmatically appropriate goal that was unforeseen by the experimenter. In short, there is no such thing as a taskless task. We therefore consider the notion of passive listening as akin to the notion of the null context, which is problematic for reasons articulated by Crain and Steedman (1985) and Altmann and Steedman (1988). Third, and perhaps most importantly, the difference between action-based (or perhaps more appropriately manipulation-based) and non-action-based variants of the visual world paradigm is really a subset of a more general question about the goal structures that control the moment-by-moment attentional state of the participants. In tasks with complex goal structures, e.g., a task-oriented dialog, multiple layers of goals will contribute to fixations, some of which may be are tied to expectations about upcoming linguistic input, some to the current subgoal, and some to higher-level planning.

Few studies to date have directly compared the action and non-action-based versions of the paradigm with the same materials (but cf., Sussman, 2006). However, to a first approximation, it appears that when anticipatory eye movements are excluded, the timing of fixations to potential referents may be slightly delayed in listening tasks compared to action-based tasks. The data from simple action-based tasks with imperatives (tasks where participants follow a sequence of instructions) is also somewhat cleaner than the data from non-actions-based tasks with declaratives, most likely because a higher proportion of the fixations are likely to be task-relevant.

2.3. Comparing Visual World and Eye-Movement Reading Studies

Many of the issues that have been investigated for decades using eye movements in reading, in particular issues in lexical processing and sentence processing are now being investigated using eye movements with spoken language. Although, some aspects of these processes will differ in reading and spoken language because of intrinsic differences between the two modalities, psycholinguists investigating issues such as syntactic ambiguity resolution and reference resolution using eye movements in reading and eye movements in spoken language believe they are testing theoretical claims about these processes that transcend the modality of the input. Thus, the psycholinguistic community will increasingly be faced with questions about how to integrate results from visual world studies with results from studies of eye movements in reading and sometimes how to reconcile conflicting results.

2.3.1. Processing load versus representational measures

In comparing reading studies to visual world studies it is useful to make a distinction between behavioral measures of language processing that measure processing difficulty and measures that probe representations. The distinction is more of a heuristic than a categorical distinction because many response measures combine aspects of both. Processing load measures assess transient changes in process complexity, and then use these changes to make inferences about the underlying processes and representations. Representational measures examine when during processing a particular type of

epresentations emerges and then use that information to draw inferences about the underlying processes and representations. Neither class of measure nor its accompanying experimental logic is intrinsically preferable to the other; the nature of the question under investigation determines which type of response measure is more appropriate.

The majority of studies that use eye movements to examine reading make use of eye movements as a processing load measure. The primary dependent measure is fixation duration. The linking hypothesis between fixation duration and underlying processes is that reading times increase when processing becomes more difficult. In contrast, the majority of visual world studies use eye movements as a representational measure. The primary dependent measure is when and where people fixate as the utterance unfolds. We can illustrate these differences by comparing reading studies of lexical and syntactic ambiguity resolution with visual world studies that address the same issues.

2.3.2. Lexical ambiguity

In a well-known series of studies, Rayner and colleagues (e.g., Duffy, Morris, & Rayner, 1988) have examined whether multiple senses of homographs, such as *bank, ball,* and *port* are accessed during reading, and if so, what are the effects of prior context and the frequency with each sense is used. Processing difficulty compared to an appropriate control is used to infer how ambiguous words are accessed and processed. For 'balanced' homographs with two more or less equally frequent senses, fixation duration is longer compared to frequency-matched controls–resulting in the inference that the multiple senses are competing with one another. This ambiguity "penalty" is reduced or eliminated for biased homographs when a 'dominant' sense is far more frequent than a 'subordinate' sense and when the context strongly favors either one of two equally frequent senses or the more frequent sense. Note that while these results do not provide clear evidence about time course per se, the overall data pattern allows one to infer that multiple senses are accessed, with the dominant sense accessed more rapidly. One can get crude time-course information by separately analyzing the duration of the initial fixation and using that as a measure of relatively early processes. More detailed information about time course can be obtained by using fixation duration as a measure, but using variations on the fast priming methods, introduced by Sereno and Rayner (1992).

A study using the visual world paradigm would adopt a similar approach to that used by Allopenna et al. Potential referents associated with the alternative senses would be displayed and the time course of looks to these referents would be used to infer degree of activation and how it changes over time. For balanced homophones, one would predict looks to the referents of both senses. For biased homophones, looks to the more frequent would begin earlier than looks to the less frequent sense. This pattern would be similar to those obtained in classic studies using cross-modal priming from the 1970s and early 1980s (Swinney, 1979; Tanenhaus, Leiman, & Seidenberg, 1979; for review see Simpson, 1984; Lucas, 1999). Note that these results would not provide direct information about processing difficulty, though one might infer from them that competing senses would result in an increase in complexity.

Thus, while the eye-movement reading studies do not provide direct information about time course and visual world studies do not provide direct information about processing difficulty, the results from reading studies that use a processing load strategy and visual world studies that probe emerging representations could converge on the same conclusions.

2.3.3. *Syntactic ambiguity*

Beginning with the classic article by Frazier and Rayner (1982), eye tracking in reading has been the response measure of choice for psycholinguists interested in syntactic processing. Frazier and Rayner's approach was to examine the processing of temporarily ambiguous sentences, using reading times within pre-defined regions to infer if and when the reader had initially pursued the incorrect interpretation. For a range of syntactic ambiguities, most of which involved disambiguating the phrase that could be "attached" to a verb phrase, thereby introducing an argument, in favor of a noun phrase attachment that modified the head noun, Frazier and Rayner found an increase in fixation duration and an increase in regressive eye movements from the disambiguating region. For current purposes we will focus on fixation duration because it is most clearly a processing load measure. The question of how to interpret regressions is more complex and beyond the scope of this chapter. The increase in fixation duration was interpreted as evidence that processing had been disrupted, thereby leading to the inference that readers had initially chosen the argument interpretation. Frazier and Rayner also introduced several different measures that divided fixations within a region in different ways. For example, 'firšt pass' reading times include all fixations beginning with the first fixation within a region until a fixation that leaves a region, and are often used as a measure of early processing.

Timing is less straightforward in eye-tracking reading when fixations are divided into multiple word regions. Most of the complexities in inferring time course in reading studies arise because the sequence of fixations need not correspond to the linear order of the words in the text. This is especially the case when one considers that arguments about timing often depend on defining regions of text and then partitioning fixations into categories in ways that separate the measure from when the input is first encountered.

Studies examining syntactic ambiguity resolution with the visual world paradigm use the timing of looks to potential referents to infer, if and, if so, when, a particular analysis is under consideration. For example, in one-referent contexts (an apple on a towel, a towel, a box and a pencil) and instructions such as, *Put the apple on the towel in the box,* Spivey et al. (2002) found that looks to the false goal (the towel without the apple) began to increase several hundred millisecond after the onset of *towel* (see Figure 3). In contrast, in two referent contexts (two apples, one on a towel and one on a napkin) fixations to the apple on the towel begin to increase several hundred millisecond after the onset of *towel.* This pattern of results suggests that the prepositional phrase *on the towel* is initially considered a goal argument in the one-referent context and a noun phrase modifier in the two-referent context. Information about time course is straightforward with the visual world logic because fixations can be aligned with the input, allowing strong inferences about what information in the input was likely to have triggered the fixation. The reason that one can align fixations and the input is, of course, because the input unfolds sequentially. Note, however, that one cannot use fixations in a straightforward way to draw inferences about processing difficulty. Thus the visual world approach is unlikely to become a paradigm of choice for investigating issues about resource demands, including increasingly important questions about what factors contribute to the complexity of sentences (e.g., Grodner & Gibson, 2005; Hale, 2003; Lewis & Vasishth, 2005).

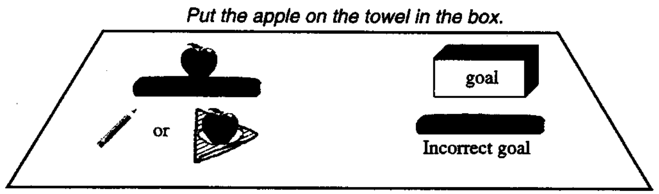

Figure 3. Schematic of the one and two-referent conditions in the Tanenhaus et al. (1995) and Spivey et al. (2002) prepositional phrase-attachment studies.

2.4. Effects of Display

The single factor that most complicates the interpretation of visual world studies of language processing is the need to use a display. First, the encoding of the display can introduce contingencies. For example, the timing of looks to a potential referent at point *t* could be affected by whether or not that referent has been fixated on during time *t-x*, either during preview or as the sentence unfolds. Thus the likelihood of a fixation may be contingent on both the input and the pattern of prior fixations. This, of course, has the potential to complicate inferences about time course, in much the same way that re-reading after a regression can complicate the interpretation of fixation duration data in eye-movement reading studies. Recent studies have begun to examine how having fixated a potential referent during preview affects the likelihood that it will be fixated when it is temporarily consistent with the input (Dahan, Tanenhaus, & Salverda, in press).

Second, use of a display with a small number of pictured referents or objects and a limited set of potential actions creates a more restricted environment than language processing in most natural contexts, while at the same time imposing more demands on the participant than most psycholinguistic tasks. In order to address these closed set issues, we will consider two cases: the first from spoken word recognition; the second from reference resolution.

2.4.1. *Spoken word recognition*

In the Allopenna et al. paradigm, the potential response set on each trial is limited to four pictured items. If participants adopted a task-specific verification strategy, such as implicitly naming the pictures, then the unfolding input might be evaluated against these activated names, effectively bypassing the usual activation process, and leading to distorted results. Even if participants do not adopt such a strategy, the visual world methodology might be limited if the effects of the response alternatives mask effects of non-displayed alternatives (e.g., neighborhood effects in the entire lexicon). This would restrict its usefulness for investigating many issues in spoken word recognition, in particular issues about the effects of lexical neighborhoods, i.e., the set of words in the lexicon that are similar to the target word. Here, an analogy might be helpful. Researchers often use lexical priming paradigms to probe for whether an exemplar of a particular class of lexical competitor is active, for example, cohorts or rhymes. However, these paradigms are not well suited for asking questions about the aggregate effects of the number and frequency of potential competitors. In order to investigate this class of question, researchers have found it more useful to measure response time to a target word, for example, auditory lexical decision, which more closely approximates a processing load measure.

2.4.2. *Implicit naming*

The issue of implicit naming has been addressed most directly by Dahan and Tanenhaus (2005) in a study that varied the amount of preview time, 300 or 1000 ms, for four-picture displays with minimal phonological overlap between the names of the distractors and the target (Figure 4). On a subset of the trials, two of the pictures were visually similar (e.g., a picture of a snake and a coiled rope) and the instruction referred to one of the pictures (e.g., *click on the snake*). The particular pictures chosen as the two referents shared some features associated with a prototypical visual representation of one or both words. For example, the pair *snake–rope* was selected because the picture of a coiled rope shares some features with the visual representation most often associated with the concept of a snake. When selecting pictures, Dahn and Tanenhaus (2005) sought to minimize their visual similarity so that the objects could be easily differentiated. For example, we chose a snake in a non-coiled position. Thus, visual similarity was maximized between the prototypical visual representation of one of the concepts, the referent, and the picture associated with the other concept, the competitor, and minimized between the competitor picture and the picture of the referent concept.

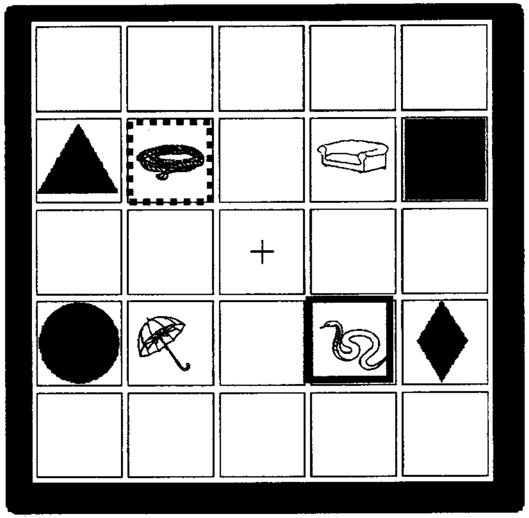

Figure 4. Sample display from Dahan and Tanenhaus (1995), illustrating a display where the visual competitor (the rope) is visually similar to a prototypical snake, whereas the picture of the target referent (snake) is somewhat less prototypical.

Several aspects of the results provide strong evidence against implicit naming. Preview duration did not affect the magnitude of visual similarity effects (looks to visually similar competitors). Moreover, even in the 1000 ms condition, the magnitude of visual similarity effects was not affected by whether or not the competitor was fixated during preview; the naming hypothesis predicts that effects would be eliminated or weakened with preview because the encoded name of the picture would not match the unfolding target. Finally, similarity effects were larger when the target had a competitor that was chosen to share visual features of its prototype representation compared to when that competitor was the referent. Thus visual similarity effects were due to the fit between the picture and the conceptual representation of the picture, not simply surface visual confusability. This suggests that mapping of the word onto its referent picture is mediated by a visual/conceptual match between the activated lexical form of the target and the picture. This hypothesis is further supported by additional analyses of the effects of fixation to a competitor during preview on the likelihood that it will be re-fixated during the speech input and evidence that a spoken word triggers looks to potential referents when the participant is engaged in a visual search task to identify the location of a dot when it appears on a random location within a schematic scene (Salverda & Altmann, 2005).

2.4.3. *Sensitivity to hidden competitors*

Perhaps, the strongest test of the sensitivity of visual world studies comes from studies that look for effects of non-displayed or "hidden competitors." For example, Magnuson, Dixon, Tanenhaus, and Aslin (in press) examined the temporal dynamics of neighborhood effects using two different metrics: neighborhood density, a frequency-weighted measure defined by the Neighborhood Activation Model (NAM), and a frequency-weighted measure of cohort density. The referent was displayed along with three semantically unrelated pictures, with names that had little phonological overlap with the referent (all names were monosyllabic). Crucially, none of the referent's neighbors were either displayed or named throughout the course of the experiment. The results showed clear effects of both cohort and neighborhood density, with cohort density effects dominating early in the recognition process and neighborhood effects emerging relatively late.

These results demonstrate that the processing neighborhood for a word changes dynamically as the word unfolds. It also establishes the sensitivity of the paradigm to the entire lexicon. To a first approximation then, when competitors are displayed, the paradigm can be used to probe specific representations, however, the aggregate effects of competitors can be observed in the timing of fixations to the target referent.

Magnuson et al.'s results complement Dahan et al. (2001b) finding that misleading coarticulatory information delays recognition more when it renders the input temporarily consistent with a (non-displayed) word, compared to when it does not. In addition, simulations using the Allopenna et al. linking hypothesis successfully captured differences between the effects of misleading coarticulatory information with displayed and non-displayed competitors. Whether the non-displayed competitor logic can be extended to higher-level sentence processing remains to be seen.

2.4.4. Sentence processing

Much trickier issues about the effects of the display come into play in higher-level processing. For example, one could argue that in the Tanenhaus et al. (1995) study displaying an apple on a towel and an apple on a napkin increases the salience of a normally less accessible sense compared to circumstances where the alternative referents are introduced linguistically. One could make a similar argument about the effects of action on the rapidity with which object-based affordances influence ambiguity resolution in studies by Chambers and colleagues (Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002; Chambers et al. 2004). In these studies, the issue of implicit naming seems prima facie to be less plausible. However, one might be concerned about task-specific strategies. For example, in Chambers et al. (2002), participants were confused, as indexed by fixations when they were told to, *Pick up the cube. Now put the cube in the can,* and there were two cans. The confusion was reduced or eliminated, however, when the cube would only fit in one of the cans. Because only one action was possible, one might attribute this to problem solving, and not as Chambers et al. argued to the effects of action and affordance on referential domains. However, the manipulation had opposite effects for instructions that used an indefinite article, e.g., *Pick up the cube. Now put it in a can.* Here participants were confused when the cube would only fit in one of the cans. This strategy of pitting linguistic effects against potential problem-solving strategies is crucial for evaluating the impact of strategies due to the display and the task.

Perhaps, the most general caution for researchers using the visual world paradigm in both production and comprehension is to be aware that while the visual world displays entities that can be used to infer the representations that the listener is developing, it also serves as a context for the utterance itself. Note that the fact that information in a display affects processing is not itself any more problematic than the observation that reference resolution, for example, is affected by whether or not potential referents are introduced linguistically in a prior discourse. One sometimes encounters the argument that the visual world paradigm can be informative about language processing only if gaze patterns to a potential referent in a display are not affected by the other characteristics of the display. This argument is no more or less valid than the comparable argument that fixations in reading can only inform us about word recognition or reference resolution if fixations to a word are unaffected by the context in which the fixated word occurs. What is crucial, however, is whether the nature of the interactions with the display shed light on linguistic processing or whether they introduce strategies that mislead or obscure the underlying processes. Thus, far investigations of potential problems has been encouraging for the approach. However, it will be crucial in further work to explore the nature of the interactions between the display and linguistic processing in much greater detail.

3. APPLICATIONS TO ISSUES IN LANGUAGE COMPREHENSION

In this section, we present a brief review of work in three domains where using eye movements is beginning to have a major impact on our understanding of spoken language processing. We begin with issues in speech, spoken word recognition, and prosody that can be addressed by using variations of the procedures we described in presenting the study by Allopenna et al. (1998). The second domain consists of issues in sentence processing, including classic issues about the role of context in syntactic ambiguity resolution, and assorted issues about referential domains. These issues are addressed by taking advantage of various features of the visual world paradigm, including having an implicit measure that can be used with simple tasks and spoken language, having a co-present referential world, and the capability of monitoring real-time processing in paradigms that bridge the language-as-product and language-as-action traditions. We then conclude with a discussion of how eye movements are beginning to provide insights into how real-time language processing develops in infants, toddlers, and young children.

3.1. Spoken Word Recognition and Prosody

3.1.1. Spoken word recognition

As we noted earlier, many classic issues about spoken word recognition can naturally be addressed using variations on the procedure used by Allopenna et al. (1998). These include questions about what types of lexical competitors become activated as a spoken word unfolds (Allopenna et al., 1998; Magnuson, 2002) and how lexical competition is modulated by context (Dahan et al., 2001b; Dahan & Tanenhaus, 2004). The Allopenna et al. (1998) procedure has also proved to be extremely useful for addressing questions about how listeners use sub-phonetic information in word recognition. Examples include the McMurray et al. (2002) study described earlier, which used looks to competitors to demonstrate fine-grained sensitivity to within-category variation and work by Dahan et al. (2001a) and Gow and McMurray (in press) on listener's use of coarticulatory information (Dahan et al., 2001a). And, in an important study, Salverda, Dahan, and McQueen (2003) used eye movements to demonstrate that listeners exploit small systematic differences in vowel duration in processing of words such as *captain,* which begin with a phonetic sequence that is itself a word, e.g., *cap* (the vowel in a monosyllabic word is typically longer than the same vowel in a polysyllabic word). Examining looks to cohort competitors to words embedded in utterances has also proved useful for examining spoken word recognition in bilinguals. For example, Spivey and Marian (1999) used looks to cohorts to demonstrate that bilingual speakers following instructions in one language, briefly consider potential referents with names that are cohort competitor in their second language (see also Ju & Luce, 2002). Finally, studies that use eye movements to measure processing of artificial lexicons and languages, initiated by

Magnuson and his colleagues (Magnuson, Tanenhaus, Aslin, & Dahan, 2003) are proving useful for addressing a range of issues in spoken word recognition and learning.

3.1.2. *Prosody*

Visual world studies are beginning to have an increasingly large impact on research investigating how listeners process information about prosody, which is carried by the pattern and type of pitch accents and realized acoustically as changes in duration, intensity, and pitch excursion on stressed vowels. Differences in vowel duration between mono, and polysyllabic words vary with the prosodic environment; they are smallest in the middle of a phrase, and largest at the end of a phrase. Salverda (2005) demonstrated that prosodic factors modulate the relative degree to which different members of a neighborhood will be activated in different environments; in medial position a polysyllabic carrier word such as *captain* is a stronger competitor than *cat* for the target *cap*, whereas the opposite pattern obtains in utterance-final position.

Cohort manipulations, in particular, are well suited for examining pitch accents because one can examine effects that are localized to the vowel that carries the pitch accent. Dahan, Tanenhaus, and Chambers (2002) examined the timing of looks to targets and cohort competitors for accented and unaccented words that referred to discourse given and discourse-new entities (e.g., *Put the candle above the triangle. Now put the CANDY/candy...*). Dahan et al. found that listeners use information about pitch accent as the vowel unfolds, initially assuming that nouns in definite referring expression with unaccented vowels refer to the most salient entity (the subject/focus) of the previous sentences, whereas words with accented vowels refer to a non-focused given entity if available, or if not, a new entity.

Arnold and colleagues (Arnold, Tanenhaus, Altmann, & Fagnano, 2004) adapted the Dahan et al. cohort design to evaluate the hypothesis that a disfluent production of a noun phrase (thee uh CANDY) would bias listeners to expect reference to a discourse-new entity. With fluent productions, Arnold et al. replicated Dahan et al.'s finding that an accented noun was preferentially interpreted as referring to a non-focused entity. However, with a disfluent production, the preference shifted to the discourse-new entity. Watson and his colleagues (e.g., Watson, Gunlogson, & Tanenhaus, in press) have also used cohort competitors to test hypotheses about the interpretation of different pitch accents, focusing on potential differences between the H* (presentational) and L+H* (contrastive) pitch accents (Pierrehumbert & Hirschberg, 1990).

Ito and Speer (described in Speer & Ito, in press) have also investigated presentational and contrastive accents, combining eye movements with a "targeted language game." The director, a naïve participant, instructs a confederate about how to decorate a Christmas tree using ornaments that need to be placed on the tree in a specified sequence. Ornaments differ in type, e.g., bells, hats, balls, houses, etc. and in color, e.g., orange, silver, gold, blue, etc. Recordings demonstrated that participants typically used a presentational accent (H*) when a color was new to the local discourse. For example, "orange" typically received a presentational accent in the instruction, "First, hang an orange ball on the left" when an orange ornament was being mentioned for the first time) for a particular row. However, if the instruction to place the orange ball followed placement of a ball of a different color, e.g., a silver ball, then "orange" was more likely to be produced with a contrastive accent (L+H*). Ito and Speer showed that the recordings using the preferred pitch accent pattern used by naïve participants facilitated listeners' time to identify the correct ornament, as measured by eye movements.

3.2. Sentence Processing

3.2.1. *Syntactic ambiguity resolution*

In a series of classic papers, Crain (1981), Crain and Steedman (1985), and Altmann and Steedman (1988) argued that many of the systematic preferences that readers and listeners exhibit when resolving temporary syntactic ambiguity are not due to differences in syntactic complexity between the alternative structures, but rather to differences in referential implications. A well-known example comes from prepositional (PP) attachment ambiguities as illustrated in sentences such as *Anne hit the thief with the wart* is one such example. The strong initial preference to consider *with the wart* (erroneously) as the instrument of *hit* rather than as a restrictive modifier of *the thief* could in part be due to the fact that the restrictive modifier is most felicitous in a context in which multiple thieves are present, one of which has a wart. In the absence of such a context, there is little reason for considering the modification analysis. Indeed, some (but not all) eye-movement studies with text have found that this referential factor (i.e., the presence/absence of referential ambiguity) has immediate effects on real-time syntactic ambiguity resolution in reading (e.g., Altmann & Steedman, 1988; Britt, 1994; Sedivy, 2003; Spivey-Knowlton & Sedivy, 1995; Spivey & Tanenhaus, 1998; but for discussion of studies finding weak or delayed effects of referential context see Rayner, this volume, and Rayner & Liversedge, 2004).

Introducing a referential world that is co-present with the unfolding language, naturally highlights these and other questions about reference. Indeed, the initial action-based visual world study (Tanenhaus et al., 1995, described earlier) examined how referential ambiguity (i.e., the presence of multiple apples in a scene) influences the listeners' initial bias when encountering a sentence with a temporarily ambiguous prepositional phrase (*Put the apple on the towel in the box.*). Recall that the presence of two apples in the scene shifted listeners' initial preference to interpret *on the towel* from a goal preference to a modifier preference. This study confirms that something like Crain's Referential Principle is an important factor when listeners interpret spoken language in the context of visually co-present referents.

Subsequent work by Snedeker and Trueswell (2004) confirms the importance of referential context, but importantly establishes that high-level expectations *contribute* to but do not solely *determine* the outcome of ambiguity resolution in visual contexts. A multiple constraint view of sentence processing predicts that lower-level linguistic factors, such as verb argument preferences, contribute simultaneously to the ambiguity resolution process. Snedeker and Trueswell (2004) confirmed this prediction in a study containing sentences that were globally ambiguous in their structure (not just temporarily ambiguous). College-age adults heard sentences like those in (1a) through (1c).

1. a. *Tickle the pig with the fan.*	(Instrument-biased Verb)
b. *Feel the frog with the feather.*	(Equi-biased Verb)
c. *Choose the cow with the stick.*	(Modifier-biased Verb)

Verbs were selected based on a separate sentence completion study, which evaluated how often a *with*-phrase would be used for these verbs as an instrument, allowing verbs to be operationally defined as: Instrument-bias, Modifier-bias, or Equi-bias. As in Tanenhaus et al. (1995), 2- and 1-Referent scenes were compared. Scenes contained, e.g., a Target Theme (a pig holding a small fan); a Competitor Theme (a pig/horse wearing a hat); a Potential Instrument (a large fan); and another object (a large hat). Here, looks to the potential instrument and the ultimate action were analyzed: i.e., participants could pick up the fan and use it to tickle the pig, or they could use their fingers to do the actions. The eye movement and action data revealed simultaneous effects of both the referential context (2-Referent versus 1-Referent) and verb argument preferences; the presence of multiple pigs reduced looks to, and use of, the potential instrument; likewise degree of verb-bias (from Instrument-biased to Modifier-biased) systematically decreased looks and actions involving the Potential Instrument. Crucially, these verb effects were observed in both *1-Referent and 2-Referent* Scenes, suggesting that the mere presence of multiple referents does not solely determine attachment preferences for listeners.

We note that it remains something of a puzzle why the effects of referential context seem so much stronger in studies examining the PP-attachment ambiguities involving goals versus modifiers (*Put the apple on the towel*) compared to Instrument versus modifiers (*Tickle the frog with the feather*) given that *put* is a verb that has a strong goal-bias. For some speculation about possible explanations, see Snedeker and Trueswell (2004); Spivey et al. (2002); Tanenhaus and Trueswell (2005), and Trueswell and Gleitman (2004).

3.3. Circumscribing Referential Domains

The studies reviewed thus far made the simplifying assumption that the referential domain for a linguistic expression comprises all of the salient entities in the environment that are temporarily consistent with the referring expression as it unfolds. However, speakers at least in their own productions consider real-world constraints like the proximity and relevance of potential referents, the relevance of other estimations of the knowledge that the listener has of the world, and several other factors (Clark, 1992; Levelt, 1989; Lyons, 1981; Stone & Webber, 1998). Put more concretely, a speaker's decision to refer to an object as *the ball, the red ball, the ball closer to you, the slightly asymmetric sphere, it, that one* or *that,* clearly depends on this wide range of spatial, perceptual, social, and cognitive factors.

A central theme of research using the visual world paradigm has been to understand how and when these factors impinge on decisions made by listeners and speakers (Chambers et al., 2001; Sedivy, 2003; Sedivy et al., 1999; Grodner & Sedivy, in press; Keysar et al., 2000; Keysar & Barr, 2005; Brown-Schmidt et al., 2005; Brown-Schmidt & Tanenhaus, in press). For instance, we have already discussed some studies demonstrating that listeners dynamically update referential domains, integrating information from the unfolding utterance in conjunction with the entities in the workspace (Chambers et al., 2002, 2004; Eberhard et al., 1995) and generating expectations about upcoming referents (Altmann & Kamide, 1999; Kamide et al., 2003), especially those that are likely to be realized as arguments (Boland, 2005). And, in an ingenious series of eye movement studies, Altmann and colleagues have recently demonstrated that actions described or implied in a narrative influence expectations about how the location of objects will change in the listener's mental model of the scene, as determined by looks to locations in the scene (Altmann & Kamide, 2004).

A listener's referential domain is also affected by intended actions and the affordances of potential objects that are relevant to those actions (Chambers et al., 2002). These affordances also affect the earliest moments of syntactic ambiguity resolution, challenging the claim that language processing includes a syntactic subsystem (module) that is informationally encapsulated, and thus isolated from high-level non-linguistic expectations (Coltheart, 1999; Fodor, 1983). For example, Chambers, Tanenhaus, and Magnuson (2004) showed that in a two-referent context that includes a liquid egg in a bowl and a liquid egg in a glass, participants will initially treat the PP *in the bowl* as a modifier with an instruction such as *pour the egg in the bowl over the flour.* However, when the egg in the bowl is solid and thus cannot be poured, then participants initially misinterpret *in the bowl* as the Goal. These results cannot be attributed to constraints lexically encoded within the linguistic representation of the verb *pour*; Chambers et al. found the same pattern of results with the verb *put* when the affordances were introduced non-linguistically by handing the participant an instrument.

3.3.1. *Scalar implicatures*

Earlier we reviewed Sedivy's finding that listeners assume that the referential domain includes a contrast set when they hear a pre-nominal scalar adjective, such as *tall.* These results are particularly striking because they represent one case in which listeners immediately generate a pragmatic inference based on a generalized implicature. There is an emerging debate about when listeners generate these types of inferences, whether they apply differently to different classes of scales, especially those that involve potential contrasts between a so-called logical interpretation (e.g., logical or inclusive *OR* versus pragmatic or exclusive *OR*) where there are claims that logical *OR* is computed (obligatorily) prior to pragmatic *OR*, and how these inferences are modulated by context (see Noveck & Sperber, 2005). Visual world eye-movement studies are beginning to feature prominently in research in this arena, though this work had not yet begun to appear in the literature as we were preparing this chapter.

Eye-movement research using pre-nominal adjectives is beginning to shed light on inference under other circumstances. Although any adjective can appear post-nominally, either in a restrictive relative clause (*the glass **that is tall***) or in a prepositional phrase (*The glass **with spots***), some adjectives are typically used pre-nominally (e.g., scalar adjectives and color adjectives), others are nearly always used post-nominally (*the shape **with diamonds***), and others occur equally often in pre-nominal and post-nominal positions (e.g., *striped, with stripes*). Using a point of disambiguation logic, Edwards and Chambers (2004) have shown that listeners make rapid use of the **absence** of a pre-nominal modifier to rule out candidate referents. Second, Grodner and Sedivy (in press) have established that listeners rapidly adjust to how reliably a speaker uses scalar adjectives contrastively, including making adjustments based on meta-linguistic information provided by an experimenter. Arnold, J. E. (personal communication) reports similar results with meta-linguistic information provided about a disfluent speaker. These results bear on questions about when in the time course of processing, and under what circumstances speakers and listeners consider the likely knowledge and intentions of their interlocutors, a topic we will return to shortly.

3.4. Word-Order Variation, Discourse, and Information Structure

The visual world paradigm has also proved to be a useful tool investigating how discourse-pragmatic factors related to information structure influence reference resolution and parsing. One such area has been an exploration of how sentence processing is achieved in languages that have highly flexible word orders (Kaiser & Trueswell, 2004; Järvikivi, van Gompel, Hyönä, & Bertram, 2005). The reason for this interest is that flexible word-order languages of this sort typically use order to communicate the information structure and discourse status (given/new distinctions). Kaiser and Trueswell (2004) used the visual world paradigm to explore how reference resolution in Finnish, a flexible word-order language with canonical SVO order and no articles. The non-canonical order OVS marks the object as given and the subject as new; SVO is more flexible, being used in multiple contexts. In the study, the eye gaze of Finnish listeners was tracked as they heard spoken descriptions of simple pictures, so as to test whether listeners use this knowledge of information structure to their advantage, to increase the efficiency with which visual information is collected. That is, upon hearing an OV… sequence, Finnish listeners should expect the upcoming noun to be discourse-new, whereas an SV… sequence makes no such prediction. The results confirmed these predictions. As compared to SVO, OVS sentences caused listeners to launch anticipatory eye movements to a discourse-new referent at the second noun onset, even before participants had enough acoustic information to recognize this word. The findings illustrate that in a flexible word-order language, a non-canonical order can result in anticipatory processes regarding the discourse status of a yet-to-be-heard constituent.

3.5. Pronouns and other Referring Expressions

Relatedly, numerous researchers have begun to use the visual world paradigm to study how syntax and information structure interact with the type of referring form (full noun phrases, pronouns, etc.) (Arnold et al., 2000; Järvikivi, van Gompel, Hyönä, & Bertram, 2005; Brown-Schmidt et al., 2005; Runner et al., 2003, in press). The visual paradigm is particularly useful for addressing these questions because the looks to potential referents, especially, when combined with a decision, allow for strong inferences about which potential referents are being considered and which referent is selected.

Several studies have examined how the order in which characters in a scene are mentioned influence the interpretation of utterances with both ambiguous and unambiguous pronouns. Arnold et al. (2000) found that English listeners upon hearing a sentence beginning with an ambiguous pronoun (he) preferentially looked to the character that had been mentioned first in the previous sentence. Kaiser and Trueswell (in press) show that this preference, at least in Finnish, reflects a preference for pronouns to refer to the grammatical subject of the previous sentence, not the object (but see also Järvikivi, van Gompel, Hyönä, & Bertram, 2005). Preferences depend though on the type of pronoun used in Finnish, another class of pronouns (demonstratives) preferentially selects referents based on surface word order rather than grammatical role. Brown-Schmidt and colleagues (Brown-Schmidt, Byron, & Tanenhaus, 2005) used eye movements and actions to demonstrate differences in the interpretation of *it* and *that,* following an instruction such as *Put the cup on the saucer. Now put it/that…..* Addressees preferentially interpret *it* as referring to the theme (the cup), whereas *that* is preferentially interpreted as referring to the composite created by the action (the cup on the saucer), which does not have a linguistic antecedent (*the cup on the saucer* is not a constituent in the instruction). Finally, as we mentioned earlier, the visual world paradigm is being used to examine the interplay between structural constraints (e.g., binding constraints), discourse, and type of referring expressions for pronouns and reflexives (Runner et al., 2003, in press).

3.6. Common Ground, Alignment, and Dialogue

Until recently, most psycholinguistic research on spoken language comprehension could be divided into one of two traditions, each with its own theoretical concerns and dominant methodologies (Clark, 1992; Trueswell & Tanenhaus, 2005). The product tradition emphasized the individual cognitive processes by which listeners recover linguistic representations, typically by examining moment-by-moment processes in real-time language processing, using carefully controlled stimuli scripted materials and fine-grained on-line measures.

In contrast, the action tradition focused on how people use language to perform acts in conversation–the most basic form of language use. Many of the characteristic features of conversation emerge only when interlocutors have joint goals and when they participate in a dialogue both as a speaker and an addressee. Thus, research within the action tradition typically examines unscripted interactive conversation involving two or more participants engaged in a cooperative task, typically with real-world referents and well-defined behavioral goals–conditions that are necessary for many of the characteristic features of conversation to emerge.

Recently, the language-processing community has begun to show increased interest in bridging the product and action traditions (Pickering & Garrod, 2004; Trueswell & Tanenhaus, 2005). However, research that aims to bridge the two traditions has rarely combined on-line measures–the methodological cornerstone of the product tradition, with unscripted cooperative conversation–the central domain of inquiry in the action tradition (see Brennan, 1990, 2005 for a notable exception). The reason is that most on-line measures interfere with dialogue. In contrast, eye movements can be monitored in most of the tasks used by researchers in the action tradition.

We believe that research monitoring eye movements in unscripted conversation is likely to play a central role in addressing at least two fundamental questions that are becoming the focus of much current research. The first is at what temporal grain do interlocutors monitor each other's likely knowledge and intentions. The second is to what degree, and at what temporal grain, do the representations of interlocutors become **aligned** during interactive conversation (Pickering & Garrod, 2004).

With respect to common ground, although keeping track of what is known, and not known, to the individual participants in a discourse would seem to be fundamental for coordinating information flow (Brennan & Hulteen, 1995; Clark, 1992, 1996), computing common ground by building, maintaining, and updating a model of a conversational partner's beliefs could be memory intensive. (Thus interlocutors may not consider common ground during initial processing; Keysar & Barr, 2005.) Some supporting evidence comes from eye-movement studies showing that addressees often fail to reliably distinguish their own knowledge from that of their interlocutor when interpreting a partner's spoken instructions (Keysar et al., 2000; Keysar, Lin, & Barr, 2003; but cf. Nadig & Sedivy, 2002; Hanna et al., 2003). However, these studies use confederates, which restricts and changes the nature of the interaction (Metzing & Brennan, 2003), the degree to which common goals are negotiated, and perhaps most importantly the types of the constructions that are used in the conversation, each of which can mask effects of perspective taking (for discussion and supporting evidence, see Tanenhaus & Brown-Schmidt, in press).

With respect to alignment, Pickering and Garrod (2004) propose that successful dialogue requires interlocutors to arrive at similar (aligned) representations across multiple linguistic and conceptual levels. They further propose that priming provides a mechanism by which alignment occurs, noting, for example, that syntactic persistence, the tendency for speakers to choose a structure they have previously heard or produced, appears to be particularly robust in dialogue (Branigan, Pickering, & Cleland, 2000). However, even if Pickering and Garrod are correct in identifying priming as an important mechanism for alignment priming will have to be supplemented by real-time measures that probe the representations of interlocutors. Otherwise, priming is being called upon to serve both as a proposed mechanism, and as a diagnostic for alignment, raising concerns about circularity.

Recent research by Brown-Schmidt and colleagues demonstrates that it is possible to use eye movements to monitor real-time processes in task-oriented dialogues with complex tasks and naïve participants (Brown-Schmidt, Campana, & Tanenhaus, 2005; Brown-Schmidt, 2005; Brown-Schmidt & Tanenhaus, 2005) For example, Brown-Schmidt et al. used a referential communication task in which participants separated by a barrier cooperated to replace stickers with blocks to match the placement of the blocks in their respective boards (see Figure 5).

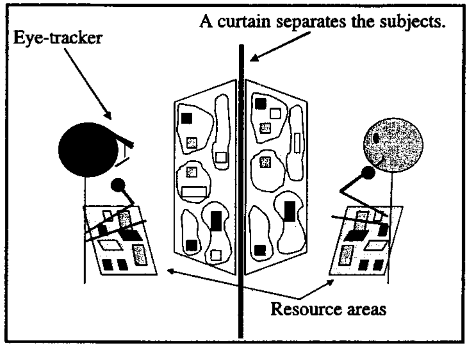

Figure 5. Schematic of the referential communication task used in Brown-Schmidt et al.'s (2005) targeted language games study with naïve participants and unscripted dialogue.

They adopted what they termed a "targeted language game" approach, placing stickers to maximize the likelihood that conditions approximating those that might be incorporated in a standard factorial design would emerge. Despite the complexity of the dialogue, they were able to see point-of-disambiguation effects for referring expressions that mirror effects observed in studies with scripted instructions and simple displays. In particular, as a speaker's referring expression unfolded, the addressee's fixations to the referent increased, and fixations to potential competitors decreased, about 200 ms after the place in the speech stream where the input first disambiguated the target from the temporarily consistent competitors.

Additional results strongly demonstrated that the addressee's referential domains were closely aligned. For example, when proximal competitors that did not match the immediate task goals were not part of the speaker's referential domain (as inferred by the form of the referring expression), they were also not considered as potential referents by the addressee (as inferred from fixations).

3.7. Development of Comprehension Abilities

Eye gaze during listening in studies with infants, toddlers, and young children is proving to be a powerful tool for addressing developmental issues in language processing (e.g., Arnold, Brown-Schmidt, Trueswell, & Fagnano, M., 2005; Song & Fisher, 2005; Fernald, Pinto, Swingley, Weinberg, McRoberts, 1998; Swingley, Pinto, & Fernald, 1999; Swingley & Aslin, 2002; Snedeker & Trueswell, 2004; Trueswell et al., 1999). (For a review, see Trueswell & Gleitman, 2004.) In these studies, the time course of children's eye movements is established either by inspecting a videotape of the child's face frame by frame (Swingley et al., 1999), or by analyzing the output of a lightweight eye-tracking visor worn by the child (Trueswell et al., 1999). These eye-movement techniques have the potential to revolutionize how we examine the child's emerging understanding of language, because they provide a natural measure of how linguistic knowledge is accessed and used in real-time interpretation.

Many initial studies demonstrate that, like adults, children rapidly access and use their linguistic knowledge in real-time processing, so long as they know the relevant words and structures. For example, Fernald, Swingley, and colleagues have shown that reference to an object with a known name (e.g., ball) results in shifts in direction of gaze to that object within 600–700 ms of the name's onset, even in children as young as 24 months (Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998). More recent research has explored the extent to which there is continuity in lexical processing over the course of development. For instance, the parallel consideration of lexical candidates appears to be a fundamental property of the spoken language comprehension system even at its earliest stages of development. Swingley et al. (1999) provided 24-month olds with spoken instructions to look at a particular object (e.g., *Look at the tree*) in the presence of either lexical cohort competitor (pictures of a tree and a truck) or some other object (pictures of a tree and a dog). Like Allopenna et al.'s (1998) adult subjects, toddlers showed temporary consideration of both the target and the cohort competitor early in the perception of the word, which resolved toward the target soon after the word's offset (also see Swingley & Aslin, 2002). Consideration of the alternative object did not occur when its name and the target name were not cohorts. These results demonstrate that the developing word-recognition system makes use of fine-grained phonemic contrasts, and from the start is designed to interface this linguistic knowledge (how the word sounds, what the word means) with knowledge about how the word might plausibly behave referentially when making contact with the ambient world.

Other work has begun to examine the development of sentence parsing abilities using eye gaze measures. This research began with studies conducted with five- and eight-year-olds, first reported in Trueswell et al. (1999) that were modeled after the adult "apple-on-the-towel" studies described earlier (Tanenhaus et al., 1995; Spivey et al., 2002). Here children's eye movements were recorded using a lightweight visor system as they acted upon spoken instructions that contained temporary ambiguities such as *Put the frog on the napkin in the box.*

The striking finding was that five-year olds showed a strong preference to interpret *on the napkin* as the Goal of *put*, even when the referential scene supported a Modifier interpretation. Upon hearing *on the napkin*, five-year olds typically looked over to a potential Goal in the scene, the empty napkin, regardless of whether there were two frogs present (supporting a modifier interpretation) or one frog present (supporting a Goal interpretation). The timing of these eyemovements were similar to those observed in the 1-Referent condition of adults, i.e., ~600 ms after the onset of the word "napkin," but for children this pattern of Goal-looks also arose in 2-Referent contexts. In fact, five-year olds' preference for the Goal interpretation was so strong that they showed little sign of revising it; upon hearing *napkin* children looked to the empty napkin as a potential goal, and then frequently moved a frog to that location. In two referent cases, children were equally likely to move the frog that was on the napkin and the frog that was not on the napkin, suggesting they never considered a Modifier interpretation. Importantly, this child-parsing behavior was localized to the ambiguity, and not to the complexity of the sentence. Five-year olds' eye movements and actions became adult-like when the temporary ambiguity was removed, as in the unambiguous modifier form, *Put the frog that's on the napkin in the box.* The nearly perfect performance with unambiguous sentences rules out a potentially mundane explanation of the results, namely that long "complicated" sentences confuse young children. Here an even longer sentence with the same intended structure does not cause difficulty, precisely because the sentence lacks the temporary ambiguity.

Both the Swingley et al. (1999) and Trueswell et al. (1999) results demonstrate that there is considerable continuity in the language-processing system throughout development: lexical and sentential interpretation proceeds incrementally and is designed to coordinate multiple information sources (e.g., linking what is heard to what is seen within milliseconds). However, the differences between five- and eight-year-old children reported by Trueswell et al. (1999) suggest that significant developmental differences do exist. These differences likely pertain to how children learn about sources of evidence pertinent to linguistic and correlated non-linguistic constraints. Highly reliable cues to structure, such as the argument-taking preferences of verbs, are learned earlier than other sources of evidence that may be less reliable or more difficult to discover, because they involve more subtle contingencies. Consistent with this hypothesis, Snedeker and Trueswell (2004) report that young children are more sensitive to verb bias manipulations than to the number of potential referents in the display. Interestingly, children of the same age do appear to be sensitive to referential constraints under some conditions, especially when the discourse guides the child toward the correct referential contrast (see Trueswell & Geltiman, 2004; Trueswell, Papafragou & Choi, in press, for further discussion). Moreover, children are also sensitive to at least some aspects of speaker perspective. Nadig and Sedivy (2003) demonstrated that 5-year-old children distinguish between common ground and privileged ground in a simplified version of the task used by Keysar et al. (2000).

4. CLOSING REMARKS

This chapter has provided an overview to the rapidly growing literature on eye movements and spoken language processing, focusing on applications to spoken language comprehension. We have reviewed some of the foundational studies, discussed issues of data analysis and interpretation, and discussed issues that arise in comparing eye-movement reading studies to visual world studies. We have also reviewed some of the major lines of research that are utilizing this method, focusing on topics in language comprehension, including spoken word recognition, use of referential constraints in parsing, interactive conversation, and the development of language processing abilities in children.

It should be clear from this review that the visual world paradigm is being employed in most traditional areas of inquiry within psycholinguistics. And in each of these areas, the visual world approach is encouraging psycholinguists to investigate uncharted theoretical and empirical issues. Within the study of spoken sentence comprehension, issues about reference have taken center stage, in part because the visual world methodology makes it possible to connect research on real-time reference resolution with social and cognitive research on pragmatics and conversation. Within the study of spoken word recognition, the time-locked nature of this measure has allowed researchers to explore phonemic and sub-phonemic and prosodic contributions of word recognition in utterances at a level of detail previously not possible with traditional methods. It is for these reasons and other reasons we are quite optimistic that eye-movement measures will continue rise in interest and use within the psycholinguistics community.

We close by noting that eye-movement measures are likely to be most powerful when combined with other measures. We have seen how combining eye movements with action and structure tasks can shed new light on real-time language processes. We expect that other measures will emerge that provide additional advantages. For instance, other body movements pertaining to gestures and actions are likely to be highly informative when connected to the timing of speech and eye gaze events. Most generally, we see the visual world approach as part of a larger movement toward connecting language and action in rich goal-directed tasks using increasingly rich and complex data arrays to understand the dynamics of comprehension and production in conversation. This approach is likely to have an increasingly important influence on theoretical development in natural language, just as it as it has begun to enrich theories in other areas of perception and cognition (Ballard, Hayhoe, Pook, & Rao, 1997; Barsalou, 1999; Hayhoe & Ballard, 2005; Land, 2004).

ACKNOWLEDGMENTS

This work was supported by NIH grants DC005071 and HD 27206 to MKT and HD 37507 to JCT.

REFERENCES

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition: 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.

Altmann, G. T. M., & Kamide, Y. (2004). Now you see it, now you don't. Mediating the mapping between language and the visual world. In: J. M. Henderson, & F. Ferreira (Eds), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 279–318). New York: Psychology Press.

Altmann, G. T. M., & Steedman, M. J. (1988). Interaction with context during human sentence processing. *Cognition*, *30*, 191–238.

Arnold, J. E., Brown-Schmidt, S., Trueswell, J., & Fagnano, M. (2005). Children's use of gender and order of mention during pronoun comprehension. In: Trueswell, J. C., & Tanenhaus, M. K. (Eds), *Processing world-situated language: Bridging the language-as-product and language-as-action traditions*. Cambridge, MA: MIT Press.

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., & Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course for pronoun resolution from eyetracking. *Cognition*, *76*(1), B13–B26.

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science*, *15*(9), 578–582.

Bailey, K. G. D. & Ferreira, F. (2005). Don't swim, hop: The timecourse of disfluency processing. Paper to be presented at the 18th annual meeting of the CUNY Conference on Human Sentence Processing, Tucson, AZ.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(4), 723–767.

Barsalou, L. (1999). Language comprehension: Archival memory or preparation for situated action? *Discourse Processes*, *28*, 61–80.

Bock, K., Irwin, D. E., & Davidson, D. J. (2004). Putting first things first. In: J. M. Henderson, & F. Ferreira (Eds), *The interface of language, vision, and action: Eye movements and the visual world.* New York: Psychology Press.

Bock, J. K., Irwin, D. E., Davidson, D. J., & Levelt, W. J. M. (2003). Minding the clock. *Journal of Memory and Language*, *48*, 653–685.

Boland, J. E. (2005). Visual arguments. *Cognition*, *95*(3), 237–274.

Branigan, H. P., Pickering, M. J., & Cleland, A. A., (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(2), B13–B25.

Brennan, S. E. (1990). *Seeking and providing evidence for mutual understanding.* Unpublished doctoral dissertation. Stanford University.

Brennan, S. E. (2005). How conversation is shaped by visual and spoken evidence. In: J. C. Trueswell, & M. K. Tanenhaus (Eds), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions.* Cambridge, MA: MIT Press.

Brennan, S. E., & Hulteen, E. (1995). Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems*, *8*, 143–151.

Britt, M. A. (1994). The interaction of referential ambiguity and argument structure in the parsing of prepositional phrases. *Journal of Memory and Language*, *33*, 251–283.

Brown-Schmidt, S. (2005). Language processing in conversation. Unpublished Doctoral dissertation, University of Rochester.

Brown-Schmidt, S., Byron, D. K., & Tanenhaus, M. K. (2005). Beyond salience: Interpretation of personal and demonstrative pronouns. *Journal of Memory and Language*, *53*(2), 292–313.

Brown-Schmidt, S., Campana, E., & Tanenhaus, M. K. (2005). Real-time reference resolution in a referential communication task. In: J. C. Trueswell, & M. K. Tanenhaus (Eds), *Processing world-situated language*: *Bridging the language-as-action and language-as-product traditions*. Cambridge, MA: MIT Press.

Brown-Schmidt S., & Tanenhaus, M. K. (2005). Real-time interpretation of referential expressions in unscripted interactive conversations, submitted for publication.

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*, 592–609.

Clark, H. H. (1992). *Arenas of language use*. Chicago: University of Chicago Press.

Clark, H. H. (1996). *Using Language*. Cambridge, UK: Cambridge University Press.

Clark, H. H. & Marshall, C. R. (1981). Definite reference and mutual knowledge. In: A. K. Joshi, B. Webber, & I. Sag (Eds), *Elements of discourse understanding* (pp. 10–63). Cambridge, UK: Cambridge University Press.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, *47*(1), 30–49.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Action-based affordances and syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *30*, 687–696.

Coltheart, M. (1999). Modularity and cognition. *Trends in Cognitive Sciences*, *3*, 115–120.

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

Crain, S. (1981). *Contextual constraints on sentence Comprehension*. Ph.D. Dissertation, University of California, Irvine.

Crain, S., & Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological parser. In: D. Dowty, L. Karttunen, & A. Zwicky (Eds), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 320–358). Cambridge, UK: Cambridge University Press.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001a). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*(4), 317–367.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001b). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*(5–6), 507–534.

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Evidence from immediate effects of verb-based constraints. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *30*, 498–513.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review*, *12*(3), 453–459.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, *47*(2), 292–314.

Dahan, D., Tanenhaus, M. K., & Salverda, A. P. (in press). The influence of visual processing on phonetically driven saccades in the "visual world" paradigm. In: R. P. G. van Gompel, Fischer, M. H., Murray, W. S., & Hill, R. L. (Eds), *Eye movements: A window on mind and brain*. Oxford: Elsevier.

Duffy, S. A., Morris, R. K, & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, *27*(4), 429–446.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye-movements as a window into spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*, 409–436.

Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, *9*(3), 228–231.

Ferreira, F., & Bailey, K. G. B. (in press). The processing of filled pause disfluencies in the visual world. In: R. P. G. van Gompel, Fischer, M. H., Murray, W. S., & Hill, R. L. (Eds), *Eye movements: A window on mind and brain*. Oxford: Elsevier.

Fodor, J. A. (1983). *Modularity of mind*. Cambridge, MA: Bradford Books.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*(2), 178–210.

Gow, D. W., & McMurray, B. (in press). Word recognition and phonology: The case of English coronal place assimilation. *Papers in Laboratory Phonology*, *9*.

Griffin, Z. M. (2004). The eyes are right when the mouth is wrong. *Psychological Science*, *15*, 814–821.

Griffin, Z. M., & Bock, J. K. (2000). What they eyes say about speaking. *Psychological Science*, *11*, 274–279.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*, 261–290.

Grodner, D., & Sedivy, J. (in press). The effect of speaker-specific information on pragmatic inferences. In: N. Pearlmutter, & E. Gibson (Eds), *The processing and acquisition of reference.* Cambridge, MA: MIT Press.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research, 33,* 101–123.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language, 49,* 43–61.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9,* 188–194.

Henderson, J. M., & Ferreira F. (2004). *The interface of language, vision, and action: Eye movements and the visual world.* Edited volume. New York: Psychology Press.

Järvikivi, J., van Gompel, R. P. G., Hyönä, J., & Bertram, R. (2005). Ambiguous pronoun resolution: Contrasting the first-mention and subject-preference accounts. *Psychological Science, 16,* 260–264.

Ju, M., & Luce, P. A. (2004). Falling on sensitive ears: Constraints on bilingual lexical activation. *Psychological Science, 15,* 314–318.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 329–354.

Kaiser, E., & Trueswell, J. C. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition, 94,* 113–147.

Kaiser, E., & Trueswell J. C. (in press). The referential properties of Dutch pronouns and demonstratives: Is salience enough? In: M. Weisgerber (Ed.) *Proceedings of Sinn und Bedeutung 8.* University of Konstanz linguistics working papers.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in Incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49,* 133–156.

Keysar, B. & Barr, D. (2005). Coordination of action and belief in conversation. In: J. C. Trueswell, & M. K. Tanenhaus (Eds), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions.* Cambridge, MA: MIT Press.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11,* 32–38.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89,* 25–41.

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

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in the depicted events. *Cognition, 95*(1), 95–127.

Kowler, E. (1995). Eye movements. In: S. M. Kosslyn, & D. N. Osherson (Eds), *Visual cognition: An invitation to cognitive science* (2nd ed., Vol. 2). Cambridge, MA: MIT Press.

Land, M. (2004). Eye movements in daily life. In: L. Chalupa, & J. Werner (Eds), *The visual neurosciences: Vol. 2* (pp. 1357–1368). Cambridge, MA: MIT Press.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Lewis, R. L, & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29,* 375–419.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences, 4,* 6–14.

Lucas, M. (1999). Context effects in lexical access: A meta-analysis. *Memory & Cognition, 27,* 385–398.

Luce, D. R. (1959). *Individual choice behavior.* Oxford: Wiley.

Lyons, J. (1981). Language, meaning and context. London: Collins/Fontana.

Magnuson, J. S. (2002). *The microstructure of spoken word recognition.* Unpublished dissertation, University of Rochester.

Magnuson, J. S. (2005). Moving hand reveals dynamics of thought. *Proceedings of the National Academy of Sciences, 102,* 9995–9996.

Magnuson, J. S., Dixon, J., Tanenhaus, M. K., & Aslin, R. N. (in press). Which words compete? The dynamics of similarity during spoken word recognition. *Cognitive Science.*

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). Time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General, 132,* 202–227.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*(1–2), 77–102.

Marslen-Wilson, W. (1990). Activation, competition, and frequency in lexical access. In: G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives.* Cambridge, MA: The MIT Press.

Marslen-Wilson, W. (1993). Issues of process and representation in lexical access. In: G. T. M. Altmann, & R. Shillcock (Eds), *Cognitive models of speech processing: The second Sperlonga meeting.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics*, *53*, 372–380.

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

McConkie, G. W., & Rayner, K. (1976). Asymmetry of the perceptual span in reading. *Bulletin of the Psychonomic Society*, *8*, 365–368.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201–213.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, *66*(2), B25–B33.

Noveck, J., & Sperber, D. (Eds). (2005). *Experimental pragmatics*. Oxford: Oxford University Press.

Pickering, M. J., & Garrod, S. C. (2004). Towards a mechanistic theory of dialog. *Behavioral and Brain Sciences*, *7*(2), 169–190.

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In: P. R. Cohen, J. L. Morgan, & M. E. Pollack (Eds), *Intentions in communication* (pp. 271–311). Cambridge, MA : MIT Press.

Rayner, K. (1998). Eye movements in reading and information processing: Twenty years of research. *Psychological Bulletin*, *124*, 372–422.

Rayner, K., & Liversedge, S. P. (2004). Visual and linguistic processing during eye fixations in reading. In: J. M. Henderson, & F. Ferreira (Eds), *The interface of language, vision, and action: Eye movements and the visual world.* New York: Psychology Press.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*, 217–236.

Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2003). Assignment of reference to reflexives and pronouns in picture noun phrases: Evidence from eye movements. *Cognition*, *89*, B1–B13.

Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2006). Assigning referents to reflexives and pronouns in picture noun phrases. Experimental tests of binding theory. *Cognitive Science*, *30*, 1–49.

Salverda, A. P. (2005). Prosodically-conditioned detail in the recognition of spoken words. Unpublished Ph.D. dissertation, Max Planck Institute for Psycholinguistics, Nijmegen.

Salverda, A. P., & Altmann, G. (2005). Cross-talk between language and vision: Interference of visually-cued eye movements by spoken language. Poster presented at the AMLaP Conference, Ghent, Belgium.

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

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*, 3–23.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

Sereno, S. C., & Rayner, K. (1992). Fast priming during eye fixations in reading. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 173–184.

Simpson, G. B. (1984). Lexical ambiguity and its role in models of word recognition. *Psychological Bulletin*, *96*(2), 316–340.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, *49*(3), 238–299.

Snodgrass, J. G., & Yuditsky, T. (1996). Naming times for the Snodgrass and Vanderwart pictures. *Behavior Research Methods Instruments & Computers*, *28*(4), 516–536.

Song, H., & Fisher, C. (2005). Who's "she"? Discourse prominence influences preschoolers' comprehension of pronouns. *Journal of Memory and Language*, *52*, 29–57.

Speer, S. R., & Ito K. (in press). Using interactive tasks to elicit natural dialogue production. In: I. Mleinek (Ed.), *Methods in empirical prosody research.* Berlin: Mouton de Gruyter.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*, 10393–10398.

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. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1521–1543.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M. & Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, *45*, 447–481.

Spivey-Knowlton, M. J. (1996). *Integration of visual and linguistic information: Human data and model simulations*. Ph.D. dissertation, University of Rochester.

Spivey-Knowlton, M., & Sedivy, J. C. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition*, *55*(3), 227–267.

Stone, M., & Webber, B. (1998). Textual economy through close coupling of syntax and semantics. In: *Proceedings of INLG*, (pp. 178–187) Niagara-on-the-Lake, Ontario, Canada.

Sussman, R. S. (2006). *Processing and representation of verbs: Insights from instruments.* Unpublished dissertation, University of Rochester.

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science, 13*(5), 480–484.

Swingley, D., Pinto, J. P., & Fernald, A. (1999). Continuous processing in word recognition at 24 months. *Cognition, 71*(2), 73–108.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning & Verbal Behavior, 18*(6), 645–659.

Tanenhaus, M. K., & Brown-Schmidt, S. (in press). Language processing in the natural world. In: B. C. M. Moore, L. K. Tyler, W. D. Marslen-Wilson (Eds), The perception of speech: from sound to meaning. *Philosophical Transactions of the Royal Society B: Biological Sciences.*

Tanenhaus, M. K., Leiman, J. M., & Seidenberg, M. S. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning & Verbal Behavior, 18*(4), 427–440.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 286*, 1632–1634.

Trueswell, J. C., & Gleitman, L. (2004). Children's eye movements during listening: Developmental evidence for a constraint-based theory of sentence processing. In: J. M. Henderson, & F. Ferreira (Eds), *The interface of language, vision, and action: Eye movements and the visual world.* New York: Psychology Press.

Trueswell, J. C., Papafragou, A., & Choi, Y. (in press). Syntactic and referential processes: What develops? In: E. Gibson, & N. Perlmutter (Eds), *The processing and acquisition of reference.* Cambridge, MA: MIT Press.

Trueswell, J. C., Sekerina, I., Hill, N., & Logrip, M. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition, 73*, 89–134.

Trueswell, J. C., & Tanenhaus, M. K. (Eds), (2005). *Processing world-situated language: Bridging the language-as-action and language-as-product traditions.* Edited volume. Cambridge, MA: MIT Press.

Viviani, P. (1990). Eye movements in visual search: Cognitive, perceptual and motor control aspects. In: E. Kowler (Ed.), *Eye movements and their role in visual and cognitive processes: Vol 4. Reviews of oculomotor research* (pp. 253–393). Amsterdam: Elsevier.

Watson, D., Gunlogson, C., Tanenhaus, M. K. (in press). Online methods for the investigation of prosody. In: I. Mleinek, (Ed.), *Methods in empirical prosody research.* Berlin: Mouton de Gruyter.

Yee, E., Blumstein, S., & Sedivy, J. C. (2000). The time course of lexical activation in Broca's aphasia: Evidence from eye movements. Poster presented at the *13th annual CUNY conference on human sentence processing*, La Jolla, CA.