1 Eye Movements as a Tool for Bridging the Language-as-Product and Language-as-Action Traditions

Michael K. Tanenhaus and John C. Trueswell

Introduction: The Product and Action Approaches

The language-as-action and language-as-product traditions, as sketched in the preface to this book, have each had their own characteristic theoretical concerns and preferred experimental methods. For the most part, the product tradition has sought to understand the individual cognitive processes by which listeners recover linguistic representations, whereas the action tradition has sought to understand how people use language to perform joint acts in interactive conversation.

Psycholinguistic research within the product tradition has typically examined moment-by-moment processes in language processing, using fine-grained reaction-time measures designed to tap processes that occur during the perception of a word or of a sentence (Tanenhaus and Trueswell 1995). The rationale for using real-time measures comes largely from the sequential nature of language comprehension. For instance, when comprehending text, readers are known to make successive fixations on individual words rather than taking in entire phrases, with attention focused on the word that is being fixated and the next word to be fixated (Rayner 1998). Fixation patterns from these studies are consistent with the hypothesis that readers assign provisional interpretations to the input essentially on a word-by-word basis (e.g., Rayner 1998; Tanenhaus and Trueswell 1995). Comprehension of spoken language necessarily involves sequential input because speech unfolds as a sequence of rapidly changing acoustic events. As in reading, experimental studies that probe the listeners' developing representations show that they make provisional commitments as soon as the input arrives (Marlsen-Wilson 1973, 1975). In both reading and listening, then, language processing is closely time-locked to the input, which is processed more or less sequentially.

The combination of sequential input and time-locked processing means that the processing system is continuously faced with temporary ambiguity. For example, the initial portion of the spoken word *beaker* is temporarily consistent with many potential lexical candidates, including *beaker*, *beetle*, *beeper*, *beagle*, and so on. An understanding of spoken-word recognition within the product tradition requires a mechanistic account of how these potential lexical candidates are activated and evaluated with respect to the unfolding input. Similarly, as the utterance *Put the apple on the towel into the box* unfolds, the phrase *on the towel* is temporarily consistent with several syntactic analyses. In one analysis, *on the towel* introduces a Goal argument for the verb *put* (the location where the apple is to be put). In another analysis, it modifies the Theme argument, *the apple*, specifying the location of the Theme (on the towel). Again, a mechanistic account of how people understand utterances requires specifying the nature of the linguistic representations that are accessed and constructed and how these representations are integrated as the utterance unfolds over time. Similar arguments for the importance of time-locked response measures can be made for studies of language production where the speaker must rapidly map thoughts onto linguistic forms that are produced sequentially (Levelt, Roelof, and Meyer 1999). Note, however, that the focus on real-time measures and mechanisms has led most researchers to study comprehension and production separately, often within limited but highly controlled contexts.

In contrast, psycholinguistic research within the action tradition has typically focused on interactive conversation involving two or more subjects engaged in a task that typically has real-world referents and well-defined goals. One reason is that many aspects of utterances in a conversation can only be understood with respect to the context of the language use, which includes the time, place, and participants' conversational goals, as well as the collaborative processes intrinsic to conversation. For example, Clark (1992) points out that in the utterance *Look at the stallion*, the expression *the stallion* could refer to a horse in a field, a painting of a horse, or even a test tube containing a blood sample taken from a stallion, depending on the context of the utterance. Moreover, many of the characteristic features of conversation emerge only when interlocutors have joint goals and when they participate in the dialogue as both a speaker and an addressee.

We can illustrate some of these characteristics by examining a fragment of a conversation from a study by Brown-Schmidt, Campana, and Tanenhaus (chapter 6, this volume). Brown-Schmidt and colleagues used a modified version of a referential communication task, originally introduced by Krauss and Weinheimer (1966). Pairs of participants, separated by a curtain, worked together to arrange blocks in matching configurations and to confirm those configurations. The excerpt includes many well-documented aspects of task-oriented dialogue, including fragments that can only be understood as combinations of utterances between two speakers, false starts, overlapping speech (marked by asterisks), and negotiated referential terms (e.g., *vertically* meaning up and down).

Speaker	Utterance
1	*ok, ok I got it* ele . . . ok
2	alright, *hold on*, I got another easy piece
1	*I got a* well wait I got a green piece right above that
2	above this piece?
1	well not exactly right above it
2	it can't be above it
1	it's to the . . . it doesn't wanna fit in with the cardboard
2	it's to the right, right?
1	yup
2	w- how? *where*
1	*it's* kinda line up with the two holes
2	line 'em right next to each other?
1	yeah, vertically
2	vertically, meaning?
1	up and down
2	up and down

Analyses of participants' linguistic behavior and actions in these tasks has provided important insights into how interlocutors track information to achieve successful communication (Clark 1992, 1996). Moreover, the findings from these studies illustrate that the establishment of a referent is not simply an individual cognitive process. Rather it is arrived at as the result of coordinated actions among two or more individuals across multiple linguistic exchanges (Clark and Wilkes-Gibbs 1986).

Why Bridge?

It is tempting to view research in the action and product traditions as complementary. Research in the product tradition examines the early perceptual and cognitive processes that create linguistic representations, whereas research in the action tradition focuses on subsequent cognitive and social-cognitive processes that build on and use these representations. Although there is some truth to this perspective, it can also be misleading. First, as we have seen, the language used in interactive conversation is dramatically different from the scripted, carefully controlled language studied in the product tradition. The characteristics of natural language illustrated in the excerpt from Brown-Schmidt and colleagues (chapter 6, this volume) are ubiquitous, yet they are rarely studied outside of the action tradition. On the one hand, they raise important challenges for models of real-time language processing within the product tradition, which are primarily crafted to handle fluent, fully grammatical well-formed language. On the other hand, it will be difficult to evaluate models of how and why these conversational phenomena arise without explicit mechanistic models that can be evaluated using real-time methods.

Second, and perhaps most importantly, the theoretical constructs developed within each tradition offer competing explanations for phenomena that have been the primary concern of the other tradition. For example, the product-based construct of *priming* provides an alternative mechanistic explanation for phenomena such as lexical and syntactic entrainment (the tendency for interlocutors to use the same words and/or the same syntactic structures). A priming account does not require appeal to the action-based claim that such processes reflect active construction of common ground between interlocutors (cf. Pickering and Garrod, forthcoming). Likewise, the tendency of speakers to articulate lower-frequency words more slowly and more carefully, which has been used to argue for speaker adaptation to the needs of the listener, has a plausible mechanistic explanation in terms of the attentional resources required to sequence and output lower-frequency forms.

Conversely, the interactive nature of conversation may provide an explanation for why comprehension is so relentlessly continuous. Most work on comprehension within the product tradition takes as axiomatic the observation that language processing is continuous. If any explanation for *why* processing is incremental is offered, it is typically that incremental processing is necessitated by the demands of limited working memory: the system would be overloaded if it buffered a sequence of words rather then interpreting them immediately. However, working-memory explanations of this type are not particularly compelling. One could alternatively argue that delaying interpretation might reduce demands on working memory, by allowing comprehenders to avoid computing multiple analyses and having to revise premature commitments that could be avoided by taking into account immediately upcoming information. In fact, the first-generation models of language comprehension—models that were explicitly motivated by considerations of working-memory limitations—assumed that comprehension was a form of sophisticated catch-up in which the input was buffered long enough to accumulate enough input to reduce ambiguity (e.g., Fodor, Bever, and Garrett 1974; Marcus 1980). However, there is a clear need for incremental comprehension in interactive conversation. Participants, who are simultaneously playing the roles of speaker and addressee, need to plan and modify utterances in midstream in response to input from an interlocutor. This type of give-and-take requires incremental comprehension.

Finally, the action and product traditions often have different perspectives on constructs that are viewed as central within each tradition. Consider, for example, the notion of *context*. Within the product tradition, context is typically viewed either as information that enhances or instantiates a context-independent core representation or as a *correlated constraint* in which information from higher-level representations can, in principle, inform linguistic processing at lower levels of representation. Specific debates about the role of context include whether, when, and how (1) lexical context affects sublexical processing, (2) syntactic and semantic context affect lexical processing, and (3) discourse and conversational context affect syntactic processing. Each of these questions involves debates about the architecture of the processing system and the flow of information between different types of representations—classic information-processing questions. In contrast, we have already noted that within the action tradition context includes the time, place, and participants' conversational goals, as well as the collaborative processes intrinsic to conversation. A central tenet is that utterances can only be understood relative to these factors. Although these notions can be conceptualized as a form of correlated constraint, they are much more intrinsic to the comprehension process than that characterization would suggest.

Given these factors, we believe that combining and integrating the product and action approaches is likely to prove fruitful by allowing researchers from each tradition to investigate phenomena that would otherwise prove intractable. Moreover, research that combines the two traditions is likely to deepen our understanding of language processing by opening up each tradition to empirical and theoretical challenges from the other tradition.

The Methodological Challenge

With the exception of an occasional shot fired across the bow (e.g., Clark and Carlson 1981; Clark 1997), the action and product traditions have not fully engaged one another. We believe that one reason is methodological. The traditional techniques in the psycholinguist's toolkit for studying real-time language processing have required using either text or prerecorded audio stimuli in contextually limited environments that cannot be used with more naturalistic tasks.

Table 1.1

Desiderata for a response measure bridging the action and product traditions

Action-based requirements:

1. Measure can be used with *conversational language*.
2. Measure can be used to monitor *language production and language comprehension*
3. Measure should not *interrupt* or *interfere* with the primary task of engaging in conversation

Product-based requirements:

4. Measure must be *sensitive* to rapid, unconscious processes underlying production and comprehension.
5. Measure should be closely *time-locked* to the input (for comprehension) and output (for production).
6. Measure should have a well-defined *linking hypothesis*.

Requirement for understanding development and deficits:

7. Measure can be used with young *children* and *special populations*.

Table 1.1 lists seven desiderata for a methodology that bridges the action and product traditions of psycholinguistic research. The first three desiderata are essential if the paradigm is to be useful in studies of interactive conversation. First, the method must be usable with *conversational language* in relatively natural behavioral contexts. Second, because both speaking and understanding are integral components of interactive conversation, the response measure should provide insights into both *language production and language comprehension*. Third, the response measure should not *interrupt* or *interfere* with the primary task of the participants—engaging in a conversation.

The next three desiderata are essential for investigating the time course of language processing with a fine-enough grain to meet the criteria of a successful product method. Specifically, the fourth desideratum states that the response measure must be *sensitive* to the rapid, typically unconscious processes that underlie comprehension and production. Fifth, the response measure must be closely *time-locked* to the input in order to provide insights into the rapidly occurring processes that underlie comprehension and production. Sixth, the response measure should have a well-defined *linking hypothesis*. By this we mean a theory, ideally one that can be formalized, that maps hypothesized underlying processes onto behavioral patterns. Without clear linking hypotheses, it is difficult to relate behavioral data patterns to theoretical constructs (Tanenhaus, Spivey-Knowlton, and Hanna 2000). Finally, if we are to understand the development of the relevant processes, the method should allow us to investigate comprehension and production processes in *children* and *special populations*.

In the remainder of this chapter we argue that monitoring saccadic eye movements as people engage in spoken-language processing in natural tasks satisfies all seven of these methodological criteria. In the next section, we briefly review the properties of saccades in natural-scene perception and investigate how they reflect momentary states of attention. We then illustrate how lightweight visor systems could be applied to common dialogue tasks in the action tradition, thus satisfying the three desiderata for an appropriate action-based response measure. In the following section, which forms the core of the chapter, we demonstrate that the eye-gaze paradigm meets the central criteria for product methods, namely, sensitivity, time locking, and availability of a linking hypothesis. We also use this section to introduce the reader to the methods employed to collect and analyze data, and provide some illustrative examples, focusing on word recognition and syntactic processing, in adults and in children. We conclude that section by discussing a potential limitation of the visual-world paradigm—the constraints imposed by a restricted task-relevant visual world—and summarize work addressing these *closed-set* concerns. We conclude the chapter with a discussion of present and future uses of eye gaze in studies of conversational language, highlighting issues that we believe will increasingly take center stage in psycholinguistic research.

Fixation as a Measure of Attention in Natural Tasks

During everyday tasks involving vision, such as reading a newspaper, looking for the car keys, making a cup of coffee, and conversing about objects in the immediate environment, people rapidly shift their gaze to bring task-relevant regions of the visual field into the central area of the fovea (e.g., for reviews see Hayhoe 2000; Kowler 1995). Eye movements are necessary because visual sensitivity differs across the retina. Acuity is greatest in the central portion of the fovea, then markedly declines. The organization of the retina can be viewed as a compromise between the need to maintain sensitivity to visual stimuli across a broad range of the visual field, while also allowing detailed spatial resolution for task-relevant aspects of the visual field. In addition, this division of labor helps restrict most processing to a relevant subset of the visual field, reducing the amount of information being made available from the visual environment. However, it also requires an eye-movement system to quickly bring new regions of the field into the fovea, where visual acuity is greatest. These gaze shifts are accomplished by saccadic eye movements (Hayhoe 2000; Kowler 1995, 1999; Liversedge and Findlay 2001).

Saccades are rapid ballistic eye movements. During a saccade, the eye is in motion for 20 to 60 ms, with the duration of the saccade related to the distance that the eye travels. At peak velocity, the eye can be moving between 500 and 1,000 degrees per second. During a saccade, sensitivity to visual information is dramatically reduced. Suppression of visual information occurs in part because of masking, and in part because of central inhibition (see the following sources and references therein: Kowler 1995; Liversedge and Findlay 2001; Rayner 1998).

A saccade is followed by a fixation that typically lasts for 200 ms or more depending on the task. The minimal latency for planning and executing a saccade is approximately 150 ms when there is no uncertainty about target location. In reading, visual search, and other tasks in which there are multiple target locations, saccade latencies are somewhat slower, typically about 200 to 300 ms. The pattern and timing of saccades, and the resulting fixations, are among the most widely used response measures in the cognitive sciences, providing important insights into the mechanisms underlying attention, visual perception, reading, and memory (Rayner 1998). Overviews of eye movements in scene perception are provided by Henderson and Hollingsworth (2003) and Henderson and Ferreira (forthcoming).

Eye Gaze in Interactive Conversation

The development of accurate, relatively inexpensive head-mounted and remote eye-tracking systems has made it possible to monitor eye movements as people perform natural tasks.[1] Eye movements naturally occur rapidly in response to even low-threshold signals, and because they are ballistic, there is little uncertainty about when a saccade has been initiated and what part of the visual field is being fixated. Crucially, they are closely linked to attention. Although attention *can* be directed to regions of space not currently being fixated, or about to be fixated, a growing body of behavioral and neurophysiological research supports a close link between fixation and spatial attention (Findlay, forthcoming; Kowler 1999; Liversedge and Findlay 2001). Thus, to the extent that attention and shifts in attention are closely time-locked to the processes that underlie comprehension and production, eye movements should be informative about real-time language processing.

Monitoring eye movements as people understand and produce language related to ongoing tasks in a circumscribed visual world would seem to meet the important criteria for an action-based measure. For example, monitoring eye movements in a referential communication task would not modify the basic task. This is illustrated in figure 1.1, which presents a schematic of a well-studied variant of a referential communication task introduced by Clark and his colleagues (e.g., Clark and Wilkes-Gibbs 1986).

Two naive participants, a matcher and a director, are separated by a barrier. Each has the same set of shapes arranged in different positions on a numbered grid. The participants' goal is for the matcher to rearrange the shapes on his grid to match the arrangement on the director's grid. In the schematic both the director and the matcher

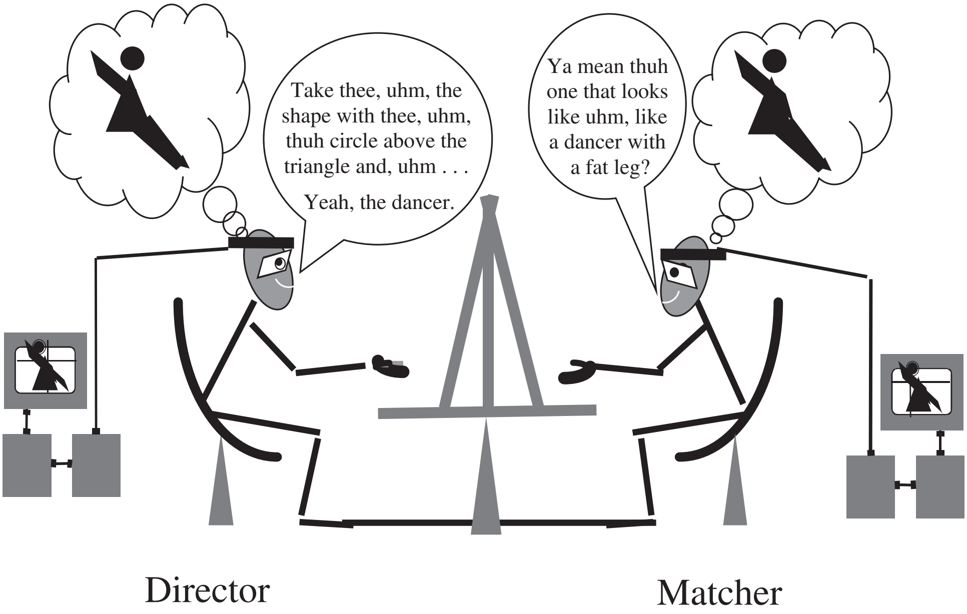

Take thee, uhm, the shape with thee, uhm, thuh circle above the triangle and, uhm . . .
Yeah, the dancer.
Ya mean thuh one that looks like uhm, like a dancer with a fat leg?
Director
Matcher

Figure 1.1
Schematic of using eye tracking in a referential communication task with Tangrams.

are wearing visor-mounted eye trackers. With a screen-based variant of the task, one could monitor eye movements using a remote eye tracker without placing anything on the head of the participants. The crucial question, then, is whether eye movements in natural tasks in a circumscribed visual world are sensitive to comprehension and production processes. Also, we would like to know if the eye movements meet the necessary criteria for a product-based measure, namely, sensitivity, time locking, and the presence of a well-defined linking hypothesis.

The use of eye movements as a real-time measure of spoken-language processing was pioneered by Cooper (1974), who demonstrated that the timing of participants' eye movements to pictures was closely time-locked to relevant information in a spoken story. More recently, Tanenhaus et al. (1995) showed that when participants follow spoken instructions to manipulate objects in a task-relevant ''visual world,'' fixations to task-relevant objects are closely time-locked to the unfolding utterance. Since then, a body of research has demonstrated that eye movements can be used to trace the time course of language comprehension and, more recently, language production (see Henderson and Ferreira, forthcoming).

The Visual-World Paradigm Applied to Issues in the Product Tradition

We now review three applications of the visual-world paradigm to address classic language-as-product questions. We begin with a review of a study by Allopenna, Magnuson, and Tanenhaus (1998) that traces the time course of lexical access in continuous speech. We use the Allopenna et al. study to illustrate how eye-movement data are analyzed. We also use this study to illustrate *sensitivity*, *time locking*, and a formalized *linking hypothesis* between underlying processes and fixations. We then review work by Spivey, Tanenhaus, and colleagues that illustrates how the paradigm can be extended to syntactic processing, suggesting that such a method taps processes at multiple levels of representation. We conclude with work by Fernald, Swingley, Trueswell, and colleagues that illustrates how the paradigm can be extended to investigations of language processing in children (desideratum 7).

Tracking Lexical Access in Continuous Speech

Allopenna, Magnuson, and Tanenhaus (1998) evaluated the time course of activation for lexical competitors that shared initial phonemes with the target word (e.g., *beaker* and *beetle*) or that rhymed with the target word (e.g., *beaker* and *speaker*). In the studies by Allopenna and colleagues, 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*).

A schematic of a sample display of pictures is presented in figure 1.2, panel (a). The pictures include the target (the beaker), the cohort (the beetle), a picture with a name that rhymes with the target (*speaker*), and the unrelated picture (the carriage). For purposes of illustrating how eye-movement data are analyzed, we will restrict our attention to the target, cohort, and unrelated pictures. The particular pictures displayed are used to exemplify types of conditions and are not repeated across trials. 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. On trial 1, the hypothetical participant initiated a fixation on 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 trial 2, the fixation on the target begins a bit later. On trial 3, the first fixation is on the cohort, followed by a fixation on the target. On trial 4, the first fixation is on the unrelated picture. Trial 5 shows another trial where the initial fixation is on the cohort. Panel (c)

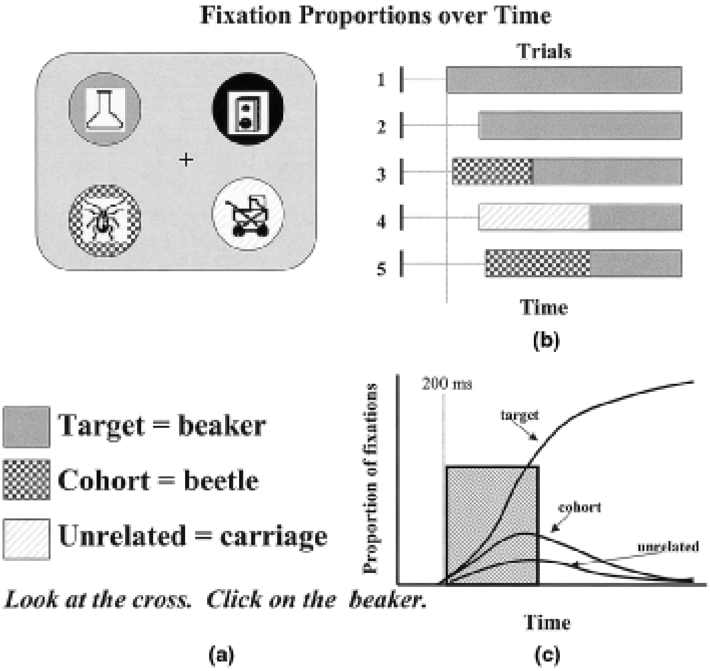


Figure 1.2
Schematic showing how proportions of fixations are calculated and plotted over time.

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 changes 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.

Researchers often define a window of interest, illustrated by the rectangle in panel (c). For example, one might want to focus on the fixations on 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 arrives. The proportion of fixations on pictures or objects, the time spent fixating on the alternative pictures (essentially the area under the curve, which is a simple transformation of proportion of fixations), and the number and/or proportion of saccades generated to pictures in this region can then be analyzed. These measures are all highly correlated.

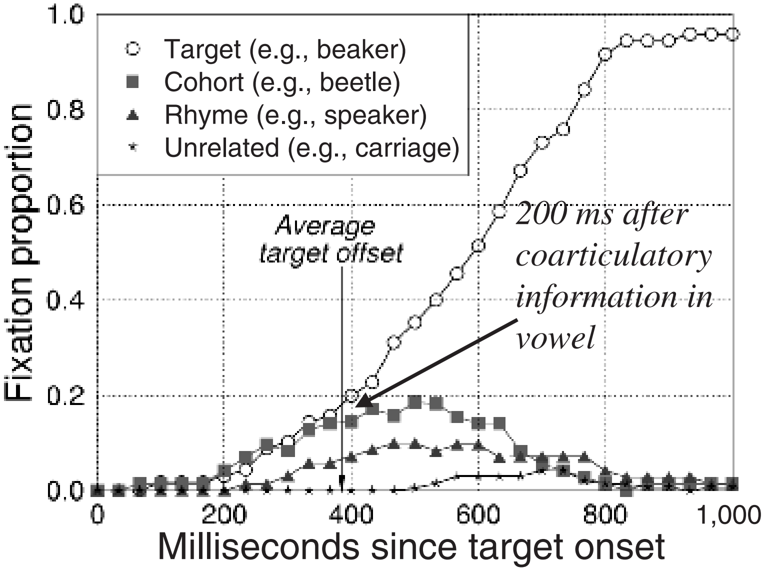


Figure 1.3
Results of Allopenna, Magnuson, and Tanenhaus 1998.

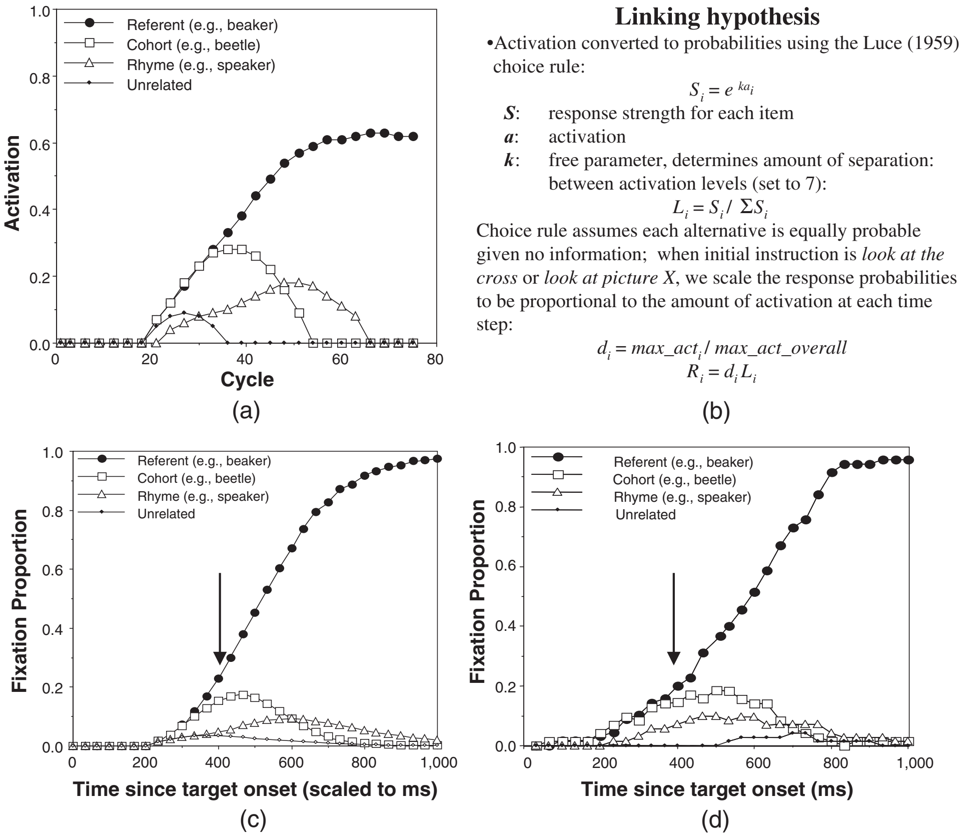
Figure 1.3 shows the actual data from the experiment by Allopenna and colleagues (1998). The figure plots the proportion of fixations on the target, cohort, rhyme, and unrelated picture. Until 200 ms, nearly all of the fixations are on the fixation cross. These fixations are not shown. The first fixations on 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 200 to 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 on the target began to diverge from the proportion of fixations on the cohort. Subsequent research has established that cohorts and targets diverge approximately 200 ms after the first phonetic input, including coarticulatory information in vowels, that provides probabilistic evidence favoring the target (Dahan et al. 2001; Dahan and Tanenhaus, forthcoming).

Shortly after fixations on the target and cohort begin to rise, fixations on rhymes start to increase relative to the proportion of fixations on the unrelated picture. This result discriminates between predictions made by the cohort model of spoken-word recognition and its descendants (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, and continuous mapping models, such as TRACE (McClelland and Elman 1986), which predict competition from similar words that mismatch at onset (e.g., rhymes). The results strongly confirmed the predictions of continuous mapping models.

We can now illustrate a simple linking hypothesis between an underlying theoretical model and fixations. 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.[2]

Allopenna and associates formalized this linking hypothesis by converting activations 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. Panel (a) in figure 1.4 shows the activation values for *beaker*, *beetle*, *carriage*, and *speaker*, generated by a TRACE simulation. Panel (b) shows the equations used in the linking hypothesis.

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*, we scale the response probabilities to be proportional to the amount of activation at each time step using the following equations, where max_t is the maximum activation at a particular time step, m is a constant equal to the maximum expected activation (e.g., 1.0), i is a particular item, and d_t is the scaling factor for time step t. 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, we introduce a 200 ms delay because programming an eye movement takes approximately 200 ms (Matin, Shao, and Boff 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. 2001). When the linking hypothesis is applied to TRACE simulations of activations for the stimuli used by Allopenna and colleagues, it generates the predicted fixations over time shown in figure 1.4, panel (c). Note that the linking hypothesis transforms the shape of the functions because it introduces a nonlinear transformation. This highlights the importance of developing and using explicit linking hypotheses. The actual data are repeated figure 1.4, panel (d). The fixations over time on the target, the cohort competitor, and a rhyme competitor closely matched the predictions generated by the hypothesis linking activation levels in TRACE to fixation probabilities over time.



(a)

Linking hypothesis

- Activation converted to probabilities using the Luce (1959) choice rule:

$$S_i = e^{ka_i}$$

S: response strength for each item

a: activation

k: free parameter, determines amount of separation: between activation levels (set to 7):

$$L_i = S_i / \Sigma S_i$$

Choice rule assumes each alternative is equally probable given no information; when initial instruction is *look at the cross* or *look at picture X*, we scale the response probabilities to be proportional to the amount of activation at each time step:

$$d_i = max\_act_i / max\_act\_overall$$

$$R_i = d_i L_i$$

(b)

(c)

(d)

Figure 1.4

Results compared to model. Adapted from Allopenna, Magnuson, and Tanenhaus 1998.

Syntactic-Ambiguity Resolution in Sentence Processing

Temporary "attachment" ambiguities like those we illustrated with the example *Put the apple on the towel* ... have long served as a primary empirical test bed for evaluating models of syntactic processing (Tanenhaus and Trueswell 1995). Crain and Steedman (1985) (also Altmann and Steedman 1988) called attention to the fact that many classic structural ambiguities involve a choice between a syntactic structure in which the ambiguous phrase modifies a definite noun phrase and one in which it is a syntactic complement (argument) of a verb phrase. Under these conditions, the argument analysis is typically preferred. For instance, in *Put the apple on the towel in the box*, readers and listeners will initially misinterpret the prepositional phrase *on the towel* as in-

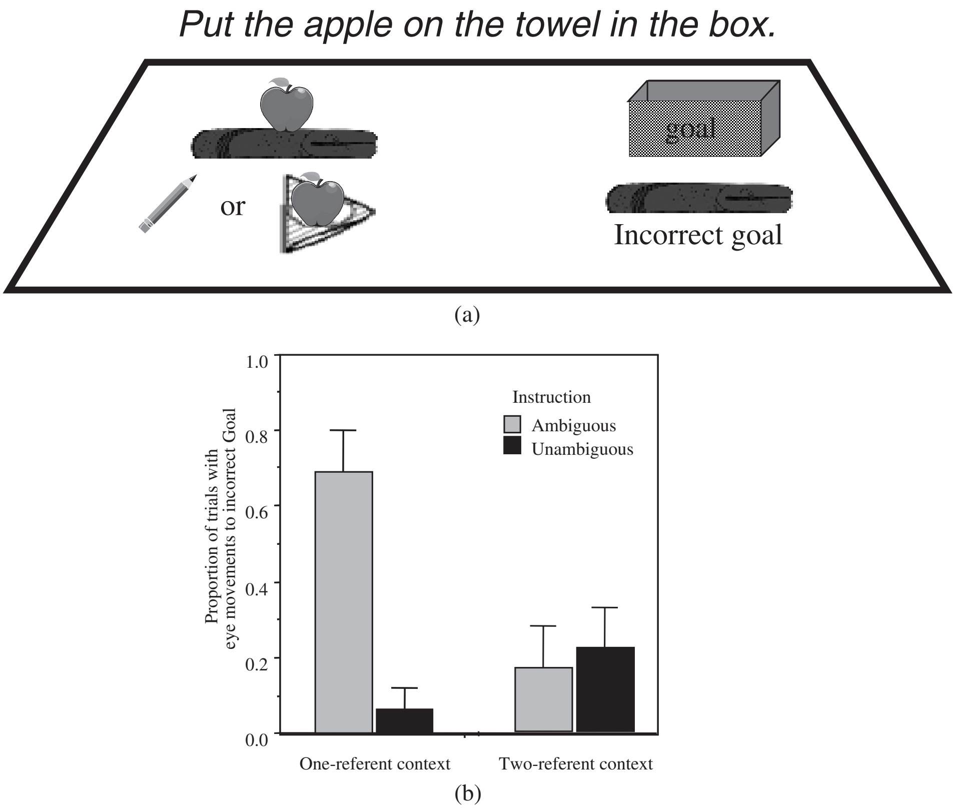


Figure 1.5
Proportion of fixations on Incorrect Goal. Adapted from Spivey et al. 2002.

troducing the Goal argument of *put*, resulting in temporary confusion if later-arriving information required treating the prepositional phrase as an adjunct modifying the Theme argument, *the apple*.

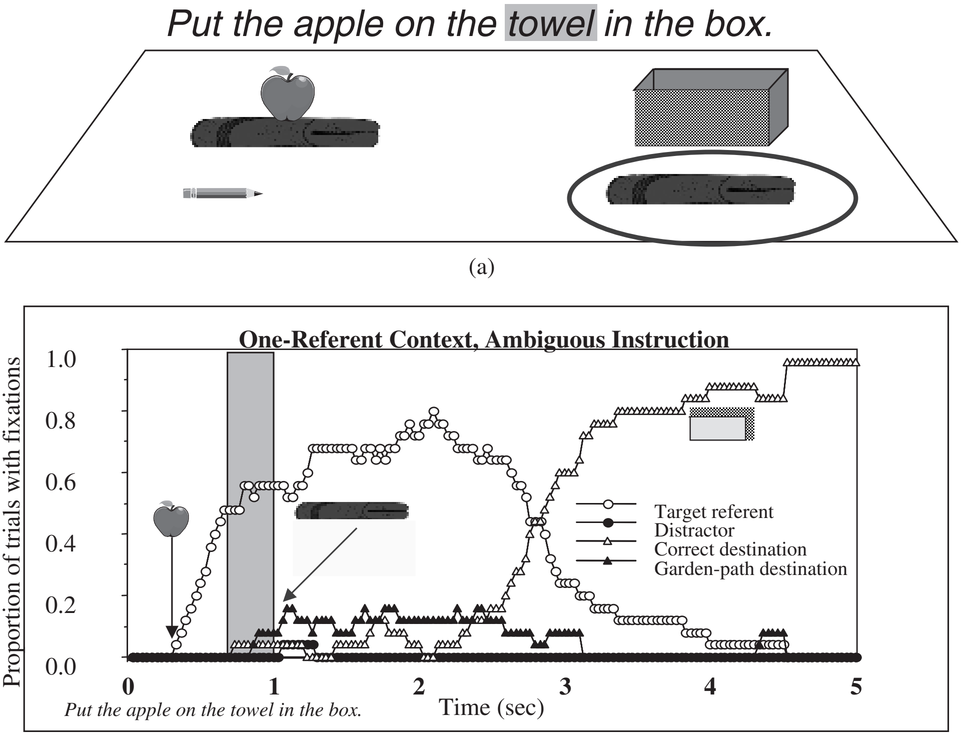
Tanenhaus et al. (1995) and Spivey et al. (2002) presented participants with temporarily ambiguous sentences such as (1) and unambiguous control sentences such as (1b), in contexts like the one illustrated in panel (a) of figure 1.5, which is adapted from Spivey et al. 2002. The objects illustrated in the figure were placed on a table in front of the participant. Participants' eye movements were monitored as they performed the action in the spoken instruction. The objects of interest are the referent of the Theme (the apple on the towel), the garden-path Goal (the empty towel), and the true Goal (the box).

(1) a. Put the apple on the towel in the box.
b. Put the apple that's on the towel in the box.

The figure presents the proportion of looks to each of these objects as the instruction unfolded. The fixations are again remarkably time-locked to the utterance. Toward the end of the word *apple*, fixations on the apple begin to rise sharply. During the middle of the word *towel*, participants begin to fixate on the empty towel, indicating that it is being considered as the Goal. The left-hand side of the graph in figure 1.5, panel (b), presents the proportion of trials with fixations on the garden-path Goal for the temporarily ambiguous (1a) and unambiguous (1b) instructions. Crucially, there were far more fixations on the garden-path Goal with the ambiguous instruction.

Crain and Steedman (1985) also noted that one use of modification is to differentiate an intended referent from other alternatives. For example, it would be odd for (1a) to be uttered in a context in which there was only one perceptually salient apple, such as the scene. However, the instruction in (1a) would be natural in a context with more than one apple—for instance, a display with two apples, one on a towel and one on a napkin. In this context, the modifying phrase *on the towel* provides information about which of the apples is the intended Theme. Crain and Steedman proposed that listeners might initially prefer the modification analysis to the argument analysis in situations that provided the appropriate referential context. Moreover, they suggested that referential fit to the context, rather than syntactic complexity, was the primary factor controlling syntactic preferences (also see Altmann and Steedman 1988).

The apple-on-the-towel experiment also included a condition with two potential referents—for example, an apple on a towel and an apple on a napkin. In the two-referent context, looks to the garden-path Goal were dramatically reduced in the two-referent context (right-hand side of the graph, panel (b) of figure 1.5). Crucially, there was not even a suggestion of a difference between the proportion of looks to the false goal with the ambiguous and the unambiguous instructions. Moreover, the timing of the fixations provided clear evidence that the prepositional phrase was being immediately interpreted as modifying the noun phrase. Participants typically looked at one of the potential referents as they heard the beginning of the instruction—for instance, *put the apple*. On trials in which participants looked first at the incorrect Theme (e.g., the apple on the napkin), they immediately shifted to the correct Theme (the apple on the towel) as they heard *towel*. Moreover, the timing was identical for the ambiguous and unambiguous instructions. Signs of garden pathing in the one-referent ambiguous instruction appeared almost immediately on hearing the potentially ambiguous *on the towel* (figure 1.6).



(a)

(b)

Figure 1.6
Proportion of fixations on Incorrect Goal over time. Adapted from Spivey et al. 2002.

Development of Language Use

Researchers have also begun to use eye gaze during listening with infants, toddlers, and young children to address developmental issues in language processing (e.g., Swingley, Pinto, and Fernald 1998, 1999; Swingley and Aslin 2002; Trueswell et al. 1999). The time course of children's eye movements is established either by inspecting a videotape of the child's face frame by frame (Swingley, Pinto, and Fernald 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. 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 et al. 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, Pinto, and Fernald (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 and colleagues' (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 and Aslin 2002). Consideration of the alternative object did not occur when the object was not a cohort member. 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 5-year-olds and 8-year-olds, first reported in Trueswell et al. 1999. The experiments were modeled after the adult apple-on-the-towel study 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 on spoken instructions that contained temporary ambiguities such as *Put the frog on the napkin in the box* (see figure 1.7). Like the apple example, the phrase *on the napkin* is briefly ambiguous between a Goal argument of the verb *put* and a Modifier of the noun phrase *the frog*, specifying a property of a particular frog. The phrase is disambiguated in favor of the Modifier interpretation by the presence of a second Goal phrase (*in the box*).

The striking finding was that 5-year-olds showed a strong preference for interpreting *on the napkin* as the Goal of *put*, even when the referential scene supported a Modifier interpretation (e.g., two frogs, one on a napkin; figure 1.7). On hearing *on the napkin*, 5-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 eye movements was similar to what was observed in the one-referent condition of adults—

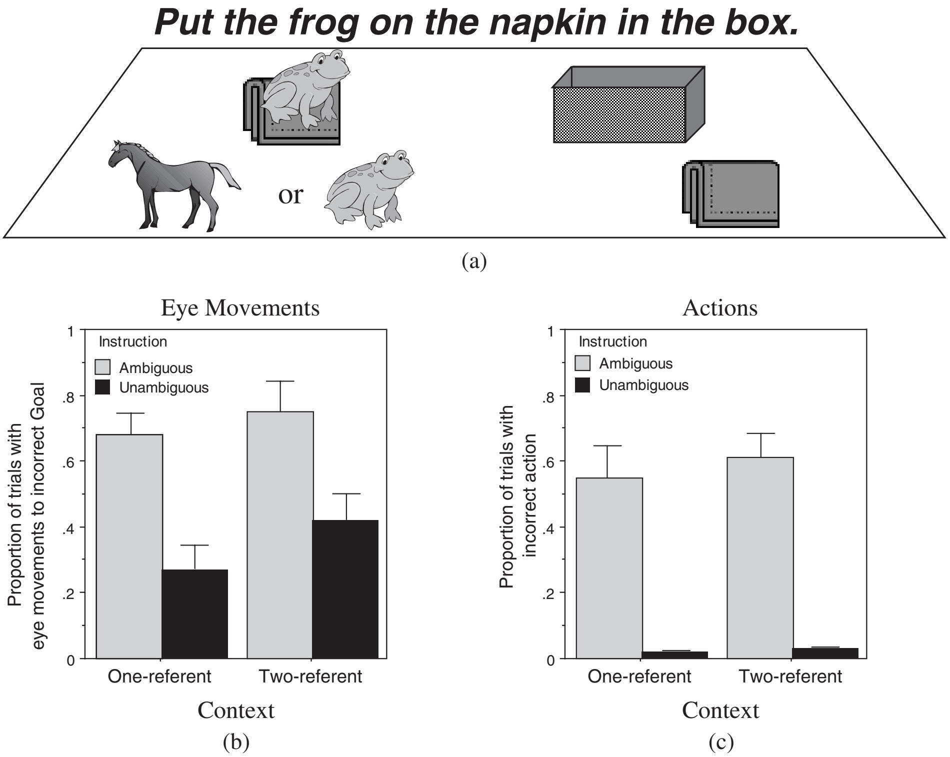


Figure 1.7

Child Parsing Data. Proportion of fixations on Incorrect Goal. Adapted from Trueswell et al. 1999.

that is, approximately 600 ms after the onset of the word *napkin*—but for children this pattern of Goal looks also arose in two-referent contexts. In fact, 5-year-olds' preference for the Goal interpretation was so strong that they showed little sign of revising it; on 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 adultlike 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.

In contrast to the responses of 5-year-olds, 8-year-olds' and adults' responses to the temporarily ambiguous stimuli were found to depend on the referential scene provided. In particular, the mere presence of a two-referent scene eliminated measurable signs of syntactic misanalysis of the ambiguous phrase: there were few looks to the potential Goal and few incorrect actions as compared to one-referent scenes. This finding is consistent with the earlier apple-on-the-napkin studies described above (Tanenhaus et al. 1995; Spivey et al. 2002).

Both of these findings (from Swingley, Pinto, and Fernald 1999 and Trueswell et al. 1999) suggest that there is considerable continuity in the language-processing system throughout development: lexical and sentential interpretation proceed incrementally and are designed to coordinate multiple information sources (e.g., linking what is heard to what is seen within milliseconds). However, the differences between 5- and 8-year-old children reported by Trueswell et al. (1999) suggest that significant developmental differences exist. These differences likely pertain to how children learn about sources of evidence relevant to linguistic and correlated nonlinguistic 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.

Closed-Set Issues

We have established that the eye-movement paradigm meets the three essential criteria for a measure of real-time spoken language processing: the response measure is sensitive, it is time-locked, and it has a clear linking hypothesis. There is, however, an aspect of the methodology that is potentially problematic. The use of a visual world with a limited set of pictured referents and a limited set of potential actions creates a more restricted environment than language processing in many, if not most, contexts. Certainly, these characteristics impose more restrictions than most psycholinguistic tasks. We will refer to this as the *closed-set* problem.

Two aspects of the closed-set problem could, in principle, limit the usefulness of the visual-world paradigm. The first is that the closed set might create task-specific strategies that result in language processing that does not generalize beyond the specific situations created within the experiment. The second is that the paradigm might not be sensitive to characteristics of linguistic knowledge and experience lying outside of the closed set that has been established on a given trial. We will illustrate these two potential problems using the experiment by Allopenna, Magnuson, and Tanenhaus (1998).

We know that as a spoken word unfolds over time, recognition takes place against a backdrop of partially activated alternatives that compete for recognition. As a consequence the recognition of a spoken word is influenced by the similarity structure created by those lexical candidates that most closely match the input. The number of competitors, their frequency of occurrence in the language, and the frequency of occurrence of the target word itself all affect recognition (e.g., Luce and Pisoni 1998; Marslen-Wilson 1987, 1990).

We can use the linking hypothesis discussed earlier to help clarify the distinction between the task-specific strategy and the sensitivity issues. Recall that the linking hypothesis assumes that the equation that determines the response strength for each lexical candidate at a moment in time is computed using the activation of its lexical representation. The activation of a lexical candidate will be affected by the entire lexicon—that is, it will be determined in part by its neighbors. However, only the items in the response set enter into calculations for response selection. The task-specific strategy concern is that processing of the input might bypass the activation process. The sensitivity concern is that the effects of response selection, or alternatively, the effects of presenting the response set, are so strong that they mask any effects of lexical neighborhoods.

Strategies In the experiment by Allopenna and colleagues, the potential response set on each trial was limited to four pictured items. If participants adopted a task-specific strategy, such as implicitly naming the pictures, then the unfolding input might be evaluated against these activated names, effectively bypassing the usual activation process. A related argument could be made for the parsing studies. Here the argument would be that listeners process the language shallowly, extracting only the information necessary to inform the action.

Along with our colleagues, we have tried to articulate the task-specific strategy concerns and to address them empirically (see Dahan and Tanenhaus, forthcoming). For example, the patterns of results observed in the Allopenna et al. studies are observed even when the preview time for the pictures is limited. In addition, we find robust effects of frequency for targets and cohort competitors (Dahan, Magnuson, and Tanenhaus 2001). These are unexpected with a closed-set strategy because the a priori probability of each of the pictured names is equated, which should eliminate or strongly reduce frequency effects. Crucially, the prenaming strategy is inconsistent with the fact that we observe input-driven looks to pictures whose names are not related to the target but are visually similar to its referent—for example, a picture of a turtle for the target *igloo* (Dahan and Tanenhaus 2003). This result is unexpected if participants are implicitly naming the pictures. However, it is predicted by the hypothesis that the link to the pictured referents is made via perceptual/conceptual representations that are accessed as the target word is being processed.

A crucial empirical test of the task-specific strategies argument in sentence processing comes from work by Craig Chambers and colleagues (Chambers et al. 2002; Chambers 2001). In Chambers et al. 2002, experiment 2, participants were presented with six objects in a workspace. On critical trials, the objects included a large and a small container—for example, a large can and a small can. The critical variable manipulated in the workspace was whether a to-be-mentioned (Theme) object, like a cube, could fit into both of the containers, as was the case for a small cube, or could only fit into the larger container, as was the case for a large cube. Thus the size of the Theme object determined whether one or two potential Goal objects were compatible referents. The instructions—for instance, *Pick up the cube. Put it inside a/the can*—manipulated whether the Goal was introduced with the definite article *the*, which presupposes a unique referent, or the indefinite article *a*, which implies that the addressee can choose from among more than one Goal.

As expected, when the definite article, which assumes a uniquely identifiable referent, was used in the instruction with the small cube, participants were confused compared to when an indefinite article was used in the instruction. Eye-movement latencies to fixate the Goal object chosen by the participant were slower in the definite condition than in the indefinite condition. However, for the large cube, confusion with the definite article (relative to a baseline with only a single large container) was eliminated. This result by itself is consistent with two explanations, both of which assume that listeners dynamically update referential domains to include only objects that afford the required action—that is, containers that the object in hand would fit into. The first explanation is that participants simply adopt a task-specific strategy. After picking up the cube and hearing *put it inside*, they focus their attention on the only Goal object compatible with the action, therefore bypassing more detailed linguistic processing. This hypothesis predicts that participants will not be confused when the indefinite article *a* is used because there is still only one possible action. The second explanation is that the participants fully process the instruction, using the information provided by each word. This explanation predicts that the indefinite article should be infelicitous when there is only one compatible Goal object because it implies that there is more than one possible Goal. The results were clearly inconsistent with the task-specific strategy explanation: latencies in the indefinite condition increased in the one-compatible-Goal condition compared to the two-compatible-Goal condition, despite the fact that the one-Goal condition afforded only one possible action.

Sensitivity Evaluating sensitivity outside of the closed set is relatively straightforward in the case of spoken-word recognition. We need to determine whether the visual-world paradigm shows effects of the nondisplayed, nonmentioned lexical neighbors. A body of such results now exists. For example, Dahan et al. (2001) introduced misleading coarticulatory information about upcoming place of articulation by creating cross-spliced tokens such as *neck* in which the onset and vowel were taken from either a word (e.g., *net*) or a nonword (e.g., *nep*). The effects of the cross-splicing was stronger when the initial portion of the target was consistent with a word compared to a nonword, even though that word was never mentioned and its referent was never displayed. This result demonstrates strong effects of a nondisplayed, nonmentioned lexical competitor.

Perhaps the most compelling evidence for sensitivity comes from a series of studies by Magnuson and colleagues (Magnuson 2001; Magnuson, Tanenhaus, and Aslin 2003). Magnuson and colleagues used a variant of the Allopenna et al. paradigm to examine the effects of lexical neighbors on recognition of spoken words. Target words matched in frequency were chosen that varied in whether they came from high- or low-density neighborhoods, where density was defined as either the number of words that differed by only a single phoneme (neighborhood density) or whether they had few or many cohorts (cohort density). The displays presented a picture of a target along with three pictures with unrelated names. Cohorts and noncohort neighbors were never pictured or mentioned throughout the course of the experiment. Nonetheless, clear effects of both cohort and neighborhood density were found, including theoretically significant time-course differences that had not been previously observed with other paradigms (also see Magnuson et al. 2003 for similar results with artificial lexicons). Despite the closed set, then, the paradigm is sensitive to effects coming from the full lexicon. Interestingly, similar conclusions have been made about the eye-gaze patterns of 24-month-olds. Swingley and Fernald (2002) looked at the speed of response to known and unknown words and found that on this task, children sought out a ball on hearing *ball* even when no ball was present. All of these results confirm the most basic claim of the Allopenna et al. linking hypothesis, namely, that the activation of the lexical candidates is determined by the entire lexicon, with the visual world operating as a response selection set.

There is also an emerging literature that addresses sensitivity concerns in sentence processing. In the adult PP-attachment studies we reviewed earlier, the two-referent contexts completely eliminated any hint of a garden path for the temporarily ambiguous instructions. This result is somewhat surprising, because the verb *put* obligatorily occurs with a goal argument. In the parallel literature in reading there is clear evidence that referential factors are partially modulated by the availability of syntactic alternatives, especially those tied to verb-based frequencies (see MacDonald, Pearlmutter, and Seidenberg 1994; Snedeker, Thorpe, and Trueswell 2001; Snedeker and Trueswell 2004; Trueswell and Tanenhaus 1994). Consider, for example, a well-known study by Britt (1994).

Britt manipulated referential context and verb bias in a study using self-paced reading. The discourse context introduced one or two potential referents (e.g., a book about the Civil War and another book). The target sentence contained a temporarily ambiguous PP-phrase (e.g., *Susan dropped the book <u>on the Civil War</u> onto the table*) that was preceded by a verb that optionally takes a Goal argument (e.g., *Susan dropped the book* . . .) or a verb that obligatorily takes a Goal argument (e.g., *Susan put the book* . . .). Britt found that two-referent contexts eliminated garden paths due to the Goal-argument bias for the optional-Goal verbs, but not for the obligatory-Goal verbs.

Why then were the context effects so strong in the visual-world, *put-the-apple-on-the-towel* studies? Certainly, referential-context effects might be stronger in visual-world situations because the context is copresent with the linguistic input rather than held in memory, as in reading studies. Nonetheless, the fact that a constraint as strong as verb bias was completely overridden raises concerns about sensitivity and/or strategies because of the highly constraining visual context and limited set of potential actions that could be performed with the objects in the workspace.

However, several visual-world studies have demonstrated clear effects of verb-based constraints. For example, using the anticipatory-looks paradigm introduced by Altmann and Kamide (1999), Boland (2002) found that as they heard a verb, participants were more likely to make anticipatory looks to referents of likely recipient arguments than to referents of plausible adjuncts. Moreover, Snedeker, Thorpe, and Trueswell (2001) have demonstrated an interaction between referential context and verb-bias in syntactic-ambiguity resolution. These studies used instructions such as *Tickle/Feel/Choose the frog with the feather*. The verb had a strong instrument bias (e.g., *tickle*) or was equibiased between taking an instrument or a modifier (*feel*) or had a strong modifier bias (*choose*). The contexts contained an instrument (e.g., a large feather), and either a single frog with a small feather (one-referent context) or two frogs, one of which was holding a small feather and one of which was holding another object (two-referent context). As expected, one-referent scenes did induce more instrument actions than two-referent scenes (i.e., two-referent scenes supported a restrictive modifier reading of *with the feather* and hence reduced the instrument interpretation). However, the extent to which the two-referent contexts reduced instrument responses was modulated by verb bias. For equibias and modifier-bias verbs, two-referent scenes resulted in very few instrument actions (i.e., picking up the feather to do the action occurred on only about 5 percent of the trials for both verb types). A substantial number of instrument responses were observed, though, in two-referent scenes when the verb was instrument biased (65 percent instrument responses). Snedeker and Trueswell (2004) suggest this pattern arises because lexical biases in this condition so strongly support an instrument reading that the competing NP-modifier interpretation is often inaccessible to the listener. Thus, like the results of Magnuson et al. 2003, these data suggest that eye-gaze responses are guided in part by the availability of linguistic alternatives; here verb-specific frequency information modulates the influence of the referential scene.

Why then would referential context have such strong effects in the *put-the-apple/frog* studies even though *put* obligatorily requires a Goal argument and thus is more strongly biased than a verb such as *tickle*, for which an instrument in not obligatory? More research is needed to provide a definitive answer but the outline of a plausible explanation is beginning to emerge (Snedeker and Trueswell 2004). First, in an instruction such as *Put the apple on the towel* . . . , the preposition introducing the PP, *on*, specifies a location, regardless of whether the PP modifies the noun phrase or the verb. In contrast, in the Britt 1994 study, the sense of *on* in the prepositional phrase differed when it introduced a Goal argument and when it modified the noun. The location sense of *on*, which corresponds to the Goal argument, is the more frequent sense, especially when *on* follows a noun phrase after a verb. The sense of the preposition *with* also differs for the modifier and the instrument attachment in *tickle the frog with the feather*, with the instrument sense more frequent when *with* follows a verb. Thus in the Britt 1994 and Snedeker and Trueswell 2004 studies, the referential constraints from the two-referent context are pitted against two opposing constraints for strongly biased Goal or Instrument verbs. One bias comes from the preposition, the other from the verb. The preposition bias is especially strong in Goal constructions like those used by Britt.

A second possible factor is specific to the noun modification/instrument ambiguity. In natural tasks, people typically fixate on an object before reaching for it (cf. Ballard et al. 1997). Instrument actions, such as tickling a frog with a feather, require the participant to first grasp the instrument (the feather) before using it to perform the action on the Theme (the frog). Thus attention for action needs to be directed to potential instruments. As a result, on hearing a Theme-instrument verb such as *tickle*, the participant's attention for action is already biased toward an instrument, even before encountering the definite noun phrase. In contrast, for Theme-Goal verbs such as *put*, attention for action will be directed toward the Theme. Thus the combination of the verb bias, preposition bias, and attention-for-action bias may conspire against the referential constraint in the *tickle-the-frog* studies. In sum, then, the strength of the referential-context effects in the *put-the-apple* and the *tickle-the-frog* studies interacts with other constraints, including lexically based verb preferences and frequency-based sense biases for prepositions. Further, visual-world parsing studies, including those involving actions, are sensitive to the lexical preferences that have been documented with other paradigms. Crucially, the differences between the Britt (1994) results and those of Tanenhaus et al. (1995), Spivey et al. (2002), and Trueswell et al. (1999) cannot be due to exaggerated context effects in visual-world tasks, in which sensitivity to well-established linguistic variables was somehow masked: Snedeker, Thorpe, and Trueswell (2001) showed sensitivity to these variables in such a task. Rather, the different findings arise precisely because both measures (reading fixations and visual-world fixations) are sensitive to subtle linguistic properties, pertaining to lexical information and lexical biases, that differed across these studies.

Bridging the Action and Product Traditions

We conclude by first briefly reviewing some ongoing work that we and our colleagues have been pursuing that uses eye gaze to bridge the product and action traditions and then outlining some future challenges.

The availability of eye gaze as a real-time response measure that can be used with nonlinguistic contexts and natural tasks makes it possible to more strongly integrate action-based constructs into product-based experimental designs. One example is the line of research initiated by Chambers and colleagues on when and how actions, intentions, and affordances affect the context with respect to which an utterance is processed. Consider, for example, a variant of the *put-the-apple* study with a two-referent context in which only one of the two potential referents is compatible with the action. In a context that includes a liquid egg in a bowl and a hard-boiled egg in a cup, the instruction *Pour the egg in the bowl over the flour* contains a temporarily ambiguous prepositional phrase (*in the bowl*) with two potential referents (two eggs) that are consistent with a context-independent sense of *the bowl*. In these circumstances, listeners are temporarily garden-pathed, mistakenly interpreting the prepositional phrase as introducing the Goal argument, just as they would in a one-referent context. This result, along with related findings, demonstrates that syntactic ambiguity is resolved with respect to a dynamically updated referential domain that takes into account plausible actions and intention-relevant affordances of objects (Chambers 2001; Chambers, Tanenhaus, and Magnuson, forthcoming). A second example is research by Arnold and colleagues on disfluency and reference resolution. There is growing interest in how characteristics of natural utterances, such as disfluent productions, affect real-time language processing (Brennan and Schober 2001; Bailey and Ferreira, chapter 14, this volume). Using the paradigm of Allopenna and colleagues (1998), Dahan, Tanenhaus, and Chambers (2002) showed that an accented definite noun phrase is preferentially interpreted as referring to a discourse-old entity that is not in focus, rather than to a new entity. Arnold noted that a disfluent production is more likely to occur when a speaker is introducing a new discourse entity than when a speaker is mentioning a given entity. Using the paradigm of Allopenna and associates (1998) to examine looks to cohort competitors, Arnold and colleagues showed that a disfluent noun phrase modulates, and sometimes reverses, the given bias for accented noun phrases (e.g., Arnold, Fagnano, and Tanenhaus 2003). An important unresolved question is whether the real-time effects of disfluency reflect learned contingencies based on statistical correlations among types of forms or whether the listener's attributions of plausible sources of the disfluency also modulate the effects. For instance, would the new bias created by a disfluency disappear if difficulty in lexical retrieval was no longer a plausible source of the disfluency, if, for example, the speaker was distracted by an external noise?

The question of how a listener's attributions about the speaker might affect the processing of a disfluent utterance raises perhaps the most hotly contested issue in current work that bridges the product and action traditions: to what extent do speakers and listeners compute common ground? Most work in the action tradition assumes that participants in a conversation monitor each other's intentions, including making distinctions between speaker and hearer knowledge. It is also frequently assumed that the speaker crafts his or her message with the listener in mind, including sending fine-grained signals about upcoming difficulty in production and choosing forms to limit ambiguity. However, an emerging literature suggests that speakers do not avoid constructions that are ambiguous or otherwise difficult for listeners (Arnold et al. 2004; Brown and Dell 1986; Ferreira and Dell 2000), though under some circumstances speakers use prosody to disambiguate an otherwise ambiguous utterance (Snedeker and Trueswell 2004).

More controversially, Keysar and colleagues (e.g., Keysar et al. 2000; Keysar and Barr, chapter 3, this volume) have presented evidence that listeners initially ignore salient aspects of common ground, such as visual copresence—an important heuristic for common ground, identified by Clark and Marshall (1981). The strongest form of the Keysar proposal was that listeners' initial interpretations are computed egocentrically, with speaker knowledge consulted only in a second stage if a misunderstanding arises (Keysar, Barr, and Horton 1998).

More recent work has qualified that conclusion. While common ground does not completely circumscribe the listeners' referential domain, it does affect even the earliest moments of reference resolution (Arnold, Trueswell, and Lawentmann 1999; Hanna and Tanenhaus, chapter 5, this volume; Hanna, Tanenhaus, and Trueswell 2003; Keysar and Barr, chapter 3, this volume). Moreover, even young children engaged in this task show a similar time course of consideration of common ground (Nadig and Sedivy 2002). Use of eye gaze has been crucial for evaluating when information about common ground is used, and we believe it will be increasingly important in linking the literature on use of common ground with the parallel literature in theory of mind, and its development (see Nadig and Sedivy 2002; Sabbagh and Baldwin 2001).

Thus far work on common ground in real-time processing has used tasks in which one of the participants is a confederate, following a script. Although this class of studies has contributed and will continue to contribute important insights, use of confederates eliminates one of the crucial ingredients of spontaneous interactive conversation: interlocutors cooperating to create a relevant discourse, in a task with joint goals. In our opinion, the question of how speakers and addressees coordinate with one another cannot be satisfactorily answered until we can monitor real-time comprehension in nonscripted interactive conversation. Doing so raises a difficult methodological challenge because traditional psycholinguistic experiments use carefully controlled stimuli. However, Brown-Schmidt and her colleagues (e.g., Brown-Schmidt, Campana, and Tanenhaus, chapter 6, this volume) have demonstrated that it is possible to use eye gaze to monitor real-time comprehension in natural interactive conversation. The initial studies replicate some effects observed in more controlled experiments, but also shed light on how common goal structures can result in closely aligned referential domains. We anticipate that extension of this line of research to more complex goal structures as well as to face-to-face interactive conversation is likely to shed light on when and how interlocutors achieve coordination. This work also parallels the environments in which children are first exposed to words, and we expect that developmental parallels will shed light on long-standing issues in language acquisition (Snedeker and Trueswell 2004; Trueswell and Gleitman, forthcoming).

As real-time work on interactive conversation develops, we are hopeful that psycholinguistic work can complement, and be complementary to, work on intelligent communicative systems that make use of spoken language (Allen et al. 2001). Dialogue systems are beginning to tackle the problem of incremental or continuous generation and understanding in domains that involve interactive conversation with a human user. Because these systems must integrate knowledge of a domain with language processing, they offer the potential for providing a theoretical test bed for explicit computational models of dialogue. We believe that such models will be necessary if psycholinguistic research on dialogue is to seriously explore interactive conversation within an explicit theoretical framework. We also anticipate that eye movements will play an important methodological role in this research.

We close by noting that eye-movement measures need not and should not be the only measures used by researchers to map the time course of processes in conversation. We expect that other methods will emerge that meet many or all of the desiderata sketched earlier in this chapter. We strongly suspect, though, that the most groundbreaking work will come from those using increasingly rich (and complex) data arrays to understand the dynamics of comprehension and production in conversation. 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. This movement toward connecting language and action in rich goal-directed tasks is likely to influence theoretical developments in natural language, just as it has begun to enrich theories of perception and cognition (Ballard et al. 1997; Barsalou 1999).

Notes

This research was partially supported by NIH grants HD-27206 and DC-05071 to MKT and NIH grant HD-37507 to JCT.

1. Studies with action-based tasks increasingly use video-based trackers that monitor the pupil and the cornea, with independent tracking of the head or compensation for head movement, when stimuli are presented on a screen. Measuring head movement can be bypassed by superimposing fixations on a head-based videorecord, though this limits analysis to videorates (60 Hz).

2. By using the word *attention*, we do not intend to suggest that participants are consciously shifting attention. In fact, people are typically unaware of making eye movements and even of exactly where they are fixating. One possibility is that the attentional shifts take place at the level of unconscious visual routines that support accessing information from a visual scene (Hayhoe 2000).

References

Allen, J. F., Byron, D. K., Dzikovska, M., Ferguson, G., Galescu, L., and Stent, A. 2001. Towards conversational human-computer interaction. *AI Magazine*, *22*, 27–35.

Allopenna, P. D., Magnuson, J. S., and 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., and Kamide, Y. 1999. Incremental interpretation of verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247–264.

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

Arnold, J., Fagnano, M., and Tanenhaus, M. K. 2003. Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, *32*, 25–36.

Arnold, J., Trueswell, J. C., and Lawentmann, S. M. 1999, November. Using common ground to resolve referential ambiguity. Paper presented at the Fortieth Annual Meeting of the Psychonomic Society, Los Angeles.

Arnold, J., Wasow, T., Asudeh, A., and Alrenga, P. Avoiding attachment ambiguities: The role of constituent ordering. Unpublished manuscript.

Austin, J. L. 1962. *How to Do Things with Words*. Oxford: Oxford University Press.

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

Barsalou, L. W. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences*, *22(4)*, 577–660.

Boland, J. E. 2002. Listeners use verb argument structure to focus visual attention on potential arguments. Paper presented at the 15th annual meeting of the CUNY Sentence Processing Conference, New York, March 2002.

Brennan, S. E., and Schober, M. F. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, *44*, 274–296.

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, P. M., and Dell, G. S. 1986. Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, *19*, 441–472.

Chambers, C. G. 2001. The Dynamic Construction of Referential Domains. Unpublished doctoral dissertation, University of Rochester.

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

Chambers, C. G., Tanenhaus, M. K., and Magnuson, J. S. Forthcoming. Action-based affordances and syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

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. 1997. Dogmas of understanding. *Discourse Processes*, *23*, 567–598.

Clark, H. H., and Brennan, S. E. 1989. Grounding in communication. In L. Resnick, J. Levine, and S. Teasley, eds., *Perspectives on Socially Shared Cognition*, 127–149. Washington, DC: American Psychological Association.

Clark, H. H., and Carlson, T. 1981. Context for comprehension. In J. Long and A. Baddeley, eds., *Attention and Performance IX*, 313–330. Hillsdale, NJ: Erlbaum.

Clark, H. H., and Marshall, C. R. 1981. Definite reference and mutual knowledge. In A. H. Joshi, B. Webber, and I. A. Sag, eds., *Elements of Discourse Understanding*, 10–63. Cambridge, UK: Cambridge University Press.

Clark, H. H., and Wilkes-Gibbs, D. 1986. Referring as a collaborative process. *Cognition*, *22*, 1–39.

Colby, C. L., and Goldberg, M. E. 1999. Space and attention in parietal cortex. *Annual Review of Neuroscience*, *22*, 97–136.

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., and Steedman, M. 1985. On not being led up the garden path: The use of context by the psychological parser. In D. Dowty, L. Karttunen, and A. Zwicky, eds., *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, 320–358. Cambridge, UK: Cambridge University Press.

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

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

Dahan, D., and Tanenhaus, M. K. 2003. Activation of visually based conceptual representations during spoken-word recognition. Unpublished manuscript.

Dahan, D., and Tanenhaus, M. K. Forthcoming. Continuous mapping from sound to meaning in spoken-language comprehension: Evidence from immediate effects of verb-based constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

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

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., and Tanenhaus, M. K. l995. 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., and McRoberts, G. 1998. Rapid gains in speed of verbal processing by infants in the second year. *Psychological Science*, *9*, 228–231. Reprinted in M. Tomasello and E. Bates, eds., *Language Development: The Essential Readings*. Oxford: Blackwell, 2001.

Ferreira, V., and Dell, G. 2000. Effects of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*, 296–340.

Findlay, J. M. Forthcoming. Eye scanning and visual search. In J. M. Henderson and F. Ferreira, eds., *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. New York: Psychology Press.

Fodor, J. A., Bever, T. G., and Garrett, M. F. 1974. *The Psychology of Language: An Introduction to Psycholinguistics and Generative Grammar*, 313–372. New York: McGraw-Hill.

Grice, H. P. 1957. Meaning. *Philosophical Review*, *66*, 377–388.

Hanna, J. E., Tanenhaus, M. K., and 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. 2000. Vision using routines: A functional account of vision. *Visual Cognition*, *7*, 43–64.

Henderson, J. M., and Ferreira, F. 2003. *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. New York: Psychology Press.

Henderson, J. M., and Hollingworth, A. 2003. Eye movements, visual memory, and scene representation. In M. A. Peterson and G. Rhodes, eds., *Analytic and Holistic Processes in the Perception of Faces, Objects, and Scenes*, 356–383. New York: Oxford University Press.

Horton, W. S., and Keysar, B. 1995. When do speakers take into account common ground? *Cognition*, *59*, 91–117.

Keysar, B., Barr, D. J., Balin, J. A., and Brauner, J. S. 2000. Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*, 32–37.

Keysar, B., Barr, D. J., and Horton, W. S. 1998. The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science*, *7*, 46–50.

Kowler, E. 1995. Eye movements. In S. M. Kosslyn and D. N. Osherson, eds., *An Invitation to Cognitive Science, Volume 2: Visual Cognition*, 2nd ed., 215–265. Cambridge, MA: MIT Press.

Kowler, E. 1999. Eye movements and visual attention. In R. A. Wilson and F. C. Keil, eds., *The MIT Encyclopedia of the Cognitive Sciences*, 306–309. Cambridge, MA: MIT Press.

Krauss, R. M., and Weinheimer, S. 1966. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, *4*, 343–346.

Levelt, W. J. M., Roelof, A., and Meyer, A. S. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

Levinson, S. C. 2000. *Presumptive Meanings*. Cambridge, MA: MIT Press.

Liversedge, S., and Findlay, J. 2001. Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, *4*, 6–14.

Luce, R. D. 1959. *Individual Choice Behavior*. New York: Wiley.

Luce, P., and Pisoni, D. 1998. Recognizing spoken words: The Neighborhood Activation Model. *Ear and Hearing*, *19*, 1–38.

MacDonald, M. C. 1994. Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, *9*, 157–201.

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676–703.

Magnuson, J. S. 2001. The Microstructure of Spoken Word Recognition. Unpublished doctoral dissertation, University of Rochester.

Magnuson, J. S., Tanenhaus, M. K., and Aslin, R. 2003. Time course of spoken word recognition: effects of frequency, cohort density and lexical neighbors. Unpublished manuscript.

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

Marcus, M. P. 1980. *A Theory of Syntactic Recognition for Natural Language*. Cambridge, MA: MIT Press.

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

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

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

Marslen-Wilson, W. D. 1990. Activation, competition, and frequency in lexical access. In G. Altmann, ed., *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*. Cambridge, MA: MIT Press.

Marslen-Wilson, W. D. 1993. Issues of process and representation in lexical access. In G. Altmann and R. Shillcock, eds., *Cognitive Models of Language Processes: The Second Sperlonga Meeting*. Hove, England: Erlbaum.

Matin, E., Shao, K. C., and Boff, K. R. 1993. *Information-processing time with and without saccades. Perception and Psychophysics*, *53(4)*, 372–380.

McClelland, J. L., and Elman, J. L. 1986. Interactive processes in speech perception: The TRACE Model. In D. E. Rumelhart and J. L. McClelland, eds., *Parallel Distributed Processing*, vol. 2. Cambridge, MA: MIT Press.

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

Miller, G. A., and Chomsky, N. 1963. Finitary models of language users. In R. D. Luce, R. R. Bush, and E. Galanter, eds., *Handbook of Mathematical Psychology*. New York: Wiley.

Nadig, A. S., and Sedivy, J. C. 2002. Evidence of perspective-taking constraints in children's online reference resolution. *Psychological Science*, *13*, 329–336.

Pickering, M., and Garrod, S. A. Forthcoming. Towards a mechanistic psycholinguistics of dialogue. *Brain and Behavioral Sciences*.

Rayner, K. 1998. Eye movement in reading and information processing: 20 years of research. *Psychological Bulletin*, *124(3)*, 372–422.

Sabbagh, M. A., and Baldwin, D. A. 2001. Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. *Child Development*, *72*, 1054–1070.

Schegloff, E. A., and Sacks, H. 1973. Opening up closings. *Semiotica*, *8*, 289–327.

Searle, J. R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, UK: Cambridge University Press.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., and Carlson, G. N. 1999. Achieving incremental processing through contextual representation: Evidence from the processing of adjectives. *Cognition*, *71*, 109–147.

Snedeker, J., Thorpe, K., and Trueswell, J. 2001. On choosing the parse with the scene: The role of visual context and verb bias in ambiguity resolution. *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, 964–969. Hillsdale, NJ: Erlbaum.

Snedeker, J., and Trueswell, J. C. 2003. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, *48(1)*, 103–130.

Snedeker, J., and Trueswell, J. C. 2004. The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in children's sentence processing. *Cognitive Psychology*. Forthcoming.

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

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

Swingley, D., and Fernald, A. 2002. Recognition of words referring to present and absent objects by 24-month-olds. *Journal of Memory and Language*, *46*, 39–56.

Swingley, D., Pinto, J. P., and Fernald, A. 1998. Assessing the speed and accuracy of word recognition in infants. *Advances in Infancy Research*, *12*, 257–277.

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

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

Tanenhaus, M. K., Spivey-Knowlton, M. J., and Hanna, J. E. 2000. Modeling thematic and discourse context effects on syntactic ambiguity resolution within a multiple constraints framework: Implications for the architecture of the language processing system. In M. Pickering, C. Clifton, and M. Crocker, eds., *Architecture and Mechanisms of the Language Processing System*, 90–118. Cambridge, UK: Cambridge University Press.

Tanenhaus, M. K., and Trueswell, J. C. 1995. Sentence comprehension. In J. Miller and P. Eimas, eds., *Speech, Language, and Communication*, 217–262. San Diego, CA: Academic Press.

Trueswell, J. C., and Gleitman, L. Forthcoming. Children's eye movements during listening: Developmental evidence for a constraint-based theory of sentence processing. In J. M. Henderson and F. Ferreira, eds., *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. New York: Psychology Press.

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

Trueswell, J. C., and Tanenhaus, M. K. 1994. Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In C. Clifton, K. Rayner, and L. Frazier, eds., *Perspectives on Sentence Processing*. Hillsdale, NJ: Erlbaum.