

Eye Movements as a Window into Real-Time Spoken Language Comprehension in Natural Contexts

**Kathleen M. Eberhard,^{1,2} Michael J. Spivey-Knowlton,¹
Julie C. Sedivy,¹ and Michael K. Tanenhaus¹**

Accepted August 11, 1995

When listeners follow spoken instructions to manipulate real objects, their eye movements to the objects are closely time locked to the referring words. We review five experiments showing that this time-locked characteristic of eye movements provides a detailed profile of the processes that underlie real-time spoken language comprehension. Together, the first four experiments showed that listeners immediately integrated lexical, sublexical, and prosodic information in the spoken input with information from the visual context to reduce the set of referents to the intended one. The fifth experiment demonstrated that a visual referential context affected the initial structuring of the linguistic input, eliminating even strong syntactic preferences that result in clear garden paths when the referential context is introduced linguistically. We argue that context affected the earliest moments of language processing because it was highly accessible and relevant to the behavioral goals of the listener.

We thank D. Ballard and M. Hayhoe for the use of their laboratory (National Resource Laboratory for the Study of Brain and Behavior). We also thank J. Pelz for his assistance in learning how to use the equipment and K. Kobashi for assisting in the data collection. Finally, we thank Janet Nicol and an anonymous reviewer for their comments and suggestions. The research was supported by NIH resource grant 1-P41-RR09283; NIH HD27206 (M.K.T); an NSF graduate fellowship (M.J.S-K.); and a Canadian SSHRC fellowship (J.C.S.).

¹ University of Rochester, Rochester, New York 14627; K. M. Eberhard, M. J. Spivey-Knowlton; M. K. Tanenhaus, Department of Brain and Cognitive Sciences; J. C. Sedivy, Department of Linguistics.

² Address all correspondence concerning this article to Kathleen M. Eberhard, Department of Brain and Cognitive Sciences, Meliora Hall, University of Rochester, Rochester, New York 14627.

The interpretation of an utterance crucially depends on the discourse context in which it occurs. Consider, for example, the sentence *He is putting the ball in the box on the shelf*. Given just the knowledge of the English language, a listener would know that when the sentence was uttered, a male person (or animal) was moving a spherical object either from a box to a shelf or from some unknown place to a box which was located on a shelf. To arrive at the complete and intended interpretation of the sentence, the listener requires knowledge of the situational context of the sentence, i.e., knowledge of the referents and their spatial relations at the particular time of the utterance.

Although all models of language comprehension acknowledge the importance of discourse context in determining the interpretation of a sentence or an utterance, they differ in their assumptions about when context exerts its influence (for recent reviews, see MacDonald, Pearlmutter, & Seidenberg, 1994; Spivey-Knowlton, Trueswell, & Tanenhaus, 1993; Tanenhaus & Trueswell, 1995). For example, some models postulate an initial stage of processing in which the rapid and automatic initial processes that structure the linguistic input are encapsulated from the slower integrative processes that relate an utterance to its discourse context (e.g., Frazier, 1978, 1987; Swinney & Osterhout, 1990). In contrast, other models have emphasized that the linguistic input is immediately mapped onto a discourse representation, with context influencing even the earliest moments of linguistic processing (Altmann & Steedman, 1988; Bates & MacWhinney, 1989; Crain & Steedman, 1985; MacDonald et al., 1994; Marslen-Wilson & Tyler, 1987; Spivey-Knowlton et al., 1993; Taraban & McClelland, 1988, 1990).

Most of the research investigating the role of context in sentence processing has focused on the comprehension of written sentences in discourse contexts that typically consist of only a few sentences describing an imaginary situation. One reason for the focus on written comprehension is that testing subtle predictions about the time course of processing requires using methodologies that can provide immediate information about how each word is interpreted as the sentence unfolds. There are a variety of methodologies for studying reading that provide a continuous processing profile, chief among them being self-paced reading and eye-movement monitoring tasks which have the advantage of allowing the subject to process the input in a relatively natural way, i.e., without requiring any explicit decisions.

The situation is different for spoken language comprehension. Although there are a variety of on-line methodologies that provide insight into the temporal characteristics of spoken language comprehension, they do not allow for continuous monitoring. In addition, most of these methodologies measure comprehension indirectly via a superimposed secondary task, e.g., detection, cross-modal priming, or probe tasks. And methodologies that do

provide a more direct measurement, e.g., monitoring tasks, require listeners to consciously attend to the linguistic input in an unnatural way. As a result, these paradigms cannot easily be used to study immediate spoken language comprehension as it typically occurs in natural real-world situations, for example in interactive conversation where the objects and events being referred to are in the immediate environment and therefore are highly salient.

In this article, we present an overview of some work involving a new methodology that allows an in-depth investigation of incremental processing and contextual dependence in spoken language comprehension. The methodology involves monitoring listeners' eye movements as they follow short discourses instructing them to move or touch common objects in a display (e.g., "Put the candy above the fork. Now put it below the pencil"), thus providing an on-line, nonintrusive measure of spoken language comprehension as it occurs in natural situational contexts. Eye movements are monitored using a light-weight eye-tracking camera that is mounted on a helmet worn by the listeners. Also mounted on the helmet is a small video camera that records the visual scene from the listener's perspective. The visual scene image is displayed on a TV monitor along with a record of the listener's eye fixations superimposed as cross hairs.³ Both the image and the experimenter's spoken instructions are synchronously recorded by a VCR which permits frame-by-frame playback.

Two essential features of this paradigm make it useful for studying spoken language comprehension in context. First, the visual context is *available* for the subject to interrogate as the spoken language message unfolds over time, and because the message directs the listener to interact with the context, the context is necessarily *relevant* to the comprehension process. This contrasts with a linguistically introduced context, which must be represented in memory and depending upon the experimental task, may or may not be immediately accessible or perceived as relevant by the reader or listener. Secondly, in all of the work we have conducted to date, we have found that subjects' eye movements to objects are closely time locked to the spoken words that refer to those objects (Sedivy, Tanenhaus, Spivey-Knowlton, Eberhard, & Carlson, 1995; Spivey-Knowlton, Tanenhaus, Eberhard, & Sedivy, 1995; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995, in press). Thus the methodology provides a natural on-line measure of how comprehension unfolds over time, and how it is influenced by the

³ Eye movements were monitored by an Applied Scientific Laboratories eyetracker which provides an infrared image of the eye at 60 Hz. This allows the tracking of the center of the pupil and the corneal reflection of one eye to be recorded every 16 msec. The accuracy of the tracking is about a degree over a range of $\pm 20^\circ$. A short calibration routine is conducted before each experimental session to map the eye-in-head coordinates from the tracker to the visual scene image coordinates.

information provided by the visual context. The work that we have conducted demonstrates that the pattern and timing of these eye movements allows for strong inferences about component processes in comprehension, including referential processing, word recognition, and parsing.

We will discuss a total of five experiments. The first two experiments lay the foundation for the other three by clearly demonstrating the incremental nature of spoken language comprehension: Listeners interpreted each word of an utterance immediately with respect to the set of co-present visual referents. More specifically, they used information from each word to reduce the set of possible visual referents to the intended one. As a result, they established reference as soon as the utterance provided enough accumulative information for them to distinguish or disambiguate the intended referent from the alternatives. In all the experiments, the visual context was manipulated to vary the point in the utterance when information for uniquely identifying an intended referent occurred. This manipulation allowed us to examine the effects of the visual context on the time course of comprehension.

In the third experiment, we manipulated the prosody of the utterance as well as the visual context. We found that listeners made immediate use of disambiguating information provided by contrastive stress, as revealed by its facilitatory effect on the point in a complex noun phrase when listeners established reference. We also discuss the results of a fourth experiment showing that the point at which reference was established within a word (e.g., candle) was influenced by whether or not the relevant visual context contained an object with a similar name (e.g., both a candle and some candy).

Finally, we present evidence that a visually co-present referential context influenced initial syntactic commitments under conditions where linguistically introduced contexts have been shown to be ineffective. The referential effects we report are consistent with Crain and Steedman's (1985) and Altmann and Steedman's (1988) claims about the importance of referential pragmatic context in ambiguity resolution, and they provide strong evidence against modular models in which syntactic commitments are made during an encapsulated first stage of processing. More generally, we argue that these effects follow naturally from the behavioral goals and Gricean expectations of the listener. We also argue that the global pattern of results across visual and linguistic contexts is most naturally accommodated by constraint-based models of language processing.

THE INCREMENTAL NATURE OF ESTABLISHING REFERENCE

For communication to be successful, a listener must arrive at the intended referents of the words of a speaker's utterance. Intuitively, the es-

tablishment of reference seems to be incremental: As soon as a word is heard, its meaning becomes available and is interpreted with respect to the entities in the discourse model. However, from a linguistic perspective, we often talk as though the interpretation of words does not occur until the end of a phrasal constituent. Consider for example the definite noun phrase *the large beach ball*. We say that the adjectives *large* and *beach* modify the noun *ball* because they provide additional relevant information to the noun. This linguistic perspective implies that the referent of a phrase like *the large beach ball* cannot be understood until the noun *ball*. While this implication may be reasonable when considering decontextualized linguistic expressions, it clashes with our intuition about expressions spoken under normal circumstances, i.e., expressions spoken in context. Words uttered in context do not refer to or modify other words, they refer to or modify entities in the discourse model. Olson (1970) further elaborated this point in the following statement: "words do not 'mean' referents or stand for referents, they have a use—they specify perceived events relative to a set of alternatives; they provide information" (p. 263).

Olson illustrated his claim by pointing out that the expression that is used to refer to a particular object depends on the context in which the object occurs. For example, the definite expression *the ball* may be used to refer to one of the objects in Fig. 1a, but a different definite expression, e.g., *the beach ball*, must be used to refer to the same object when it occurs in the context of Fig. 1b. And yet another definite expression must be used when the object occurs in the context of Fig. 1c, e.g., *the large beach ball*. Speakers use different expressions to refer to the same object in different contexts in order to provide their listeners with the necessary information for distinguishing the intended referent from the set of alternatives. Listeners expect speakers to walk a middle line between providing too little information for distinguishing the referent and providing too much information (Grice, 1975). An important factor affecting a speaker's ability to walk this line is the extent to which his or her set of relevant discourse referents is the same as the listener's (Clark, 1992), i.e., the extent to which the information that the speaker intends the listener to consult in understanding the utterance is the information the listener actually consults. The speaker is apt to be most successful at providing the right information when the relevant set is visually co-present and therefore is highly accessible and salient to both the speaker and the listener (Clark, 1992).

From the listener's perspective, when the relevant set of referents is visible and co-present with the speaker's utterance, the listener should be able to immediately interpret each word of the utterance with respect to the

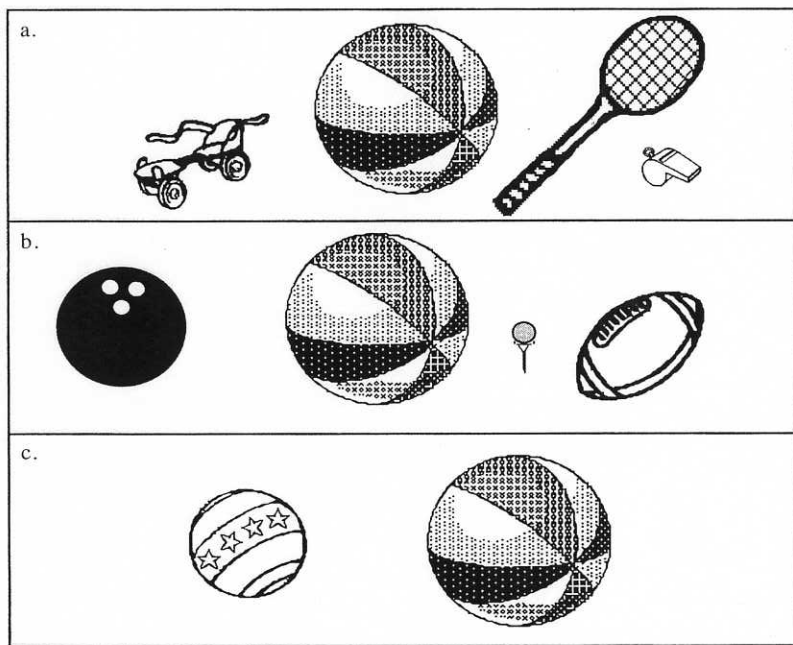


Fig. 1. The noun phrase, *the ball*, may be used to refer to one of the objects in Fig. 1a, but a different noun phrase, e.g., *the beach ball*, must be used to refer to the same object when it occurs in the context of Fig. 1b. And yet another noun phrase must be used when the object occurs in the context of Fig. 1c, e.g., *the large beach ball*.

set of referents.⁴ For example, if a listener is told to *kick the large beach ball* in a co-present visual context depicted in Fig. 1c., where there are two beach balls but only one that is large, he or she may be able to identify the referent as soon as the adjective *large* is heard. This general prediction was investigated using the head-mounted eye-tracker paradigm (Tanenhaus et al., in press). We reasoned that because the paradigm allows us to observe the listeners' eye movements to the referent objects as the spoken input unfolds, it would provide important insight into not only how but also when listeners establish reference.

THE INCREMENTAL PROCESSING OF COMPLEX SPOKEN NOUN PHRASES

In the first experiment (see Tanenhaus et al., in press), subjects were given spoken instructions to touch various blocks arranged in simple dis-

⁴ For a discussion of the relation between co-presence and mutual knowledge and their effect on the establishment of definite reference see Clark and Marshall (1992).

plays like those depicted in Fig. 2. The blocks differed along the dimensions of marking, color, and shape, and the spoken noun phrases that referred to the blocks specified each of those dimensions. For each display, the experimenter read aloud from a script a critical instruction like "Touch the starred yellow square" and several filler instructions (e.g., "Touch the plain red square. Now touch the plain blue square. Now touch it again"). Every display contained a centrally located fixation cross, and the subjects were instructed at the beginning of the experiment to look at the cross and rest their hands in their lap after they perform each requested action.

Critical instructions were given in three display conditions. The conditions determined the point in the instructions when disambiguating information occurred. Disambiguating information was provided by the marking adjective in the early condition, by the color adjective in the mid condition, and by the noun in the late condition. We predicted that if subjects interpreted the words in a noun phrase incrementally with respect to the relevant set of referents in the display, their eye movements to the target objects should occur shortly after the word that disambiguated the intended referent from the alternatives, rather than after the noun. In addition, we expected that the timing of the eye movements relative to the onset of the disambiguating words would reveal the speed with which nonlinguistic visual information was integrated with the spoken linguistic input.

The data were analyzed using frame-by-frame playback of the synchronized video and audio recordings of the experimental sessions. Eye-movement latencies to target objects were obtained by locating the frame on which a critical word in an instruction began and then counting the number of frames until the eye position, which was indicated by the crosshairs, moved from the central cross in the display to the target object. Each audio-video frame consisted of a 33-msec segment of time.

Figure 3 contains a graph of the mean eye-movement latencies to the target blocks measured from the onset of the spoken determiner *the* in each

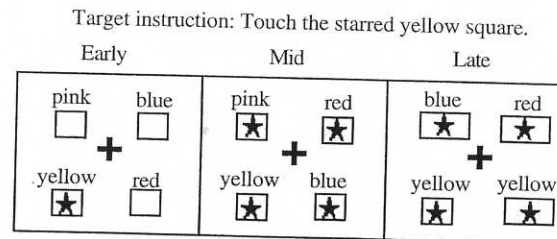


Fig. 2. Example displays representing the three point-of-disambiguation conditions in Experiment 1.

