Eye Movements Accompanying Language and Action in a Visual Context: Evidence Against Modularity

Michael Spivey-Knowlton
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627
spivey@psych.rochester.edu

Michael Tanenhaus
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627
mtan@psych.rochester.edu

Kathleen Eberhard
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627
eberhard@psych.rochester.edu

Julie Sedivy
Department of Linguistics
University of Rochester
Rochester, NY 14627
sedivy@psych.rochester.edu

Abstract

It is commonly assumed that as a spoken linguistic message unfolds over time, it is initially processed by modules that are encapsulated from information provided by other perceptual and cognitive systems. We were able to observe the effects of relevant visual context on the rapid mental processes that accompany spoken language comprehension by recording eye movements using a head-mounted eye-tracking system while subjects followed instructions to manipulate real objects. Under conditions that approximate an ordinary language environment, incorporating goal-directed action, the visual context influenced spoken word recognition and mediated syntactic processing, even during the earliest moments of language processing.

Introduction

It is often claimed that early stages of language comprehension are comprised of informationally-encapsulated modules devoted to processing particular sub-domains of the linguistic input without influence from other perceptual and cognitive systems (e.g., Ferreira & Clifton, 1986; Fodor, 1983). In contrast, constraint-based approaches assume that "correlated constraints" from various information sources are immediately integrated during the processing of linguistic input (MacDonald, Pearlmutter & Seidenberg, 1994; McClelland, 1987; Spivey-Knowlton & Sedivy, in press; Tanenhaus & Trueswell, in press). The temporary ambiguities that arise because language unfolds over time have provided the primary empirical testing ground for evaluating these contrasting theoretical perspectives, with the strongest evidence for information encapsulation coming from studies in which potentially relevant constraints that are introduced by a prior linguistic context appear to have delayed effects (cf. Ferreira & Clifton, 1986; Rayner, Garrod & Perfetti, 1992). However, the "context" in such studies is frequently rather impoverished, consisting of a few sentences that precede the target sentence. In addition, the subject's task is often vague, with no well-defined behavioral goal. Under these conditions, the context may not be perceived as relevant by the comprehender, and even if it is, it must be stored in memory, thus it may not be immediately accessible when the ambiguity is first encountered. Moreover, because the context is introduced linguistically, it is always possible to preserve modularity by expanding the scope of the linguistic module.

The research presented here explores temporary ambiguity resolution during the comprehension of spoken language under conditions in which: 1) there is a strong test of modularity because the context comes from a completely different perceptual modality: vision, 2) the context is immediately relevant because an action must be carried out that directly relates the utterance to the context, and 3) the context, because it is visual, is *co-present* with the linguistic input, and thus can be interrogated when the ambiguity is first encountered. We make use of an experimental paradigm we have recently developed in which the listener follows spoken instructions to manipulate real objects in a display, while we record their eye movements to those objects, time-locked with corresponding referents in the spoken instruction -- *all under conditions that approximate an ordinary language environment* (Spivey-Knowlton, Sedivy, Eberhard & Tanenhaus, 1994; Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, in press a, in press b). With this methodology, we are able to observe the effects of visual context on early moments of comprehension, and tap into partial incremental interpretations that are often observable only in eye-movement patterns (not in hand movements or subjects' intuitions). We describe experiments on ambiguity resolution *within individual words* (Experiment 1) and *across syntactic relationships* (Experiment 2), that reveal early commitments to interpretations based on partial input, and immediate integration of relevant visual information with linguistic information in both word recognition and syntactic parsing.

Method

Before reaching for an object, people typically move their eyes to fixate it (Ballard, Hayhoe & Pelz, in press; Epelboim, Collewijn, Kowler, Erkelens, Edwards, Pizlo, & Steinman, 1994). Thus, when we instruct a subject to "pick up the candle," she makes a saccadic eye movement to the candle, and our methodology allows us to measure the time elapsed from the beginning of the word "candle" to the initiation of the saccade, as well as record any intermediate fixations to other objects. Eye movements are especially informative about early moments of processing because saccades are relatively automatic (devoid of strategic influences) and almost entirely ballistic. Thus, an initial misinterpretation of the spoken input, from which the listener rapidly recovers, is still observable as a brief fixation of the "incorrect" object.

Eye movements were monitored by an Applied Scientific Laboratories (ASL) eyetracker mounted on top of a lightweight helmet. The camera provides an infrared image of the eye at 60Hz. The center of the pupil and the corneal reflection are tracked to determine the orbit of the eye relative to the head. A scene camera, mounted on the side of the helmet, provides an image of the subject's field of view. Gaze position (indicated by crosshairs) is superimposed over the scene camera image and recorded onto a Hi8 VCR with frame-by-frame playback. Accuracy of the gaze position record is about a degree over a range of +/- 20 deg. The video record was coordinated with the audio record for all data analysis.

Subjects were seated at arm's length from a 3' by 3' table workspace that was divided into 25 squares (see Figure 1), and were given spoken instructions to move everyday objects around. A black cross in the center square served as a neutral fixation point, where the subject's gaze was directed at the onset of an instruction set. No more than one object was placed in any square, so that noise in the gaze position signal was never enough to mistake a fixation of one object for another. The majority of instructions did not involve the experimentally relevant objects.

Experiment 1: Ambiguity Within Words

Background

In a classic set of experiments, Marslen-Wilson and colleagues demonstrated that, to a first approximation, recognition of a word occurs shortly after the auditory input uniquely specifies a lexical candidate. For polysyllabic words, this is often prior to the end of the word. For example, the word "elephant" would be recognized shortly after the "phoneme" /f/. Prior to that, the auditory input would be consistent with the beginnings of several words, including "elephant", "elegant", "eloquent" and "elevator". Thus, recognition of a spoken word is strongly influenced by the words that it is phonetically similar to, especially those words that share initial phonemes. Marslen-Wilson referred to the set of lexical candidates that is activated in the same phonetic environment as a **cohort** (for review, see Marslen-Wilson, 1987).

Evidence from several experimental paradigms indicates that these candidates are partially activated as a word is being processed. For example, cross-modal lexical priming experiments demonstrate that semantic information associated with cohort members is temporarily activated as a word unfolds. The prior context of the utterance and subsequent input provide evidence that is used to evaluate the competing alternatives. While current models differ in how they account for these data, nearly all models incorporate the idea that the time it takes to recognize a word depends on a set of potential lexical candidates. (See Cutler, in press, for a recent review.)

This experiment had two goals. The first goal was to determine how closely time-locked eye movements to a target object would be to the name of the object in a spoken instruction. The second goal was to determine whether the presence of a "competitor" object with a similar name would influence eye-movement latencies to the referred-to object. A visually-mediated "cohort competitor" effect would provide strong evidence that lexically-based information associated with multiple lexical candidates is partially activated during spoken word recognition. In addition, it would demonstrate that relevant visual context affects even the earliest moments of language processing.

Procedure

Eight naive subjects participated in this experiment. They were given instructions to pick up an object and then put it in the square above or below another object. A sample instruction set is given below:

(1) Look at the cross.
Pick up the candle.
Now hold it over the cross.
Now put it above the mouse.

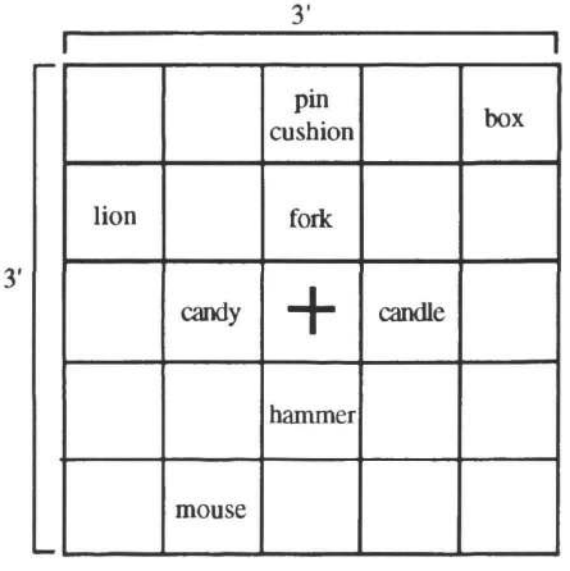

Figure 1. Example display in which both members of the cohort pair are present in the workspace. (The words in this figure indicate locations of actual objects on the table.)

We used four pairs of objects with names that were phonetically similar until late in the word: *candy/candle, car/carton, penny/pencil,* and *doll/dolphin.* Each critical object appeared with its "cohort competitor" on some trials and with only distractor objects on other trials. Each subject was exposed to two of the four cohort pairs. For instructions involving cohort members, the objects were always in one of the central eight squares (excluding the square with the cross).

Figure 1 shows the workspace at the beginning of the instruction set given in example 1. The instructions and the positions of the objects were varied to prevent strategies. In particular, we avoided creating any contingencies that would have resulted in predictable instructions.

Results

On all of the critical trials, the movement of the hand to pick up the target object was preceded by a saccade to that object. On 33% of the trials, fixation of the referred-to object was preceded by a saccade launched to an "incorrect" object. For trials without such "false launches", the mean saccade latency for the "pick-up" instruction was 487ms from the onset of the target word (e.g., "candle"). Saccade latencies were reliably longer when the display contained a "cohort competitor" (530 ms) than when it did not (445 ms); $F(1,7)=9.27$, $p<.02$. The average duration of a target word was 300ms. If we assume that the interval between the onset of programming a saccade and the initiation of the saccade is about 200ms (Matin, Shao & Boff, 1993), then we can estimate that the programming of a saccade to the target object began an average of 55ms before the end of the word in the competitor-absent condition.

More false launches were made when a competitor was present than when it was absent, but this difference was not reliable (37% compared to 29%). Of the false launches in the competitor-present condition, 61% were to the competitor object. In contrast, in the competitor-absent condition, only 25% of the false launches were to the object that occupied the same square as the competitor object in its corresponding competitor-present display. This difference was reliable in an analysis for six subjects; $F(1,5)=14.90$, $p<.02$. (Two subjects did not make any false launches in the competitor-absent condition and thus were excluded from the analysis.)

Discussion

Three critical results emerged from this experiment. First, eye movements to the target object were closely time-locked to the linguistic expression that referred to that object. Thus, the eye movements provide an informative measure of ongoing comprehension. Second, the latency with which the saccades to the target object were launched provides clear evidence that activation of lexical representations begins before the end of a word. The high rate of false launches to competitors lends further support to the idea that multiple lexical candidates are activated early on in recognition. Third, the names of the possible referents in the visual context clearly influenced the speed with which a referent in the speech stream was identified. This demonstrates that the instruction was interpreted incrementally, taking into account the set of relevant objects present in the visual workspace.

Experiment 2: Ambiguity In Syntax

Background

Clearly, the place where the modularity hypothesis has found the most purchase is in the realm of syntactic processing (cf., Ferreira & Clifton, 1986). The strongest evidence for the modularity of syntactic processing has come from studies using sentences with brief syntactic ambiguities in which readers have clear preferences for particular interpretations that persist momentarily *even when preceding linguistic context supports the alternative interpretation* (e.g., Britt, 1994; Ferreira & Clifton, 1986). In this type of experiment, the context is typically comprised of a few sentences preceding the target sentence. However, when the context is a visual display that is immediately relevant to the linguistic input (because an action is expected), and storing the context in memory is unnecessary (because the visual context is co-present with the spoken input), syntactic parsing may indeed show immediate effects of context. If so, this would provide definitive evidence against the encapsulation of syntactic parsing.

Procedure

We used instructions containing the temporary syntactic ambiguity with perhaps the strongest syntactic preference in English, as illustrated by the examples in (2).

(2) a. Put the saltshaker on the envelope in the bowl.
b. Put the saltshaker that's on the envelope in the bowl.

In sentence (2a), the first prepositional phrase (PP), "on the envelope", is ambiguous as to whether it modifies the noun phrase ("the saltshaker") thus specifying the Location of the object to be picked up, or whether it denotes the Goal of the event, i.e. where the saltshaker is to be put. As they are processing this sentence, readers and listeners initially interpret the first prepositional phrase as specifying the Goal, resulting in momentary confusion when they encounter the second preposition ("in"). In example (2b) the word "that's" disambiguates the phrase as a modifier, serving as an unambiguous control condition.

Six naive subjects were presented with six instances of each type of instruction (ambiguous and unambiguous) illustrated in example 2, with a one-referent visual context that supported the Goal interpretation or a two-referent context that supported the Location-based modification interpretation. In the one-referent context for this example, the workspace contained an saltshaker on a envelope, another envelope, a bowl, and an apple. Upon hearing the phrase "the saltshaker", subjects can immediately identify the object to be moved because there is only one saltshaker and thus they are likely to assume that "on the envelope" is specifying the Goal of the putting event. In the two-referent context, however, the apple was replaced by a second

saltshaker which was on a napkin. Thus, "the saltshaker", could refer to either of the two saltshakers and the phrase "on the envelope" provides modifying information that specifies which saltshaker is the correct referent.

Results

Strikingly different fixation patterns between the two visual contexts revealed that the ambiguous phrase "on the envelope" was initially interpreted as a Goal in the one-referent context, but as a modifier in the two-referent context. In the one-referent context, subjects looked at the incorrect Goal (e.g., the irrelevant envelope) on 55% of the trials shortly after hearing the ambiguous PP, whereas they never looked at it during the unambiguous instruction; $t(5)=4.11$, $p<.01$. In contrast, when the context contained two possible referents, subjects rarely looked at the incorrect Goal (17% of the trials), and there was no difference between the ambiguous and unambiguous instructions. The statistical interaction between Context and Ambiguity was reliable; $F(1,5)=8.24$, $p<.05$.

Figures 2 and 3 summarize the most typical sequences of eye movements in the ambiguous and unambiguous instructions for the one-referent and the two-referent contexts. In the one-referent context (Figure 2), subjects first looked at the target object (the saltshaker) 500ms after hearing "saltshaker", then looked at the incorrect Goal (the rightmost envelope) 484ms after hearing "envelope". In contrast, with the unambiguous instruction, the first look to a Goal did not occur until 471ms after the subject heard the word "bowl". Example 3 indicates the approximate timing of saccades with the speech stream via subscripted indices of the eye movements in Figure 2.

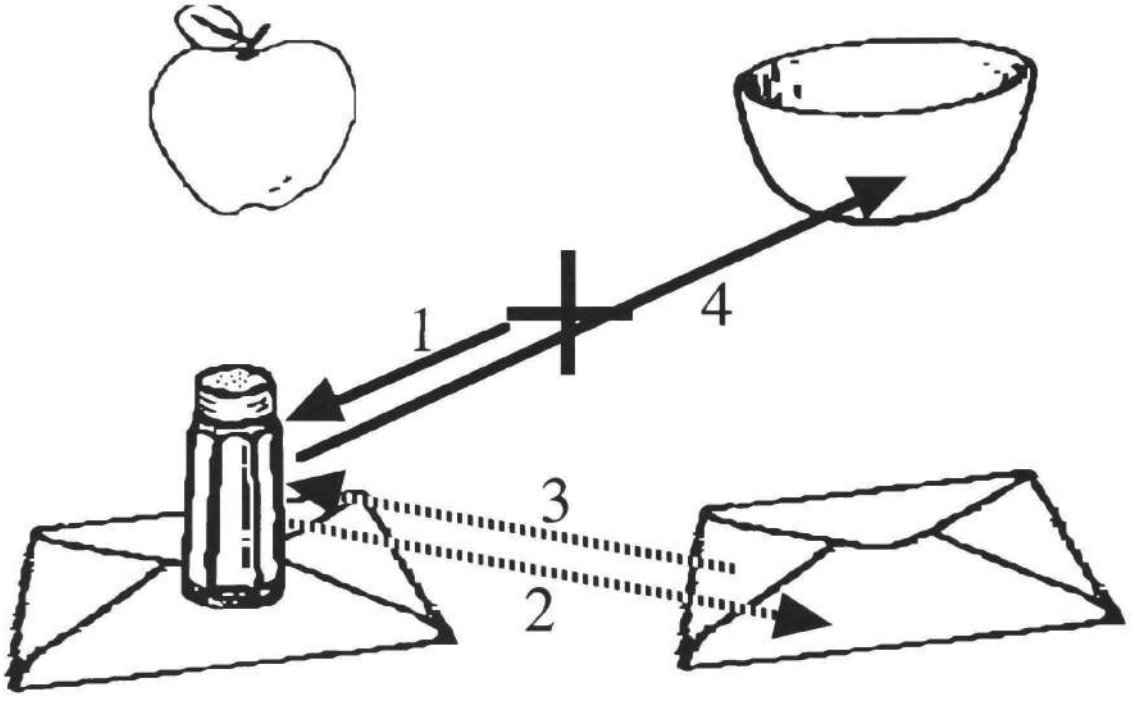

Figure 2. Typical sequence of eye movements in the one-referent context for the ambiguous and unambiguous instructions. Dashed arrows show the intermediate saccades to the incorrect Goal and back to the referent object that occur *only in the ambiguous instruction*. (See examples 3a and 3b for the temporal relationship between eye movements and words in the speech stream.)

(3) a.Put the saltshaker on the$_1$ envelope in$_2$ the bowl.$_3$ $_4$
b.Put the saltshaker that's $_1$on the envelope in the bowl.$_4$

In the two-referent context, subjects often looked at both saltshakers, reflecting the fact that the referent of "the saltshaker" was temporarily ambiguous. Subjects looked at the incorrect object on 42% of the unambiguous trials and on 61% of the ambiguous trials. In contrast, in the one-referent context, subjects rarely looked at the incorrect object (0% and 6% of the trials for the ambiguous and unambiguous instructions, respectively). The three-way interaction among Context, Ambiguity, and Type of Incorrect Eye movement (object vs. Goal) revealed the bias toward a Goal interpretation in the one-referent context and toward a Location-based modification interpretation in the two-referent context; $F(1,5)=18.41$, $p<.01$.

In the two-referent context, the timing of eye movements relative to the speech stream was nearly identical for ambiguous and unambiguous instructions. This indicates that subjects were interpreting the PP ("on the envelope") as an NP-modifier (instead of as a Goal) equally quickly in both ambiguous and unambiguous instructions.

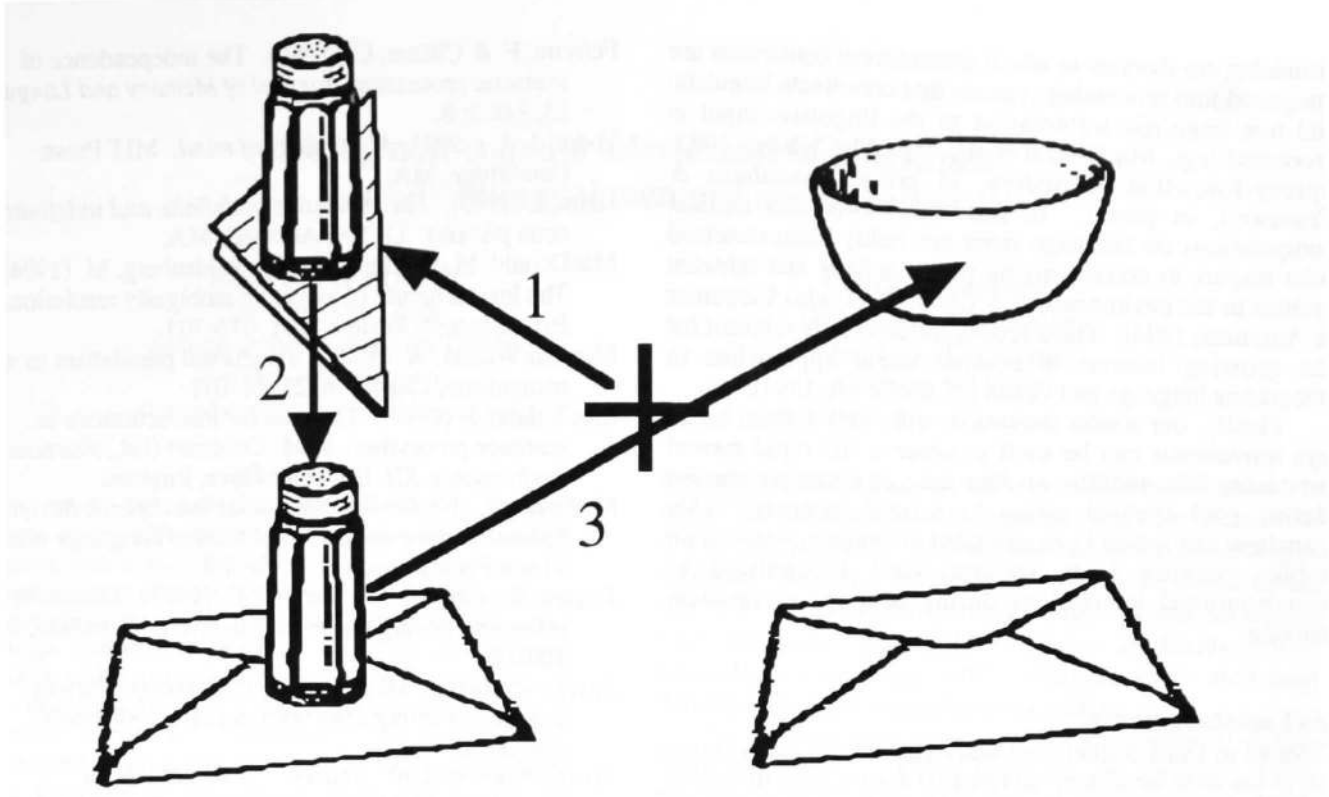

Figure 3. Typical sequence of eye movements in the two-referent context. Note that, in this context, the eye-movement pattern did not differ for the ambiguous and unambiguous instructions. (See subscripted indices in examples 4a and 4b for the temporal relationship between eye movements and words in the speech stream.)

(4)a.Put the saltshaker on the$_1$ envelope $_2$in the bowl.$_3$
b.Put the saltshaker that's $_1$on the envelope $_2$in the bowl.$_3$

In addition to examining the effects of one- versus two-referent contexts on syntactic processing, we examined the effects of a "three-and-one" referent context, i.e., instead of having *one* additional referent, there were *three* additional referents. So, in place of the saltshaker on a napkin in Figure 3, there were *three saltshakers* in that square. We did this to examine whether the presupposition of uniqueness associated with the definite determiner *the* (cf. Heim, 1982) in "Put the saltshaker" would bias the subject toward the lone saltshaker (on the envelope). Indeed, such a bias was observed in saccade latencies to the target saltshaker, which resembled those of the one-referent context. This is because subjects rarely looked at the three saltshakers. However, unlike the one-referent condition, the ambiguous PP "on the envelope" was still interpreted as a modifier instead of a Goal. This is seen in the eye-movement pattern, which resembled the two-referent condition: subjects rarely looked at the incorrect Goal. Thus, with both ambiguous and unambiguous instructions, this 3-and-1 referent context elicited an overall eye-movement pattern and timing similar to that for the *unambiguous* instruction in the one-referent context.

Discussion

It is clear from these results that the relevant aspects of the visual scene influence even the initial moments of syntactic analysis. When an object referred to in the speech stream is unique in the visual input, further specification of it is deemed unnecessary, resulting in a bias toward interpreting an ambiguous PP as describing the event and not the object. In contrast, the visual presence of multiple referents (e.g., two saltshakers) biases the listener toward interpreting the ambiguous PP as describing *which object* is being referred to, instead of *where to put it*.[1] Crucially, this effect of visual context is observed at the earliest measurable point in processing.

General Discussion

Our results demonstrate that, in natural contexts, people seek to establish reference with respect to their intended actions during the earliest moments of linguistic processing. Moreover, referentially relevant non-linguistic information immediately affects how the linguistic input is initially structured. Given these results, approaches to language comprehension that assign a central role to encapsulated linguistic subsystems are unlikely to prove fruitful. More

[1] Crain and Steedman (1985) develop a theory of syntactic ambiguity resolution in which referential context of this sort (but in linguistic form) plays a central role.

promising are theories in which grammatical constraints are integrated into processing systems that coordinate linguistic and non-linguistic information as the linguistic input is processed (e.g., MacDonald et al., 1994; McClelland, 1987; Spivey-Knowlton & Sedivy, in press; Tanenhaus & Trueswell, in press). In this view, even the earliest computations on language input are richly contextualized with respect to accompanying plans, actions and relevant entities in the environment (cf. Clark, 1992; also Carpenter & Alterman, 1994). These results are especially relevant for the growing interest in computational approaches to integrating language and vision (cf. McKevitt, 1994).

Finally, our results show that, with well-defined tasks, eye movements can be used to observe the rapid mental processes that underlie spoken language comprehension during goal-directed action in natural contexts. This paradigm can naturally be extended to explore questions on topics ranging from spoken word recognition to conversational interactions during cooperative problem solving.

Acknowledgments

Thanks to Dana Ballard and Mary Hayhoe for encouraging us to use their laboratory, to Jeff Pelz for teaching us how to use the equipment, and to Kenzo Kobashi for assistance in data collection. Supported by NIH resource grant 1-P41-RR09283; NIH, HD27206, to MKT, an NSF Fellowship to MS-K and a Canadian SSHRC fellowship to JCS.

References

Ballard, D., Hayhoe, M. & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, 7, 68-82.

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.

Carpenter, T. & Alterman, R. (1994). A taxonomy for planned reading. *Proceedings of the 16th Annual Conference of the Cognitive Science Society.*

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

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

Crain, S. & Steedman, M. (1985). On not being led up the garden path. In Dowty, Kartunnen & Zwicky (eds.), Natural Language Parsing. Cambridge U. Press: Cambridge, MA.

Cutler, A. (1995). Spoken word recognition. In J. Miller & P. Eimas (Eds.), *Handbook of Cognition and Perception*. Academic Press.

Epelboim, J., Collewijn, H., Kowler, E., Erkelens, C., Edwards, M., Pizlo, Z., Steinman, R. (1994). Natural oculomotor performance in looking and tapping tasks. *Proceedings of the 16th Annual Conference of the Cognitive Science Society.*

Ferreira, F. & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, 25, 348-368.

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

Heim, I. (1983). *The semantics of definite and indefinite noun phrases.* GLSA: Amherst, MA.

MacDonald, M. Pearlmutter, N & Seidenberg, M. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review* , 101, 676-703.

Marslen-Wilson, W. (1987). Functional parallelism in word recognition. *Cognition*, 25, 71-102.

McClelland, J. (1987). The case for interactionism in sentence processing. in M. Coltheart (Ed.) *Attention & Performance XII* Erlbaum: Hove, England.

McKevitt, P. (1994). (Ed.), *Artificial Intelligence Review, Special Volume on the Integration of Language and Vision Processing*, Vol. 8, No.1-3.

Rayner, K., Garrod, S. & Perfetti, C. (1992). Discourse influences during parsing are delayed. *Cognition*, 45, 109-139.

Spivey-Knowlton, M. & Sedivy, J. (in press). Parsing attachment ambiguities with multiple constraints. *Cognition.*

Spivey-Knowlton, M., Sedivy, J., Eberhard, K. & Tanenhaus, M. (1994). Psycholinguistic study of the interaction between language and vision. In *AAAI-94 Workshop Notes on the Integration of Natural Language and Vision Processing.*

Tanenhaus, M. & Trueswell, J. (1995). Sentence comprehension. In J. Miller & P. Eimas (Eds.), *Handbook of Cognition and Perception*. Academic Press.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K. & Sedivy, J. (in press a). Using eye movements to study spoken language comprehension. In T. Inui & J. McClelland (Eds.), *Attention & Performance XVI: Integration in Perception and Communication.*

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K. & Sedivy, J. (in press b). Integration of visual and linguistic information in spoken language comprehension. *Science.*