Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language 57 (2007) 455–459
Journal of Memory and Language
www.elsevier.com/locate/jml

# Introduction to the special issue on language–vision interactions

Fernanda Ferreira [a,*], Michael K. Tanenhaus [b]

[a] University of Edinburgh, Chair in Language and Cognition, 7 George Square, Psychology, PPLS, Edinburgh, UK EH8 9JZ, USA
[b] University of Rochester

Received 3 August 2007

**Abstract**

Researchers in psycholinguistics are increasingly interested in the question of how linguistic and visual information are integrated during language processing. In part, this trend is attributable to the use of the so-called "visual world paradigm" in psycholinguistics, in which participants look at and sometimes manipulate objects in a visual world as they listen to spoken utterances or generate utterances of their own. In this introductory article to the Special Issue on Language-Vision Interactions, we briefly describe the history of attempts to look at the integration of language and vision, and we preview the articles appearing in the special issue. From those articles, it is clear that recent work has dramatically expanded our understanding of this important question, a trend that will only accelerate as theoretical and methodological advances continue to be made.


Language and vision are the two primary systems available for studying human perception and cognition, including those "central" processes that are involved in all cognitive domains, such as attention, memory, and learning. The two systems often operate in concert, as when we discuss aspects of the world around us. Researchers studying each system often address similar problems, for example, temporary ambiguity, and researchers in language and vision often adopt similar language, for example, focus and salience. One would think, then, that there would be a great deal to learn by studying the interactions between these systems, and that investigations of language and vision would feature prominently in the literature. For the most part, however, language and vision have been studied independently of one another, especially when the target of inquiry is spoken language. (The situation is clearly different in reading, where the study of language processing is inextricably linked to vision.) This has been the case even when studies in one domain make use of response measures from the other. For example, in studies of visual perception, linguistic responses are sometimes treated as a dependent variable, with naming success or naming latency treated as a transparent measure of perceptual success, without consideration of the processes involved in planning and executing an utterance (see Bock, 1996). Likewise, in studies of spoken language processing—the focus of the papers in this special issue—performance on visual tasks has been treated more or less as just a dependent variable. This makes the implausible assumption that fixations and saccades to objects mentioned in speech transparently reflect linguistic processing, and that the problem of understand-

* Corresponding author.
*E-mail address:* fernanda.ferreira@ed.ac.uk (F. Ferreira).

0749-596X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2007.08.002

ing the interface between the two systems could be postponed indefinitely.

There have been some notable exceptions. There is a rich tradition of examining how visual information from the face, especially the lips, combines with auditory information to form phonetic percepts (Massaro, 1998; Massaro, Cohen, & Smeele, 1996; Massaro & Stork, 1998; McGurk & MacDonald, 1976). And, in the early, heady days of the cognitive revolution in the 1960s, psycholinguists explored a variety of tasks designed to provide insights into the representations and processes that enable language users to produce and comprehend language. These paradigms occasionally involved language–vision interactions. One well-known example is the sentence–picture verification paradigm. For example, in an influential application of this paradigm, Clark and Chase (1972) measured reaction times as participants judged whether sentences such as "The cross is above the star" matched schematic pictures in a display. The results were used to argue for propositional representations, including a common representational format for knowledge acquired from linguistic and visual input. Another influential set of studies examined the processing of negation (Carpenter & Just, 1975), demonstrating that utterances with a negative operator such as *not* tend to be difficult to understand.

At about the same time, in work that did not have as immediate an impact, Cooper (1974) tracked participants eye movements as they listened to stories while looking at a display of pictures. Cooper found that participants initiated saccades to pictures that were named in the stories, as well as pictures associated to those words. Moreover, fixations were often generated before the end of the word. Cooper's remarkable article was presciently titled. "The control of eye fixation by the meaning of spoken language: a new methodology for the real-time investigation of speech perception, memory and language processing". This study went relatively unnoticed, however, perhaps in part because the findings did not seem relevant given the theoretical debates taking place at the time. For example, one hot topic was the representation and processing of syntactically complex sentences (e.g., Forster, 1970; Hakes, 1972; Holmes & Forster, 1972), and the Cooper findings and paradigm did not obviously bear on that issue.

More recently, as reliable and inexpensive eye-trackers have become available, there has been an exponential increase in research using eye movements to study spoken language processing, beginning with Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy (1995). Psycholinguists are currently using what Tanenhaus and colleagues dubbed the "visual world" paradigm to address questions that run the gamut of topics in language processing, ranging from speech perception to real-time language production and comprehension during interactive conversation. Many of these studies focus on classic issues in language processing; others are exploring relatively uncharted territory (for recent reviews see Tanenhaus, 2007; Tanenhaus & Trueswell, 2006). There is also some nascent work on how language processing might affect visual search (e.g., Spivey, Tyler, Eberhard, & Tanenhaus, 2001). Why then dedicate valuable pages of the *Journal of Memory and Language* to a special issue devoted to language–vision interactions?

The seeds of this special issue were planted in a workshop organized by John Henderson and Fernanda Ferreira at Michigan State University in May, 2003. This workshop brought together a handful of psycholinguists who had begun to use eye movements to study spoken language comprehension and production, and a small group of psychologists who use eye movements to study scene perception, reading and visual attention (for an edited volume of papers from that workshop, see Henderson & Ferreira, 2004). One of the organizers' goals was to increase interaction among researchers in these areas. Another of their less explicit goals was to use the workshop as an intervention—think short stay at the academic equivalent of a boutique recovery center—in which psycholinguists using the visual world paradigm were confronted with the highly relevant literature in scene perception and visual attention, and with their numerous unexamined and often erroneous assumptions (for a review see Henderson & Ferreira, 2004). For example, researchers using the visual world paradigm often did not appreciate differences between pictures and scenes, were unaware of recent developments in our understanding of visual search, and rarely considered how characteristics of the display might influence how much information was encoded prior to the onset of an utterance.

Our goal in this special issue is to pick up on the challenge initiated by Henderson and Ferreira. Although we would have been delighted to have received contributions that use language to investigate visual processing, all of the authors focus primarily on spoken language processing. Each of the articles examines language processing in the context of a visual display, typically a set of pictures or a depicted scene, and typically using eye movements as the primary response measure (but cf. the contributions by Farmer, Cargill & Spivey and Bard et al.). For the most part, however, the authors do more than use eye movements to a visual display as an index of attention during language processing. They also examine how visual and linguistic representations interact to address representational issues at the language–vision interface. In doing so, they begin to raise questions that we hope will serve as a catalyst for more systematic research on language–vision interactions. In the remainder of our introduction we briefly introduce the papers in this issue, placing each in context, and highlighting the questions that each raises about language–vision interactions.

The first two articles examine spoken word recognition, building upon the line of research initiated by Allopenna, Magnuson, and Tanenhaus (1998). Allopenna et al. showed that fixations to pictures provide a fine-grained measure of lexical access that can be used to evaluate predictions from competing models, via a linking hypothesis relating fixations to accumulating activation (evidence) for alternative lexical candidates. The sensitivity of eye movements to fine-grained acoustic detail coupled with the close time locking between fixations and the unfolding speech stream has made the visual world paradigm a useful psychophysical measure for studying the phonetic/lexical interface. The strength of the inferences that can be drawn from such studies will, however, depend on understanding the nature of the representations that link hearing a spoken word with a rapid shift in attention to its referent.

Huettig and McQueen consider three possible types of representations that could mediate fixations to a target picture: phonological representations, visual/perceptual representations (e.g., information about shape), and semantic conceptual representations. Hearing a word or seeing a picture can presumably access each of these types of representations, a conclusion that is supported by a growing literature in brain imaging, and each could mediate the link between hearing a word and fixating a picture. Huettig and McQueen use the time course of fixations to phonological, visual, and semantic competitors to argue that all three types of representations can influence looks to pictures (and printed words), with the earliest fixations mediated by phonological representations, except when preview time is limited.

Dahan and Gaskell address the role of frequency in spoken word recognition, combining detailed analyses of the timing of fixations with information arriving from the speech input. They conclude that frequency effects continue to persist even after disambiguating information arrives, interpreting their results within the Bayesian framework discussed in Norris (2006). Dahan and Gaskell also examine how having previously looked at a picture whose name is a phonological competitor affects the likelihood that people will again fixate this picture when they hear a spoken word. They conclude that shape/object file representations are more likely to link fixations between pictures and words than picture names.

More generally, developing a plausible model for how listeners link a spoken word to a visual referent is an ambitious but tractable challenge for cognitive science and cognitive neuroscience. Meeting that challenge will require sophisticated studies and models of the vision–language interface, including more detailed knowledge of the nature of the perceptual representations that become activated as a spoken word is recognized. We hope that the contributions by Huettig and McQueen and Dahan and Gaskell will inspire others to take up the challenge.

Altmann and Kamide's contribution also addresses the nature of the representations that link language comprehension to looks to depicted objects, focusing on sentences rather than words. Beginning with Altmann and Kamide's (1999) influential study, anticipatory eye movements have been widely used to infer the nature of the expectations that listeners are generating. But what is the nature of the representations that drive listeners expectations, and how do those expectations mediate looks to elements of a display or scene? Altmann and Kamide address this question in empirical studies examining anticipatory looks to depicted objects that differ in their plausibility as participants in a future or a past event (the man will drink/has drunk, etc.). They also develop an affordance-based account of why people look while listening that makes important links with related research in vision.

Knoeferle and Crocker present an explicit model of how linguistic and visual information are integrated online. Their Coordinated Interplay Account assumes that referents and linguistic terms get coindexed incrementally during processing, and this coindexation is accomplished by making eye movements to objects during linguistic processing. Knoeferle and Crocker describe three experiments designed to test the basic assumptions of this model, using a paradigm in which successive scenes relating to the same event are presented in sequence, leading to a quasi-dynamic depiction of an event. These experiments demonstrate that people tend to look at stereotyped agents of depicted actions, even when the scene is replaced with a blank screen. The blank screen manipulation together with the use of scenes presented in sequence also allow Knoerfle and Crocker to explore, how working memory is involved in the linking of visual and linguistic information.

Researchers studying language production have long used pictures and depicted events to study utterance formulation. In an important study, Griffin and Bock (2000) found that speaker's eye movements to pictures are closely time-locked to utterance production (also see Meyer, Sleiderink, & Levelt, 1998). Surpisingly, they found little evidence that the speaker's initial fixation during apprehension of the depicted event influenced the form of the speaker's utterance. Gleitman, January, Nappa and Trueswell revisit this issue using an ingenious attentional capture manipulation to control where the speaker looks first. Gleitman and colleagues demonstrate that, for a variety of linguistic constructions, the first look to a depicted object influences the form of the speaker's utterance, so that concepts in the focus of attention tend to be grammatically encoded as subjects. This finding is consistent with previous work showing that concepts made available through semantic priming tend to become subjects (Bock, 1996), but extends the result by linking semantic priming and attentional capture for visual events. We expect that this

study will usher in a new era of research that uses sophisticated manipulations of visual attention to study both language production and language comprehension.

Farmer, Cargill and Spivey re-examine some old terrain, how referential context affects syntactic ambiguity resolution, using a potentially powerful new tool, tracking the trajectory of mouse movements. In a replication of Tanenhaus et al. (1995), they find that the trajectory of mouse movements is different for unambiguous sentences and ambiguous sentences that induce temporary garden-paths. Mouse movements for ambiguous sentences in biasing visual contexts are similar to those for unambiguous sentences—a result they interpret as evidence for constraint-based parsing models, and against race-based serial models. Crucially, they provide *prima facie* evidence based on distributional analyses that mouse movements, which they argue provide a more continuous measure of processing than eye movements, can, in principle, distinguish between a continuous garden-path recovery process and a two-stage serial process.

One important source of information that vision makes available in interactive conversation, arguably the most basic arena of language use (Clark, 1992), is information about what an interlocutor is likely to be attending to. Where a person is fixating provides reasonably reliable information about what they are attending to. Therefore, interlocutors might make use of each others eye gaze to monitor each others attentional states. In a clever set of experiments that uses speaker eye gaze as an independent variable and addressee eye gaze as a dependent variable, Hanna and Brennan show that listeners will use information about the direction of a speaker's gaze to help resolve the referent of a temporarily ambiguous referring expression.

Bard, Anderson, Chen, Nicholson, Havard and Denzel-Job also examine how interlocutors might use eye gaze. Bard et al. embed their investigation of eye gaze within the Edinburgh map task (Anderson et al., 1991), an interactive task involving an instructor and a follower (listener), that Bard has employed to great effect to study unscripted interactive conversation within a structured, well-understood task environment. Using a cursor to simulate the listeners eye gaze, Bard et al. track the instructor eye gaze. The instructor's eye gaze is used to determine when the instructor monitors the listener's eye gaze and how that monitoring affects her utterances. Bard et al. use sophisticated analyses of the dialogue to compare predictions of three models of the speaker's responsibility for monitoring listener-privileged information: duplicated responsibility, shared responsibility and cognitive load, concluding that shared responsibility provides the best account of their data.

Two articles on prosody will appear in JML in February of 2008. Investigating prosody in natural language poses special challenges for investigators, which is perhaps one reason why such work is woefully underrepresented in psycholinguistics. In English, for example, phrasing, stress, and the placement and nature of pitch accents interact in complex ways to produce effects that despite being intuitively striking are difficult to describe in formal (or even informal) terms. The challenges extend to the laboratory: the type of information prosody conveys is as difficult to manipulate experimentally as it is to define formally. Snedeker and Yuan examine the effects of prosody on syntactic ambiguity resolution in children and adults. After first demonstrating that previous failures to find effects of prosody with children are likely due to blocked designs that can encourage perseverative behavior, Snedeker and Yuan explore how children and adults combine lexical and prosodic constraints. Their contribution is an outstanding example of the emerging literature in real-time sentence processing in young children, initiated by Trueswell, Sekerina, Hill, and Logrip (1999), that leverages the visual world paradigm to study sentence processing in pre-literate children.

Ito and Speer address another aspect of prosody, the interpretation of different pitch accents—the building blocks for the "tune" of an utterance that are realized on stressed syllables for accented words. Ito and Speer use an ingenious Christmas tree decorating task to explore the hypothesis that listeners develop different expectations for pitch accents that are treated as distinctive within the influential Pierrehumbert and Beckman TOBI framework (Beckman, Hirschberg, & Shattuck-Hufnagel, 2005; Beckman & Pierrehumbert, 1986). Ito and Speer's contribution illustrates a growing trend towards examining language processing in richer more naturalistic tasks, which parallels a similar trend in vision (for recent reviews see Hayhoe & Ballard, 2005; Land, 2007).

As we hope these brief summaries make clear, the nine papers included in this special issue of JML cover a broad range of topics in language processing, ranging from speech perception to coordination during conversation, and they provide valuable insights into the way that information from linguistic and visual systems are combined. Although, our understanding of how these two key cognitive systems communicate has increased dramatically over the last 12 years or so, it is also obvious that a great deal of work remains to be done. For example, little is known about the way truly dynamic visual information is integrated with an ongoing linguistic utterance. Also, the visual displays are still not nearly as complex as the real world, so that it is difficult to assess how integration occurs when the environment is cluttered and busy, and when objects are difficult to locate or occluded by other objects. But progress is clearly being made, and we are optimistic that these and other critical questions about the interface of language and vision will be the focus of many researchers' attention over the next few decades. In the future we also

hope that Henderson and Ferreira's first goal might be realized more fully, with researchers studying visual cognition beginning to explore the possibility that investigation the language–vision interface might help illuminate issues in vision. For example, investigations of the perceptual representations that become available as a spoken word unfolds might shed light on some of the representations involved in object recognition. But, that would be a topic for a future special issue.

## References


Allopenna, P., Magnuson, J., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419–439.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247–264.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., et al. (1991). The HCRC map task corpus. *Language and Speech, 34*, 351–366.

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S. A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 9–54). Oxford, UK: Oxford University Press.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook, 3*, 255–309.

Bock, K. (1996). Language production: Methods and methodologies. *Psychonomic Bulletin & Review, 3*, 395–421.

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review, 82*, 45–73.

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

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology, 3*, 472–517.

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.

Forster, K. I. (1970). Visual perception of rapidly presented word sequences of varying complexity. *Perception & Psychophysics, 8*, 215–221.

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

Hakes, D. T. (1972). Effects of reducing complement constructions on sentence comprehension. *Journal of Verbal Learning and Verbal Behavior, 11*, 278–286.

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

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 1–58). New York: Psychology Press.

Holmes, V. M., & Forster, K. I. (1972). Perceptual complexity and underlying sentence structure. *Journal of Verbal Learning and Verbal Behavior, 11*, 148–156.

Land, M. (2007). Fixation strategies during active behavior: A brief history. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 75–98). Oxford: Elsevier.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America, 100*, 1777–1786.

Massaro, D. W., & Stork, D. G. (1998). Speech recognition and sensory integration. *American Scientist, 86*, 236–244.

McGurk, H., & MacDonald, J. W. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

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

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113*, 327–357.

Spivey, M. J., Tyler, M. J., Eberhard, K. M., & Tanenhaus, M. K. (2001). Linguistically mediated visual search. *Psychological Science, 12*, 282–286.

Tanenhaus, M. K. (2007). Eye movements and spoken language processing. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 443–469). Oxford: Elsevier.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). The interaction of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632–1634.

Tanenhaus, M. K., & Trueswell, J. C. (2006). Eye movements and spoken language comprehension. In M. Traxler & M. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (2nd ed., pp. 863–900). New York: Academic Press, Elsevier.

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