Acta Psychologica 137 (2011) 208–216

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

I see what you're saying: The integration of complex speech and scenes during language comprehension

Richard Andersson [a], Fernanda Ferreira [b], John M. Henderson [b,*]

[a] Lund University Cognitive Science, Lund University, Sweden
[b] Department of Psychology, University of South Carolina, United States

ARTICLE INFO

Article history:
Received 29 March 2010
Received in revised form 7 January 2011
Accepted 11 January 2011
Available online 8 February 2011

PsycINFO classification:
2720 Linguistics & Language & Speech
2326 Auditory & Speech Perception
2346 Attention

Keywords:
Language comprehension
Scene perception
Eye movements
Attention
Visual world

ABSTRACT

The effect of language-driven eye movements in a visual scene with concurrent speech was examined using complex linguistic stimuli and complex scenes. The processing demands were manipulated using speech rate and the temporal distance between mentioned objects. This experiment differs from previous research by using complex photographic scenes, three-sentence utterances and mentioning four target objects. The main finding was that objects that are more slowly mentioned, more evenly placed and isolated in the speech stream are more likely to be fixated after having been mentioned and are fixated faster. Surprisingly, even objects mentioned in the most demanding conditions still show an effect of language-driven eye-movements. This supports research using concurrent speech and visual scenes, and shows that the behavior of matching visual and linguistic information is likely to generalize to language situations of high information load.

One powerful method for investigating the integration of language and vision is the practice of monitoring the eye movements people make as they listen to speech while simultaneously looking at a visual world containing relevant objects. This technique allows psycholinguists to study how information sources are integrated in real-time to allow comprehenders to form interpretations and link linguistic forms to real-world referents (see Tanenhaus & Brown-Schmidt, 2008, for a review). For example, research has shown that listeners use the visual scene context to constrain the set of possible target referents (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Knoeferle, Crocker, Scheepers, & Pickering, 2005). Altmann and Kamide (1999) showed that listeners use verb information to anticipate a postverbal object, and they later demonstrated the use of real-world information as well (Kamide, Altmann, & Haywood, 2003; see also Ferreira & Tanenhaus, 2007).

These studies demonstrate that linguistic interpretations are used to guide the eyes almost immediately to relevant objects in the visual world. Moreover, listeners are highly likely to fixate an object within about a one-second window following the onset of a word, even when nothing about the task seems to demand that the word and the object be linked. What accounts for this tendency to fixate on objects mentioned in speech? One possibility is that this link allows the comprehender to form a much richer and detailed representation than would be possible otherwise (see e.g., Altmann & Kamide, 2007; Ferreira, Apel, & Henderson, 2008; Richardson, Altmann, Spivey, & Hoover, 2009).

To understand the nature of the eye movements in the so-called Visual World Paradigm (Tanenhaus, Spivey, Eberhard, & Sedivy, 1995) and the strength of this link, it is important to conduct investigations using stimuli that are sufficiently complex to tax the language–vision interface. This is necessary in order to see whether this link weakens in demanding language situations, for example by the comprehender prioritizing processing resources elsewhere. Up to now, most experiments have involved the presentation of a single sentence per trial, and typically only one word in that sentence is identified as a potential target of eye movements. In natural speech, of course, people often hear multiple sentences containing several objects that may be of interest and may therefore become the target of an eye movement. In addition, many of the stimuli that have been presented have been simple line drawings of scenes, or scenes created from pasting clip-art images together in such a way that an event such as a wizard painting a princess is strongly implied. A simple display may allow the participant to preview all objects and possible targets, subvocalize them, and thus pre-generate the linguistic labels that may appear in the speech (for visual search, see Zelinsky & Murphy, 2000, but see also Dahan & Tanenhaus, 2005). Conscious encoding of the objects by the participants is normally disregarded (Tanenhaus, Magnuson, Dahan, & Chambers, 2000:564), but still, typical stimuli

* Corresponding author. Tel.: +1 803 777 41 37.
E-mail address: jhender@mailbox.sc.edu (J.M. Henderson).

0001-6918/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.actpsy.2011.01.007

displays in the visual world paradigm contain clearly identifiable objects in limited numbers, which provide every possibility to do precisely this pre-processing. As well as the flow of information can move from phonological form to visual form, it may as well move in the opposite direction (see Huettig & McQueen, 2007, for a discussion). The pre-processing may also involve memorizing the object locations or visual aspects of the objects. This would imply that simple displays have a processing advantage compared to complex scenes which do not allow this pre-processing.

However, there are studies using real-world objects as targets which have investigated the effect of somewhat complex scenes, but also with limitations to the demand on the language–vision interface. For example, a set-up by Hanna and Tanenhaus (2004) used 10 possible visual targets and referents, but allowed the participant to preview all objects and keep them highly activated. Similarly, a study by Brown-Schmidt, Campana, and Tanenhaus (2005) used a 5×56 grid of possible referential targets. However, the study used only four participant pairs (who may not be representative) and the same visual scene was used throughout the entire experiment (~2.5 h), allowing participants to become more and more familiar with the display and allowing gradually reduced complexity as portions of the display were used up. A study by Brown-Schmidt and Tanenhaus (2008) used an irregular display of 57 different objects and showed how a conversation, as opposed to merely calling out the names of the objects, helps to restrict the referential domain. The authors identify the proximity, relevance and recency of referents as helpful factors in restricting the referential domain. In this experiment, however, the display was semi-permanent in the sense that the available game board was always present and all objects to be used, except one, were also present (either as blocks or stickers). This allowed for a continuity in the visual scene and as such, the display was not as complex as an equivalent display of 57 objects where the object types are freshly generated every trial. Many real scenes are quite different (see Henderson & Ferreira, 2004, for discussion), as the reader can verify by simply looking around his or her immediate environment. Scenes may contain almost uncountable numbers of objects, some predictable, but many not, and often only temporary present never to return again. And in a situation in which objects in the scene are mentioned in speech, a very large proportion of the scene content will be irrelevant to the utterance, or at least will not be mentioned. As a result, the comprehender attempting to link words and objects in the world may have a far more demanding task than has so far been considered in visual world experiments: Utterances are multi-sentence and may contain multiple referents; and scenes are complex and may contain hundreds or thousands of objects, only a few of which are relevant at a given moment in linguistic processing. This is not to say that *all* scenes and utterances are complex, but they represent a subset of the possible scene and utterance combinations that we believe has been neglected.

Of course, it is also important to note that the properties of real-world utterances and scenes do not only make the situation for the comprehender more challenging; they may also make the task easier, because natural stimuli are constrained in ways that likely facilitate processing. For example, connected sentences tend to be coherent, and so a series of utterances may help to converge on the possibility that a particular object will soon be mentioned; and real scenes allow the rapid extraction of gist (e.g., this is a playground scene), allowing listeners to anticipate which object will be mentioned and where in the scene it is likely to be found (Castelhano & Henderson, 2007; Torralba, Oliva, Castelhano, & Henderson, 2006). Also, as shown by Brown-Schmidt and Tanenhaus (2008), a real two-way conversation may help to restrict the referential domain.

To understand to what extent people look at objects when they are mentioned in extremely complex settings, we conducted a study in which participants viewed photographs of complex real-world scenes. A representative example is shown in Fig. 1. The scenes contained a large number of objects arranged in a typically cluttered and busy manner.

The linguistic material presented to participants was also more complex than in typical studies, consisting of three sentences, the second of which was designated as the target sentence. These passages were spoken at either a slow or fast rate of speech. The purpose of this rate manipulation was as to allow the participant less or more time to navigate the scene and find the target object. This added visual search task on top of the linguistic processing task served to increase the information processing demands. Moreover, the eye movement system requires a minimum latency of about 150–170 ms to program a saccade to a fixed target (Rayner, 1998). Thus, with faster speech, the probability increases that the eye movement system will have trouble keeping up with the input because it must locate referents, program saccades to them, and fixate on them long enough for identification and integration (Gibson, Eberhard, & Bryant, 2005).

Fig. 1. A typical stimulus scene with multiple objects.