Acta Psychologica 182 (2018) 91–99

Contents lists available at ScienceDirect

Acta Psychologica

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

Mapping language to visual referents: Does the degree of image realism matter?☆

Raheleh Saryazdi*, Craig G. Chambers

University of Toronto, Canada

ARTICLE INFO

Keywords:
Spoken language comprehension
Visual cognition
Iconicity
Reference
Attention

ABSTRACT

Studies of real-time spoken language comprehension have shown that listeners rapidly map unfolding speech to available referents in the immediate visual environment. This has been explored using various kinds of 2-dimensional (2D) stimuli, with convenience or availability typically motivating the choice of a particular image type. However, work in other areas has suggested that certain cognitive processes are sensitive to the level of realism in 2D representations. The present study examined the process of mapping language to depictions of objects that are more or less realistic, namely photographs versus clipart images. A custom stimulus set was first created by generating clipart images directly from photographs of real objects. Two visual world experiments were then conducted, varying whether referent identification was driven by noun or verb information. A modest benefit for clipart stimuli was observed during real-time processing, but only for noun-driving mappings. The results are discussed in terms of their implications for studies of visually situated language processing.

As a spoken sentence unfolds, listeners are able to rapidly map language to objects and actions in the external visual environment. This is reflected in the timing and pattern of eye movements that are generated as participants listen to spoken utterances (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Here, we examine the details of this mapping process by exploring the *kind* of visual entity to which language is mapped, namely photographs versus clipart images of individual objects. Although both kinds of images are routinely used in studies of visually situated language processing, the rationale behind using a particular stimulus type is rarely motivated, and instead appears to reflect issues of convenience or experimenters' personal preference. Nevertheless, work in visual cognition and cognitive development has outlined various perceptual and conceptual factors that could plausibly influence the time course of referential mapping with different kinds of object depictions. The aim of the present study is to examine if and to what extent the manner in which objects are depicted affects the course and nature of on-line language processing.

As background, various studies have demonstrated that performance on non-linguistic tasks can differ with real objects versus 2-dimensional (2D) depictions of these objects. For example, Snow, Skiba, Coleman, and Berryhill (2014) examined recall and recognition performance in a study comparing real objects, coloured photographs, and black and white line drawings. Participants' scores on both measures were higher for real objects compared to the 2D image types. Of greater relevance to the current study is the question of whether performance varies within different kinds of 2D images that differ in their visual iconicity, namely the degree to which they are perceptually faithful representations of real-world objects they are intended to depict. A colour photograph of an apple, for example, resembles its real-world analogue to a greater degree than a black-and-white photograph, which is in turn a truer representation than an artist's pencil sketch of the same object. Digital images also fall along this continuum (e.g., realistic-looking photo objects, coloured clipart, black–and-white clipart).

The question of whether the iconicity of 2D images affects cognitive processes has been particularly important for work in the area of cognitive development. In one study, Pierroutsakos and DeLoache (2003) presented 9-month-olds with four different types of stimuli (i.e., colour photographs, black-and-white photographs, colour line drawings, and black-and-white line drawings) and observed how the infants interacted with the images. The greater the overlap between the depicted and the real object it represented (i.e., colour photograph), the more the child attempted to manually interact with the image. Over the course of development, young children gain an appropriate understanding of the symbolic nature of 2D representations, acquiring what is often referred to as *pictorial competence* (DeLoache, 2004; DeLoache, Pierroutsakos, & Uttal, 2003; DeLoache, Pierroutsakos, Uttal, Rosengren, & Gottlieb, 1998). Related work has examined children's understanding of the 2D

☆ Funding for this research was provided by the Social Sciences and Humanities Research Council of Canada (496721).
* Corresponding author at: University of Toronto Mississauga, 3359 Mississauga Road, Mississauga L5L1C6, Ontario, Canada.
E-mail address: raheleh.saryazdi@mail.utoronto.ca (R. Saryazdi).

https://doi.org/10.1016/j.actpsy.2017.11.003
Received 23 February 2017; Received in revised form 10 October 2017; Accepted 4 November 2017

images in pictures books, which are an important vehicle for learning about object concepts. Indeed, children learn many concepts they have not yet seen or will never see (imaginary things or extinct animals) from picture books (Troseth, Pierroutsakos, & DeLoache, 2004). These studies have shown that learning is facilitated by more realistic depictions of objects (e.g., Ganea, Pickard, & DeLoache, 2008; Simcock & DeLoache, 2006; Tare, Chiong, Ganea, & DeLoache, 2010).

Advantages for more realistic 2D images have also been reported in adult studies. Salmon, Matheson, and McMullen (2014) reported that images of manipulable objects were better recognized when presented as photographs than as line drawings. This was attributed to the activation of motor representations in the photograph condition due to the more true-to-life depiction. Further, even within a given image type (e.g., drawings), the degree of realism can affect performance. One clear illustration comes from Rossion and Pourtois (2004), whose addition of surface detail and colour to the widely used black-and-white line drawings by Snodgrass and Vanderwart (1980) significantly increased naming accuracy and reduced response times (see also Bramão, Reis, Petersson, & Faísca, 2011 for a review and meta-analysis relating specifically to colour). Similarly, even with photographs, objects located in scene regions that are comparatively less "real" (e.g., in a mirror reflection), are correspondingly less likely to be labelled or noticed in a change detection task than those located in scene regions depicting genuine 3D space (Sareen, Ehinger, & Wolfe, 2015). However, in contrast to this work, some studies comparing performance across different image types have reported little or no effect, in either behavioural measures (e.g., Biederman & Ju, 1988; Snow et al., 2014) or in patterns of brain activity (e.g., Kourtzi & Kanwisher, 2000; Walther, Chai, Caddigan, Beck, & Fei-Fei, 2011). For example, in the Snow et al. (2014) study mentioned earlier, the benefit observed for 3D objects over 2D depictions did not carry over to 2D stimuli that differed in their degree of realism (photographs vs. line drawings).

Even with a more stable pattern of findings, however, the implications for real-time linguistic processing would be uncertain. This is because visual recognition and categorization processes could arguably occur before the point where referential language is encountered. For example, the linguistically-relevant category of a 2D object presented on a screen, and its corresponding spatial position, might be robustly encoded in an internal representation that abstracts beyond the perceptual features that underlie realism differences. However, in its strong form, this assumption may not be entirely secure. Studies of visual cognition have suggested that perceivers do not generate or maintain especially robust internal representations of objects and their properties when this information is accessible in the external visual world (Cohen, Dennett, & Kanwisher, 2016; Hayhoe, Bensinger, & Ballard, 1998; Triesch, Ballard, Hayhoe, & Sullivan, 2003). In addition, studies of real-time language processing have demonstrated listeners' apparent openness to different descriptions for the same visual object (e.g., Heller & Chambers, 2014; Pontillo, Salverda, & Tanenhaus, 2015). These flexible expectations for how objects will be described suggest that visual encoding involves a certain degree of fuzziness or in-parallel representation. Moreover, other studies have shown that listeners use various coarse-level features (e.g., shape, colour) to identify target referents at the moment when a description is heard. For example, participants are likely to briefly fixate the image of a rope upon hearing *snake* in view of the shared shape characteristics (Dahan & Tanenhaus, 2005; see also Huettig & Altmann, 2007; Huettig & McQueen, 2007; Rommers, Meyer, & Huettig, 2015; Rommers, Meyer, Praamstra, & Huettig, 2013; Yee, Huffstetler, & Thompson-Schill, 2011). Similar effects have also been observed in studies examining the role of colour (e.g., Huettig & Altmann, 2011; Johnson & Huettig, 2011; Johnson, McQueen, & Huettig, 2011, see Huettig, Rommers, & Meyer, 2011, for a review). These effects would be unexpected if eye movements were programmed solely using a mental representation based on viewed objects' conceptual labels and their associated spatial coordinates. Instead, the surface features of objects seem to have some importance at the point when the noun is encountered. Given that clipart and photographic images differ in the quality of their surface features, these findings suggest yet another way in which the degree of realism of an image could influence the course of referential mapping, if perhaps only to a subtle degree.

To our knowledge, no study of real-time referential processing has specifically explored the question of iconicity per se. However, a handful of studies have asked more general questions about whether effects found with clipart scenes extend to tasks using photographic scenes. Andersson, Ferreira, and Henderson (2011) found that the incremental interpretation of unfolding language (as reflected in eye movement patterns) is still observed when photographs of cluttered scenes are used as stimuli. Similarly, Staub, Abbott, and Bogartz (2012) and Coco, Keller, and Malcolm (2015) found that listeners show verb-driven anticipatory fixations when photorealistic scenes are used as stimuli (complementing earlier work with clipart stimuli, e.g., Altmann & Kamide, 1999). These studies, however, focused on scene-level realism and did not involve a direct comparison of photographs versus clipart using the same experimental task and materials. As a result, they do not address the more fine-grained question asked here, namely whether the degree of realism for an individual object can have an effect on aspects of referential mapping (independent of the additional contextual and complexity-related factors involved in natural scenes).

If there is a difference between more and less iconic images, in what direction would the difference lie? Consistent with some of the studies reviewed above, one possibility is that stimuli with greater visual iconicity would facilitate real-time reference resolution due to their inherent naturalness. A related argument comes from linguistic pragmatics. Consider, for example, that neither a clipart image nor a photograph of a cup is an actual instance of the object category. An utterance like *Look at the cup* therefore serves as a linguistic shortcut for something like *Look at the picture of the cup*. Listeners' ability to map the unadorned description *the cup* to a 2D referent may involve a type of pragmatic accommodation performed in the service of being a cooperative language user (Grice, 1975). This accommodation might be easier for photographs because they are a more veridical reflection of a real-world object, and are also intended as such.

Conversely, it is also possible that a processing benefit might be found with clipart images in view of their clearer contours and simpler textures/colouring. An outcome in this direction would be consistent with the notion of "edge-based" representations of objects, which is a central element of certain theories of visual cognition. On this account, recognition is driven first and foremost by basic structural components of an object, with additional surface features (i.e., colour, texture, shadows) playing a minimal and often secondary role (Biederman, 1987; Biederman & Ju, 1988). The notably clearer edge contours of clipart stimuli (e.g., defined black outlines, no shadows) could therefore be beneficial. Language-referent mappings could also be easier with clipart stimuli for conceptual reasons. Given that linguistic terms are understood to encode semantic meanings that abstract beyond episodic experiences (i.e., real-world encounters involving specific objects and actions, see Altmann, 2016), the correspondingly abstract and "episodically neutral" character of clipart images might be facilitative for mapping linguistic terms to their corresponding referents.

In addition to addressing the theoretical questions outlined above, an improved understanding of the role of iconicity has methodological value. Researchers interested in optimizing their experimental designs would benefit from knowing if a particular image type facilitates referential processing. Similarly, the absence of a difference would suggest researchers can freely choose the image type that is convenient for the study at hand. To address these questions, the current study explores potential differences in eye movement patterns as listeners hear language referring to either photographs or clipart images. We use a visual world methodology to examine the effect of image type on the real-time mapping of language to 2D visual referents (photographs of everyday objects vs. clipart images). All critical stimuli began as custom