Neuropsychologia 95 (2017) 21–29

Contents lists available at ScienceDirect

Neuropsychologia

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

Linking language to the visual world: Neural correlates of comprehending verbal reference to objects through pointing and visual cues

David Peeters[a,*], Tineke M. Snijders[a,b], Peter Hagoort[a,b], Aslı Özyürek[a,b]

[a] Max Planck Institute for Psycholinguistics, 6500 AH, Nijmegen, The Netherlands
[b] Radboud University Nijmegen, 6525 HP, Nijmegen, The Netherlands

ARTICLE INFO

Keywords:
Multimodal communication
Gesture
Pointing
Conceptual matching
Referential communication
Pragmatics
Mentalizing

ABSTRACT

In everyday communication speakers often refer in speech and/or gesture to objects in their immediate environment, thereby shifting their addressee's attention to an intended referent. The neurobiological infrastructure involved in the comprehension of such basic multimodal communicative acts remains unclear. In an event-related fMRI study, we presented participants with pictures of a speaker and two objects while they concurrently listened to her speech. In each picture, one of the objects was singled out, either through the speaker's index-finger pointing gesture or through a visual cue that made the object perceptually more salient in the absence of gesture. A mismatch (compared to a match) between speech and the object singled out by the speaker's pointing gesture led to enhanced activation in left IFG and bilateral pMTG, showing the importance of these areas in conceptual matching between speech and referent. Moreover, a match (compared to a mismatch) between speech and the object made salient through a visual cue led to enhanced activation in the mentalizing system, arguably reflecting an attempt to converge on a jointly attended referent in the absence of pointing. These findings shed new light on the neurobiological underpinnings of the core communicative process of comprehending a speaker's multimodal referential act and stress the power of pointing as an important natural device to link speech to objects.

1. Introduction

In everyday talk, people often refer to things in their immediate surroundings. In such situations, an important prerequisite for communicative success is for speaker and addressee to establish joint attention to the object, person, or event they are talking about. Imagine you are sitting at the window in a restaurant and your friend says "Look at that car". How do you identify the specific car your friend is talking about? In many such cases, a speaker may connect her communication to the entity she is referring to by manually pointing at it (Bühler, 1934; Clark, 1996; Kita, 2003), helping the addressee to single out the intended referent (one specific car). In other cases a pointing gesture may not be necessary because one object in the environment is clearly perceptually most salient, such that the addressee may infer that the speaker refers to the salient object (Clark et al., 1983). In both cases, the addressee needs to match the visual object that is referred to (the car) to the spoken label by which it is described ("car"). The aim of the current study is to advance our understanding of the neural architecture supporting this everyday communicative process, both when an object is singled out by a pointing gesture and when it is made perceptually salient by non-communicative physical properties.

Comprehending our interlocutors' pointing gestures is a core feature of everyday communication (Baron-Cohen, 1989; Clark, 1996; Kendon, 2004; Tomasello et al., 2007). Previous neuroimaging studies have looked at the neural correlates of observing pointing gestures outside a referential speech context and at their integration with cues such as the gesturer's gaze direction (e.g., Brunetti et al., 2014; Conty et al., 2012; Gredebäck et al., 2010; Materna et al., 2008; Redcay et al., 2015; Sato et al., 2009). Perceiving a pointing hand compared to perceiving a non-directional closed hand elicits enhanced activation in a set of mainly right-hemisphere regions, including right inferior frontal gyrus (IFG), right angular gyrus, right parietal lobule, right thalamus, and bilateral lingual gyri (Sato et al., 2009). Following the direction of someone's pointing finger elicits bilateral posterior superior temporal sulcus (pSTS) activation (Materna et al., 2008). Integrating someone's pointing gestures with their gaze direction recruits parietal and supplementary motor cortices in the right hemisphere (Conty et al., 2012). Together, these findings suggest an extensive right-hemisphere dominant network that is activated when one perceives a manual pointing gesture that shifts one's attention.

* Correspondence to: Max Planck Institute for Psycholinguistics, P.O. Box 310, NL-6500 AH, Nijmegen, The Netherlands.
E-mail address: david.peeters@mpi.nl (D. Peeters).

http://dx.doi.org/10.1016/j.neuropsychologia.2016.12.004
Received 5 September 2016; Received in revised form 25 November 2016; Accepted 5 December 2016
Available online 06 December 2016
0028-3932/

In everyday communication, however, pointing gestures are not observed in isolation and often shift one's attention toward a visible entity such as an object (Kita, 2003). Pierno et al. (2009) compared the observation of an image of a hand pointing at an object to the observation of an image of a hand grasping an object and to a control condition of an image of a hand resting next to an object. In comparison to the control condition, the perception of the pointing hand and object elicited enhanced activation in left middle temporal gyrus (MTG), left parietal areas (postcentral gyrus and supramarginal gyrus) and left middle occipital gyrus. The pointing condition did not elicit additional activation compared to the grasping condition. Nevertheless these results suggest that, in addition to the right-hemisphere dominant network involved in perceiving a pointing hand that shifts one's attention, a left-lateralized set of cortical areas may subsequently be involved in visually integrating the pointing hand and an object-referent.

The studies described above each contribute valuable information towards a better understanding of the neural architecture involved in observing pointing gestures, but do not reflect the richness of everyday acts of human referential communication. Pointing gestures often occur in a context in which one perceives not only visual information such as an interlocutor's pointing hand and one object, but also the speech that she may concomitantly produce. Furthermore, the pointing gesture may be produced to single out one specific object from a larger set of visible potential referents. In such situations, an addressee needs to combine incoming information from visual (speaker, pointing gesture, and objects) and auditory (speech) modalities to comprehend the referential act. Furthermore, the perceived spoken label needs to be matched to the specific object the speaker intended to refer to for communication to be successful. The current study focuses on the comprehension of pointing gestures in such richer audiovisual contexts.

The main aim of the current study is to get a better understanding of the neural infrastructure involved in the conceptual matching of a spoken word with a visible object as induced by a referential pointing gesture in comprehension. Pointing gestures may single out an object from a larger set of potential referents while speech may concomitantly describe the object (Bühler, 1934; Clark and Bangerter, 2004), as in someone pointing at an apple while saying "I have bought this apple at the market this morning" (Peeters et al., 2015b). Previous work suggests that conceptual matching between auditory and visual information may recruit pMTG (e.g., Dick et al., 2014). It has been found, for instance, that observing a mismatch (versus a match) between a pantomime gesture and a concurrently encountered spoken word leads to enhanced activation in pMTG (Willems et al., 2009). This suggests that pMTG may be involved in mapping different sources of information onto a common memory representation, a process that has also been called semantic integration (Hagoort et al., 2009; Willems et al., 2009). A typical everyday situation in which semantic integration of auditory (the spoken label) and visual (the identified object) information takes place is presumably referential communication via pointing.

Additionally, in the case of complementary or mismatching signals, a novel conceptual representation may have to be construed. Evidence suggests that this process is subserved by left inferior frontal gyrus (LIFG). Observation of images (e.g. of a dog) paired with an incongruent sound (e.g. meowing), for instance, leads to enhanced activation in LIFG compared to observation of images (e.g. of a dog) paired with a congruent sound (e.g. barking; Hein et al., 2007). In the gestural domain, Dick et al. (2014) compared the perception of supplemental iconic gestures with speech to the perception of “redundant” iconic gestures with speech. The former gestures added information to the speech they accompanied (e.g. the verb in the phrase “Sparky attacked” was combined with a “peck” gesture) whereas the latter gestures (“Sparky pecked” combined with a “peck” gesture) did not. An increase in activation was found in LIFG for the gestures that added information to speech. Both such gestures commonly occur in everyday interactions (Holler and Beattie, 2003; Kendon, 2004; McNeill, 1992), suggesting that enhanced activation in LIFG is not restricted to unnatural mismatch situations. Rather, these findings suggest that LIFG is recruited in the online construction of a novel semantic representation, a process that has also been referred to as semantic unification (Hagoort et al., 2009; Willems et al., 2009). Unlike the iconic co-speech gestures used in previous studies, pointing gestures do not convey semantic information. Nevertheless, they do often relate semantic information in speech to (properties of) a physical object in one's immediate environment. Therefore a conceptual mismatch, induced by a pointing gesture, between a spoken word and a visual object might also recruit LIFG. Activation in LIFG and pMTG may be preceded by activation in pSTS linking auditory and visual information at a pre-lexical level (Dick et al., 2014).

A secondary aim of the current study is to investigate the neural underpinnings of referential communication in situations in which a speaker refers to an object that is perceptually salient, in the absence of pointing. In everyday conversations, addressees may identify a particular referent on the basis of its perceptual salience in the absence of a pointing gesture that singles out the object. Clark et al. (1983) showed participants a picture with four types of flower in it and asked "how would you describe the color of this flower? ", without pointing at one of the specific flowers in the picture. When daffodils were perceptually more salient than the other types of flower, participants described the color of the daffodils. Arguably, the addressee in such cases inferred that the speaker was referring to the object that was perceptually most salient. The neural underpinnings subserving such inferential processes in the comprehension of referential communication are unclear. One possibility is that such situations activate the mentalizing system (medial prefrontal cortex, temporo-parietal junction, and possibly precuneus; Frith and Frith, 2006; Schurz et al., 2014; Van Overwalle and Baetens, 2009), because addressees may attribute a belief or intention to the speaker in relation to their common ground. They both know that they both know that, in the absence of a pointing gesture, the most salient object is most likely the intended referent. This mentalizing process may be less necessary in more straightforward cases where a speaker expresses her communicative intent by simply pointing at an object while concurrently describing it in speech.

1.1. The present study

The present study aims to shed more light on the functional roles of different cortical areas recruited in basic communicative situations in which a speaker refers in speech and/or gesture to an object for an addressee in a visual context. In an event-related functional magnetic resonance imaging (fMRI) study, participants were presented with images of a speaker and two different objects while they listened to her speech. In each picture, one of the objects was singled out, either through the speaker's index-finger pointing gesture or through a visual cue that made the object more salient, in the absence of gesture. We employed a mismatch paradigm, such that the object that was singled out was either congruent (on match trials) or incongruent (on mismatch trials) with concurrent speech. In addition we included two separate unimodal runs (audio-only and visual-only; cf. Willems et al., 2009).

The main aim of the study was to get a better understanding of the neural infrastructure involved in the conceptual matching of a spoken word with a visible object as induced by a referential pointing gesture in comprehension. We predicted that brain areas involved in processing combinatorial semantic information through verbal and gestural channels as found in previous studies might also be relevant in the current manipulation. More specifically, we hypothesized that LIFG would be activated more in the case of a mismatch (compared to a match) between speech and the object that was singled out by the pointing gesture (see Dick et al., 2014 and Özyürek, 2014, for overview; Willems et al., 2007). This is in line with a view of LIFG, more