ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 39 (2008) 2010–2024

# Neural correlates of the processing of co-speech gestures

Henning Holle,[a,*] Thomas C. Gunter,[a] Shirley-Ann Rüschemeyer,[a] Andreas Hennenlotter,[a] and Marco Iacoboni[b]

[a] *Max-Planck-Institute of Human Cognitive and Brain Sciences, Stephanstr. 1a, 04103 Leipzig, Germany*
[b] *Ahmanson-Lovelace Brain Mapping Center, Dept. of Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, Brain Research Institute, David Geffen School of Medicine at UCLA, Los Angeles, USA*

Received 8 May 2007; revised 19 October 2007; accepted 26 October 2007
Available online 13 November 2007

**In communicative situations, speech is often accompanied by gestures. For example, speakers tend to illustrate certain contents of speech by means of iconic gestures which are hand movements that bear a formal relationship to the contents of speech. The meaning of an iconic gesture is determined both by its form as well as the speech context in which it is performed. Thus, gesture and speech interact in comprehension. Using fMRI, the present study investigated what brain areas are involved in this interaction process. Participants watched videos in which sentences containing an ambiguous word (e.g. *She touched the mouse*) were accompanied by either a meaningless grooming movement, a gesture supporting the more frequent dominant meaning (e.g. *animal*) or a gesture supporting the less frequent subordinate meaning (e.g. *computer device*). We hypothesized that brain areas involved in the interaction of gesture and speech would show greater activation to gesture-supported sentences as compared to sentences accompanied by a meaningless grooming movement. The main results are that when contrasted with grooming, both types of gestures (dominant and subordinate) activated an array of brain regions consisting of the left posterior superior temporal sulcus (STS), the inferior parietal lobule bilaterally and the ventral precentral sulcus bilaterally. Given the crucial role of the STS in audiovisual integration processes, this activation might reflect the interaction between the meaning of gesture and the ambiguous sentence. The activations in inferior frontal and inferior parietal regions may reflect a mechanism of determining the goal of co-speech hand movements through an observation–execution matching process.**


## Introduction

Meaningful hand movements (i.e. gestures) are an integral part of everyday communication. It seems once people become involved in a conversation they inevitably start to move their hands to illustrate certain contents of speech. Some of these co-speech gestures bear a formal relationship to the contents of speech and have therefore been termed *iconic* in the literature (McNeill, 1992). For example, a speaker might form a precision grip and make a quick turning movement when uttering a sentence such as: "I tightened the screw". When producing such an iconic gesture, the speaker transmits meaning in two channels simultaneously. Previous research has shown that listeners make use of additional gesture information in comprehending language (Alibali et al., 1997; Beattie and Shovelton, 1999, 2002). Furthermore, progress has been made in specifying the temporal characteristics of gesture–speech interaction in comprehension (Holle and Gunter, 2007; Kelly et al., 2004; Özyürek et al., 2007; Wu and Coulson, 2007). The present study investigates the neural systems involved in the interaction of gesture and speech in comprehension.

Iconic gestures are a special subcategory of gestures (McNeill, 1992). It is helpful to see what delimits iconic gestures from other gesture types (e.g. pantomime, emblems, pointing). Iconic gestures and pantomime have in common that they often illustrate actions. A crucial difference, however, is that there is no speech during pantomime whereas iconic gestures are mostly produced in combination with speech (McNeill, 1992, 2000, 2005). Probably related to this difference is the fact that iconic gestures are less elaborate in their form than pantomimed movements. Iconic gestures are actions recruited in the context of another domain (i.e. speech, cf. Willems et al., 2006), therefore the timing of the gestures is deeply intertwined with the timing of speech (for more on this, see below). Their co-speech timing results in a limited time available for the production of an iconic gesture. Accordingly, pantomimed actions can give very detailed (or even exaggerated) descriptions, whereas iconic gestures tend to be much more casual in form and are often abstractions of the performed actions. Furthermore, in contrast to iconic gestures, a sequence of pantomimed movements can be joined together to create sentence-like constructions (Rose, 2006). Iconic gestures differ from another subcategory of gesture called emblems in their degree of conventionalization. Emblems, which are meaningful hand postures such as the victory sign, are so conventionalized in their form that they can be effortlessly understood in the absence of speech (Gunter and Bach, 2004). In comparison,

* Corresponding author. Fax: +49 0 341 9940 260.
*E-mail address:* holle@cbs.mpg.de (H. Holle).
**Available online on ScienceDirect (www.sciencedirect.com).**

1053-8119/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2007.10.055

iconic gestures are much less conventionalized. Studies investigating iconic gesture production typically find a great degree of interindividual variability in the form of these gestures (e.g. Kita and Özyürek, 2003). Nevertheless, iconic gestures contain additional information that is not found in speech. In the example described above, only the gesture gives an indication about the size of the screw (probably a rather small screw, because a precision grip was used). In a series of previously conducted ERP experiments, we have shown that such additional gesture information can modulate how the two word meanings of lexically ambiguous words (e.g. *ball*) are processed in a sentence context (Holle and Gunter, 2007). Thus, iconic gestures interact with speech during online language comprehension. Not much is known, however, about which brain areas are involved when gesture and speech interact.

The interaction of iconic gestures and speech in comprehension can be approached from at least two different perspectives. First, given that many iconic gestures constitute re-enacted actions, one can adopt an action perspective. From this point of view, an empirical question is the extent to which the processing of iconic gestures recruits the brain network associated with action comprehension. Based on the findings that area F5 and PF of the macaque brain contain neurons that fire both during the observation as well as the execution of goal-directed hand movements (Gallese et al., 1996, 2002; Umiltà et al., 2001), it has been proposed that these so-called mirror neurons form the neural circuitry for action understanding (Rizzolatti et al., 2001). Although direct evidence (via single-cell recording) for mirror neurons in the human brain is still lacking, there is a substantial body of indirect evidence that a similar system exists in humans as well (for recent overviews, see Binkofski and Buccino, 2006; Iacoboni and Dapretto, 2006; Molnar-Szakacs et al., 2006). In particular, the inferior frontal gyrus (IFG) including the adjacent ventral premotor cortex and the inferior parietal lobule (IPL) have been suggested as the core components of the putative human mirror neuron system (MNS) (Rizzolatti and Craighero, 2004). According to a recent theoretical suggestion, the human MNS is able to determine the goal of observed actions by means of an observation–execution matching process (for a more detailed description, see Iacoboni, 2005; Iacoboni and Wilson, 2006). Because many iconic gestures are re-enacted actions, it is therefore plausible that the MNS also participates in the processing of such gestures.

Second, one can adopt a multimodal perspective on iconic gesture comprehension. As has been argued above, iconic gestures show little conventionalization, i.e. there is no "gestionary" that can be accessed for their meaning. Instead, the meaning of iconic gestures has to be generated online on the basis of gesture form and the co-speech context in which the gesture is observed (Feyereisen et al., 1988; McNeill, 1992, 2005). Thus, comprehending a co-speech iconic gesture is a process which requires a listener to integrate auditory and visual information. Within the multimodal view on iconic gestures, a further distinction can be made between local and global gesture–speech integration (see Willems et al., 2006). Because co-speech gestures are embedded in spoken utterances that unfold over time, one can investigate the integration processes between gesture and speech both at a local level (i.e. the integration of temporally synchronized gesture and speech units) as well as on a global sentence level (i.e. how greater meaning ensembles are assembled from smaller sequentially processed meaningful units such as words and gestures).

Local integration refers to the combination of simultaneously perceived gestural and spoken information. Previous research indicates that the temporal relationship between gesture and speech in production is not arbitrary (McNeill, 1992; Morrel-Samuels and Krauss, 1992). Instead, speakers tend to produce the peak effort of a gesture, the so-called stroke, simultaneously with the relevant speech segment (Levelt et al., 1985; Nobe, 2000). This stroke–speech synchrony might be an important cue for listeners in comprehension, because it can signal to which speech unit a gesture belongs. Returning to the example given previously, the speaker uttering the sentence "*He tightened the screw*" might produce the gesture stroke simultaneously with the verb of the sentence. In this example, local integration would refer to the interaction between the simultaneously conveyed visual information (i.e. the turning-movement gesture) and auditory information (the word *tightened*).

Although related to such local processes, the global integration of gesture and speech is a more complex phenomenon. Understanding a gesture-supported sentence relies not only on the comprehension of all individual constituents (i.e.: words and gestures), but also on a comprehension of how the constituents are related to one another (i.e.: who is doing what to whom, cf. Grodzinsky and Friederici, 2006). This relational process requires integrating information over time. The multimodal aspect in this integration over time is the extent to which the process recruits similar or different brain areas depending on whether the to-be-integrated information is a spoken word or a gesture. Thus, local integration refers to an instantaneous integration across modalities, whereas global integration describes an integration over time, with modality as a moderating variable. Whereas interactions at the global level can be examined in an epoch-related analysis, an analysis of gesture–speech interactions at the local level can only be performed in an event-related design. More precisely, in order to investigate how gesture and speech interact at the local level, one first has to objectively identify the point in time at which gesture and speech start to interact. As will be outlined below, the gating paradigm may be used to determine such a time point.

Willems et al. (2006) investigated the neural correlates of gesture–speech interaction on a global sentence level. In this experiment, subjects watched videos in which an initial sentence part (e.g. *The items that he on the shopping list*[1]) was followed by one of four possible continuations: (1) a correct condition, where both gesture and speech matched the initial sentence context (e.g. saying *wrote* while producing a writing gesture), (2) a gesture mismatch (e.g. saying *wrote* while producing a hitting gesture), (3) a speech mismatch (e.g. saying *hit*, gesturing writing) and (4) a double mismatch (saying *hit*, gesturing hitting). In the statistical analysis, the complete length of the videos was modeled as an epoch. When contrasted with the correct condition, only the mid- to anterior portion of the left IFG (BA 45/47) was consistently activated in all three mismatch conditions. On the basis of this finding, Willems and co-workers suggested that the integration of semantic information into a previous sentence context (regardless whether the to-be-integrated information had been conveyed by gesture or speech) is supported by the left IFG.

Whereas the Willems study investigated the interaction of gesture and speech at a global level, it is an open issue what brain areas are involved in local gesture–speech interactions. One candidate area might be the superior temporal sulcus (STS). There is a substantial amount of literature supporting the notion of the STS as an important integration site of temporally synchronized audiovisual stimuli (Beauchamp, 2005). For example, the STS seems to

[1] The example is a literal translation of the original Dutch stimuli.

be involved in the integration of lip movements and speech sounds (Calvert et al., 2000; Wright et al., 2003). Furthermore, Skipper et al. (2005) observed that the activation in the posterior STS elicited by the observation of talking faces is modulated by the amount of visually distinguishable phonemes. In an experiment by Sekiyama et al. (2003) it was found that the left posterior STS is particularly involved in the McGurk effect, e.g. the fusion of an auditory /ba/ and a visual /ga/ into a perceived /da/. While in these examples, visual and auditory information can be mapped onto each other on the basis of their form, there is evidence that the STS is also involved in more complex mapping processes at a higher semantic level, such as the integration of pictures of animals and their corresponding sounds (Beauchamp et al., 2004b). Saygin et al. (2003) have reported that patients with lesions in the posterior STS are impaired in their ability to associate a picture (e.g. a cow) with a corresponding sound (e.g. *moo*).

On the basis of these results, it is not unreasonable to assume that the STS is also involved in the multimodal interactions between gesture and speech. The integration of iconic gestures and speech during comprehension has some similarities with the integration of pictures and their associated sounds, as it was for instance investigated by Beauchamp et al. (2004b). In both cases, there is a temporal synchrony between auditory and visual information. In the audiovisual condition of the Beauchamp study, the pictures and the corresponding sounds were presented simultaneously. Likewise, as it has been introduced above, the stroke of a gesture tends to coincide with the relevant speech unit. Another similarity is that for both stimulus types, what is being integrated are not the forms of gesture and speech (or the forms of pictures and sounds), but the interpretations of the respective forms. That is, in both cases the integration is said to occur on a semantic-conceptual level. The stimuli used in the Beauchamp study required participants to identify the depicted visual object (e.g. *a telephone*). On basis of their world knowledge, participants could then activate a number of possible sounds associated with the perceived visual object and decide whether the currently perceived sound matched one of these expectations.[2] Similarly, an iconic gesture first has to be processed unimodally to some extent before it can be associated with the co-expressive speech unit. Thus, the interactions of pictures and sounds and gesture and speech have in common that the unimodal information first has to be processed and semantically interpreted to some extent individually, before an interaction between auditory and visual information can occur. However, the two audiovisual interaction types differ in complexity. In the Beauchamp study, the semantic relationship between auditory and visual information was fixed. The visual object was always presented with the sound that such an object typically creates. In contrast, the semantic relationship between iconic gestures and speech is not fixed. A sentence such as *During the game, he returned the ball* can be accompanied by a gesture that depicts the form of the ball, or a gesture that focuses on the returning motion. Moreover, the gesture might primarily depict the trajectory of the ball's movement, the manner (rolling, sliding, ...) or a combination of trajectory and manner. Finally, the gesture can depict the scene from a character viewpoint (i.e. the person returning the ball) or from an observer viewpoint. How the gesture is related to speech is not defined a-priori, but has to be detected by the listener on an ad-hoc basis. Thus, the comprehension of iconic gestures requires complex semantic interactions between gestural and auditory information. So far there are no studies that have investigated whether the STS also houses these complex multimodal processes underlying co-speech iconic gesture comprehension.

[2] The sequential description of information flow from visual to auditory information suggested here is only for illustration. After an initial unimodal processing phase, there is probably a continuous interaction in both directions between auditory and visual information in the processing of such complex stimuli. The same holds true for the interaction between gesture and speech.

### *The present study*

The present experiment aimed to locate brain areas involved in the processing of co-speech iconic gestures. As has been described above, one can approach the comprehension of iconic gestures from a multimodal perspective. Investigating the putative multimodal integration sites for gesture and speech would entail an experimental design with a gesture-only, speech-only as well as a bimodal gesture+ speech condition, as it was for instance suggested by Calvert and Thesen (2004). However, the problem with such a manipulation is that it neglects the one-sided dependency between the two information channels. Whereas understanding speech does not depend on gesture, iconic gestures are dependent upon the accompanying speech in that these gestures are only distinctly meaningful in their co-speech context. It is generally agreed upon that the meaning of decontextualized iconic gestures is very imprecise (e.g. Cassell et al., 1999; Krauss et al., 1991). Thus, when presenting a gesture-only condition to participants, one runs a great risk of inducing artefactual processing strategies. As McNeill has stated: "It is profoundly an error to think of gesture as a code or 'body language', separate from spoken language. [...] It makes no more sense to treat gestures in isolation from speech than to read a book by looking at the 'g's." (McNeill, 2005, p. 4). Another independent group of researchers around Robert Krauss have also come to the conclusion that decontextualized iconic gestures convey little meaning to the listener and that the relationship between auditory, visual and audiovisual information is not well captured by a linear model (Krauss et al., 1991, Experiment 3).

Rather than adopting a strict multisensory perspective, the present study approaches the comprehension of co-speech iconic gestures by means of a disambiguation paradigm, where lexically ambiguous sentences (e.g. *Sie berührte die Maus, She touched the mouse*) are accompanied either by disambiguating iconic gestures or meaningless grooming movements. Such a disambiguation paradigm has several advantages. First, it has some external validity. Holler and Beattie (2003) have shown that speakers spontaneously produce a substantial amount of iconic gestures when asked to explain the different word meanings of a homonym. Second, in a disambiguation paradigm, the iconic gestures are not removed from their co-speech context, which excludes the possibility of a gesture-only condition inducing artefactual processing strategies. Third, the influence of the speech channel, which is certainly the channel with the highest information content, is perfectly controlled for, because the sentences are physically identical in the critical experimental conditions.

Thus, all of the observed differences in a disambiguation paradigm can only be due the accompanying hand movement (i.e. the main effect) or the interaction between the hand movement and the spoken sentence. The challenge in interpreting the results is to determine which one – main effect or interaction – actually caused an observed activation difference. One can think of the present study as an exploratory study in the evolving field of co-speech gesture

comprehension. It identifies regions possibly involved in the interaction between iconic gestures and speech in a paradigm with a high external validity.

In the present experiment, only the gestures but not the meaningless grooming movements bias the interpretation of the sentence, resulting in a disambiguation of the homonym. That is, only in the case of gesture there is an interaction between the visually and the auditorily conveyed information. On the basis of the literature, we hypothesized that the processing of co-speech gestures would elicit greater levels of activation in the STS than the processing of the meaningless co-speech grooming movements.

To elucidate the role of the left IFG (i.e. BA 44, 45 & 47) in local gesture–speech interactions, we additionally included a manipulation of word meaning frequency in the present study. All sentences could either be interpreted in terms of a more frequent dominant meaning (e.g. the *animal* meaning of *mouse*) or the less frequent subordinate meaning (e.g. the *computer device* meaning of *mouse*). Because previous studies have shown that the processing of lexically low frequent words recruits the left IFG to a stronger degree than high frequent words (Fiebach et al., 2002; Fiez et al., 1999; Joubert et al., 2004), we hypothesized that the processing of subordinate gestures would elicit greater levels of activation in the left IFG than dominant gestures. Alternatively, if the left IFG (and in particular the anterior inferior portion) is not only the site of multimodal gesture–speech interactions at the global level, as suggested by Willems et al. (2006), but also at the local level, greater levels of activation for gestures as compared to grooming should be observed in this region.

## Materials and methods

### *Participants*

Seventeen native speakers of German (10 females), age 21–30 (mean age 25.7, *SD*=2.8) participated in this experiment after giving informed written consent following the guidelines of the Ethics committee of the University of Leipzig. All participants were right-handed (mean laterality coefficient 92.7, SD=11.3, Oldfield, 1971) and had normal or corrected-to-normal vision. None reported any known hearing deficits.

### *Materials*

#### *Homonyms*

The present study is based on a set of unbalanced German homonyms (for a description of how the set was obtained, see Gunter et al., 2003). Each of the homonyms had a more frequent dominant and a less frequent subordinate meaning, which shared identical phonological and orthographical surface features (e.g. *ball* — dominant meaning: *game*; subordinate meaning: *dance*). Target words representing the dominant meaning as well as target words representing the subordinate meaning were assigned to each of the homonyms. The relatedness of the target words to the homonyms has been tested previously. In that visual priming experiment, participants made lexical decisions to targets that were preceded by homonym primes (see also Wagner, 2003). The results showed that the lexical decision time for each target word was significantly shorter as compared to an unrelated item.

A total of 52 two-sentence utterances were constructed. The utterances consisted of an introductory sentence introducing a character followed by a second sentence describing an action of that character. The second sentence always contained the homonym. The two-sentence utterances were constructed to be globally ambiguous, i.e. a given sentence could be interpreted both in terms of the dominant as well as the subordinate meaning (see Table 1. Further stimulus examples can be found in the online supplement).

#### *Gesture recording*

A professional actress was videotaped while uttering the sentences. The exact recording scenario was as follows. The actress stood in front of a video camera with her hands hanging comfortably in a resting position. In a first step, she memorized one two-sentence utterance until she could utter it fluently. Then she was asked to utter the sentence and simultaneously perform a gesture that supported one of the two possible meanings of the sentence. The gestures were created by the actress and not choreographed in advance by the experimenter. She was instructed to perform the gesture to coincide with the second sentence (e.g. *Sie berührte die Maus / She touched the mouse*) and to return her hands to the resting position afterwards. About two thirds of all gestures re-enacted the actions in the sentence from a first-person perspective (typing on a keyboard, swatting a fly, peeling an apple) while the remainder of gestures typically depicted salient features of objects (the shape of a skirt, the height of a stack of letters). To exclude the possibility that participants might use cues provided by the facial expression for disambiguation, the head of the actress was covered with a nylon stocking. All gestures resembling emblems or gestures directly related to the target words were excluded.

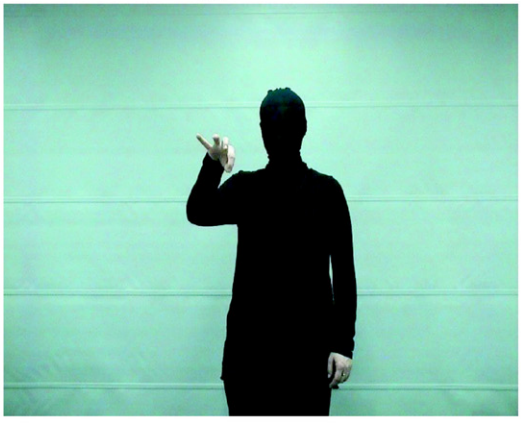
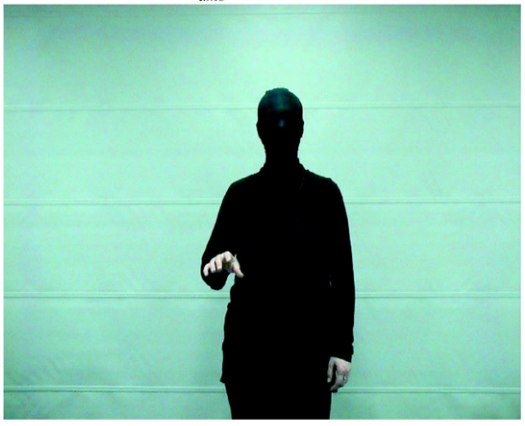
Additionally, the actress uttered each sentence set in combination with a simple grooming movement. The grooming was targeted at various body positions across the stimulus set (e.g. chin, ear, nose, back of head, chest, forearm, upper arm, stomach) and always coincided with the second sentence of each two-sentence utterance.

#### *Audio recording*

The speech of the sentences was re-recorded in a separate session to improve the sound quality. In order to maintain a comparable audiovisual synchrony between the three experimental conditions, the re-recorded audio was synchronized with the video stream according to the phonological synchrony rule, which states that "the stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech" (McNeill, 1992, p. 26). The exact procedure for combining the re-recorded sentences with the gesture videos was as follows. First, the recording of each sentence that seemed most compatible with both word meanings was selected. Next, the most strongly stressed syllable in the second sentence was determined. For instance, in the stimulus example this was the second syllable of the verb *berührte* (see Table 1). Following this, a video segment was combined with the selected re-recorded sentence so that the onset of the stroke of each hand movement (dominant gesture, subordinate gesture, grooming) coincided with the peak syllable. In the resulting audiovisual stream, the onset of the sentence always marked the onset of the video. Thus, each sentence set was realized by combining one audio recording with three different types of hand movements: (1) a gesture supporting the dominant meaning, (2) a gesture supporting the subordinate meaning, (3) a grooming movement.

Because the phonological synchrony rule was used, the gesture stroke coincided with different sentence positions across the stimulus set. While most stroke onsets coincided with the verb (59.5%) or the immediately preceding pronoun subject (32.5%),

Table 1
Stimulus examples

| | |
|---|---|
| Introduction: | Korinna streckte die Hand aus.<br>*Korinna reached her hand out.* |
| **Type of hand movement** | **Ambiguous sentence** |
| Dominant meaning | Sie berührte die $\text{Maus}_{amb}$<br>*She touched the* $\text{mouse}_{amb}$ |
| |  |
| Subordinate meaning | Sie berührte die $\text{Maus}_{amb}$<br>*She touched the* $\text{mouse}_{amb}$ |
| |  |
| Grooming | Sie berührte die $\text{Maus}_{amb}$<br>*She touched the* $\text{mouse}_{amb}$ |
| |  |

Introductory sentence was identical for all three conditions. Literal translation in italics.

only a smaller portion of strokes had their onset on or near the homonym (7.9%). Since the main focus of the present study is on the local interaction of gesture and speech, it is important to provide an explanation of how the gesture may locally (i.e., at the position of the verb) bias the sentence interpretation, although the to-be disambiguated homonym has not yet been processed. One way of conceptualizing the effect that the gestures may have locally on the interpretation of the verbs is in terms of adding selectional restrictions (Chomsky, 1965). For instance, in a sentence, a verb like *to phone* places selectional restrictions on the upcoming direct object, because the object must be something 'phone-able': A violation of the selectional restriction such as *He phoned the book* is not acceptable, and it has been proposed that during comprehension, listeners make immediate use of such selectional information (Tanenhaus and Trueswell, 1995).

Because the sentences of the present stimulus set were constructed to be compatible with both word meanings, the verbs themselves placed very little selectional restrictions on the direct object. Examples of such low-constraining verbs used in our stimuli are *to control, to touch, to observe, to adjust* and so on. However, in combination with a gesture, the verbs received some selectional restrictions. Consider for instance the sentence *Sie entfernte die Schale / She removed the bowl / peel* (see supplementary example items). The dominant gesture depicted carefully removing a fairly large bowl, whereas the subordinate gesture reenacted peeling an apple. In both cases, the stroke of the gesture coincided with the second syllable of the verb *entfernte*. Perceiving the dominant gesture in combination with the sentence fragment *Sie entfernte* places some selectional restrictions on the upcoming direct object. The listener already knows that an entity with a diameter of approx. 30 cm, which requires delicate handling, is being removed. This rules out the possibility that an apple's peel is being removed even before the homonym is encountered, and eventually results in a selection of the dominant word meaning of *Schale*. In this way, the gesture locally constrains the interpretation of the verb, which subsequently enables a disambiguation of the homonym.

### *Pre-test*

The selected video material was edited using commercial editing software (Final Cut Pro 5). A pre-test was conducted to assess how effective the gestures were in disambiguating the homonyms. In this pre-test, the videos were displayed to thirty German native speakers. At the offset of each video, the dominant and the subordinate target word were displayed on the screen. The participants had to select the target word that fit best into the previous video context. Gestures (and the corresponding homonyms) which did not elicit the selection of the correct target word in at least 50% of all subjects were excluded from the experimental set. In this final set of 42 homonyms, dominant gestures elicited a total of 88% correct responses (*SEM* 2.23) whereas subordinate gestures elicited a total of 85% correct responses (*SEM* 2.30). The difference was not significant ($t(1,82)=1.1$, $p>.27$). After a grooming video, participants selected the dominant meaning in 56% (*SEM* 4.44) of all cases. The meaning selection after grooming was not significantly different from chance level ($t(1,41)=1.3$, $p>.19$).

### *Gating*

The point in time at which the gesture information reliably contributed to selecting the appropriate meaning of the corresponding homonym was determined in a separate experiment using a

modified gating procedure (Grosjean, 1996).[3] In this experiment, forty native speakers of German watched the gesture video clips without sound in segments of increasing duration. A trial started with the visual presentation of a homonym. Subsequently, a segment of a gesture video was displayed. Finally, the two target words of the homonym representing the dominant and the subordinate meaning were displayed on the screen. The task of the participants was to determine whether the homonym referred to the dominant or the subordinate meaning based on gesture information. The gesture segment at which participants chose the correct meaning without any changes in response thereafter was determined as the gating point. In this way, we obtained the gating points for all gesture clips used in the current experiment.

*Procedure*

The experimental items were randomly divided into three blocks with the constraint that each homonym appeared only once within a block. Each block was then pseudo-randomized separately with the constraints that (1) no more than two consecutive videos belonged to the same condition and (2) the regularity with which one conditions followed another was matched. The experimental lists were assembled from all three blocks. All possible block orders were realized yielding a total of six experimental lists. These were distributed randomly across participants.

An experimental session consisted of three 10-min blocks. Blocks consisted of an equal number of trials and a matched number of items from each condition. Each session contained 168 trials, consisting of 126 critical trials (42 x each critical condition) plus 42 null events, in which no stimulus was presented and the BOLD response was allowed to return to a baseline state.

The length of the trials in the critical conditions depended on the length of the video clip and ranged from 9.92 sec to 15.08 sec (mean 11.0 sec, *SD* 0.55 sec). The length of the video clips did not differ significantly between the three experimental conditions ($F(2,123) < 1$). Each trial started with the presentation of a video clip. Following this, two target words were presented visually for 3000 ms and cued participants to judge which of the two words fit better into the context of the previous video clip. Participants held a response box in their right hand and were requested to push one of two buttons depending on the relatedness of the target words. The side on the screen at which the related target word was presented (left or right) was randomly determined for each trial. Hence, participants could not anticipate during the video which button they were to press in the upcoming response phase. Participants were allowed 3 s to respond to the target words. Performance rates and reaction times were recorded. Following the presentation of the target words, the trial was ended by the presentation of a fixation cross for 4000 ms.

Null events consisted of a continuous presentation of a fixation cross for 10500 ms.

*fMRI data acquisition*

Participants were placed in the scanner in a supine position. Visual stimuli (i.e. the videos and the subsequent target words) were presented on a computer screen outside of the scanner, which participants could see via mirror-glasses. Simultaneously with the videos, the corresponding sentences were presented via a set of specialized headphones (Resonance Technology Inc.) that attenuate the scanner noise about 30 dB. Furthermore, each participant wore ear plugs, which act as an additional low-pass filter. Before the experiment was conducted, the primary investigator (HH) tested whether the auditory sentences were clearly audible in the noisy scanner environment. Additionally, each participant was questioned after the experiment whether all the stimuli had been clearly audible and visible. Nobody reported any problems.

Eighteen axial slices (4 mm thickness, 1 mm inter-slice distance, FOV 19.2 cm, data matrix of 64 × 64 voxels, in-plane resolution of 3 × 3 mm) were acquired every 2 s during function measurements (BOLD sensitive gradient EPI sequence, TR = 2 s, TE = 30 ms, flip angle = 90, acquisition bandwidth = 100 Hz) with a 3 T Siemens TRIO system. Prior to functional imaging T1-weighted MDEFT images (data matrix 256 × 256, TR 1.3s, TE 10 ms) were obtained with a non-slice-selective inversion pulse followed by a single excitation of each slice (Norris, 2000).[4] These images were used to co-register functional scans with previously obtained high-resolution whole head 3D brain scans—128 sagittal slices, 1.5 mm thickness, FOV 25.0 × 25.0 × 19.2 cm, data matrix of 256 × 156 voxels (Lee et al., 1995).

*FMRI Data Analysis*

The accuracy data was analyzed by means of a repeated-measure ANOVA with the factor GESTURE_TYPE (dominant, subordinate) and BLOCK (1, 2, 3). The reaction time data was analyzed using a repeated-measure ANOVA with the factors MOVEMENT_TYPE (dominant gesture, subordinate gesture, grooming) and BLOCK (1, 2, 3). Greenhouse–Geisser correction was applied where appropriate. In these instances, we report the uncorrected degrees of freedom, the correction factor ε and the corrected *p* value.

The functional imaging data was processed using the software package LIPSIA (Lohmann et al., 2001). Functional data were motion-corrected offline with the Siemens motion correction protocol (Siemens, Erlangen, Germany). Data were subsequently corrected for the temporal offset between slices acquired in one scan using a cubic-spline interpolation based on the Nyquist–Shannon-Theorem. Low-frequency signal changes and baseline drifts were removed by applying a temporal highpass filter to remove frequencies lower than 1/120 Hz. A spatial Gaussian filter with 8 mm FWHM was applied.

To align the functional dataslices with a 3D stereotactic coordinate reference system, a rigid linear registration with six degrees of freedom (3 rotational, 3 translational) was performed. The rotational and translational parameters were acquired on the basis of the MDEFT-T1 (Norris, 2000) and EPI-T1 slices to achieve an optimal match between these slices and the individual 3D reference data set,

[3] Please note that the gating experiment will be part of a separate publication (Holle, Gunter & Obermeier, in preparation). Here we use the information obtained from gating to enable a more precise statistical modeling of the experimental video clips.

[4] MDEFT (modified driven equilibrium Fourier transform) refers to the pulse sequence used to obtain the T1-weighted images prior to functional scans and was originally developed by Ugurbil et al. (1993). In comparison with the manufacturer's pulse sequence for T1-weighted images (MP-RAGE), MDEFT as it is implemented in our institute has the advantage that the contrast is less dependent on the rf field. In MDEFT, each image has identical contrast and point spread function (Norris, 2000). The reduced power multislice MDEFT imaging sequence by Norris (2000) can be used at MRI scanners of several vendors (Siemens and Bruker) and it is therefore part of the standard protocol in our institute.

which was acquired during a previous scanning session. The MDEFT-T1 volume data set with 160 slices and 1 mm slice thickness was standardized to the Talairach stereotactic space. The rotational and translational parameters were subsequently transformed by linear scaling to a standard size. The resulting parameters were then used to transform the functional slices using trilinear interpolation, so that the resulting functional slices were aligned with the stereotactic coordinate system. The transformation parameters obtained from the normalization procedure were subsequently applied to the functional data. Voxel size was interpolated during co-registration from $3 \times 3 \times 4$ mm to $3 \times 3 \times 3$ mm.

Because we were interested in the neural correlates of the interaction between gesture and speech, we modeled the gating points of the gestures as determined in the gating experiment as an event. In the case of grooming the mean gating point of the dominant and subordinate gesture of a sentence-triplet was used as event. Note that the gating point did not differ significantly between the three experimental conditions ($F(2,123)<1$). The design matrix was generated with a synthetic hemodynamic response function (Friston et al., 1998; Josephs et al., 1997). The subsequent statistical analysis was based on a linear model with correlated errors (Worsley et al., 2002). Trials which were followed by an incorrect response were excluded from the statistical analysis.

For each participant three contrast images were generated: (1) Dominant gestures vs. Grooming, (2) Subordinate gestures vs. Grooming, (3) Subordinate gestures vs. Dominant gestures. Because individual functional datasets had been aligned to the standard stereotactic reference space, a group analysis based on the contrast images could be performed. Single-participant contrast images were entered into a second-level random effects analysis for each of the contrasts. The group analysis consisted of a one-sample *t*-test across the contrast images of all subjects that indicated whether observed differences between conditions were significantly distinct from zero. Subsequently, t-values were transformed into *Z*-scores. To protect against false positive activation a double threshold was applied, by which only regions with a *Z*-score exceeding 3.09 ($p<0.001$, uncorrected) and a volume exceeding 12 voxels (324 $mm^3$) were considered. This non-arbitrary voxel cluster size was determined by using the program AlphaSim (Ward, 2000) and is equivalent to a significance level of $p<0.05$ (corrected). Larger clusters of activation were checked for the existence of local maxima. A voxel was defined to be a local maximum if its *z*-value exceeded 3.09, if it was largest within a 12 mm radius and if the local volume of spatially contiguous activated voxels exceeded the cluster size threshold of 324 $mm^3$.

The time course of MR signal intensity was extracted for the most significant voxel of each cluster for each individual participant. Percent of signal change was calculated by dividing the MR signal by the constant of the linear model. Because the BOLD response typically peaks 6 s after stimulus onset, we decided on the basis of the mean percent signal change between 4 and 8 s post stimulus onset whether a given activation difference was due to either a positive or a negative BOLD response.

### *Motion tracking*

Because of the possibility that some of the observed activations might be partially driven by kinematic differences between conditions, a post-hoc analysis of the amount of hand motion in the video clips was conducted in the following way. First, the position of the right hand was manually marked in each video frame. More precisely, the pixel coordinate of the junction point between index finger and thumb was recorded (or estimated, if occluded from sight). Next, the procedure was repeated for the left hand. Subsequently, the euclidian distance between adjacent frames was calculated, yielding the mean amount of distance traveled by the hand. For each video, the mean across both hands was modeled as an epoch into the design matrix. The parametric effect of hand motion vs. the constant of the linear model, as well as the potential impact of hand motion on the data are presented at the end of Results.

## Results

### *Behavioral results*

Accuracy of responses and reaction times were recorded during the functional measurement. Here, we first report the accuracy data, following by the reaction time data.

In general, participants reliably selected the intended target word after both the dominant as well as the subordinate gesture videos (dominant: 91.6% correct, subordinate: 88.2% correct, see Fig. 1). Differences in the performance of participants were analyzed in a repeated-measures ANOVA with the dependent variable performance rate and the independent variables GESTURE_TYPE (dominant, subordinate) and BLOCK (1, 2, 3). The ANOVA yielded a significant main effect of BLOCK ($F(2,32)=5.4$; $\varepsilon=0.83$; $p<0.05$) indicating that accuracy increased across the experimental run. The main effect of GESTURE_TYPE ($F(1,16)=3.0$; $p=0.10$) and the interaction between GESTURE_TYPE and BLOCK ($F(2,32)<1$) were not significant.

Because there was no correct response possible after grooming videos, these data were analyzed separately. Overall, participants selected the dominant target word after 54.0% (*SEM* 1.87) of all grooming videos. The corresponding ANOVA indicated that the selection of dominant target word was significantly above chance level ($F(1,16)=4.4$; $p=0.05$). No other effects of this ANOVA were significant.

The reaction time data (see Fig. 2) was analyzed in a repeated-measures ANOVA with the factors MOVEMENT_TYPE (dominant gesture, subordinate gesture, grooming) and BLOCK (1, 2, 3). The ANOVA yielded a significant main effect of BLOCK ($F(2,32)=28.00$; $\varepsilon=0.77$; $p<0.0001$), indicating that the reaction time

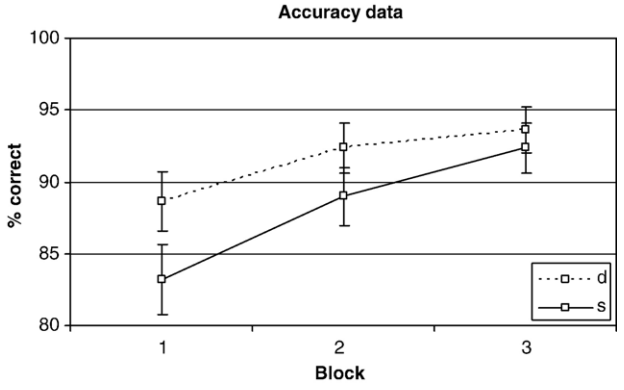



Fig. 1. Percentage of correctly selected target words for dominant and subordinate gestures. The dotted line represents responses following dominant gestures, the straight line responses following subordinate gestures. The error bars indicate the standard error of the mean.

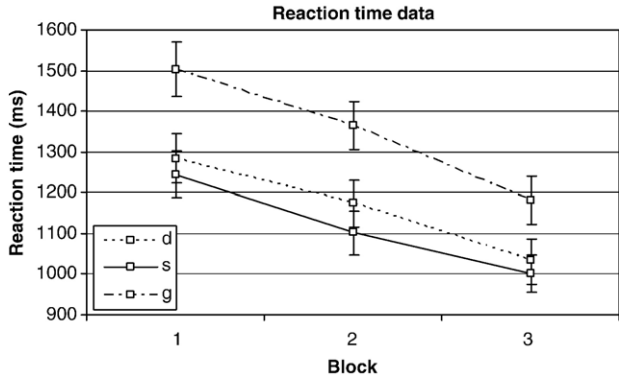



Fig. 2. Mean reaction time in ms for dominant gestures, subordinate gestures and grooming.

decreased over the experimental run. Additionally, a significant main effect of MOVEMENT_TYPE ($F(2,32)=35.24$; $\varepsilon=0.79$; $p<0.0001$) was observed. The interaction between MOVEMENT_TYPE and BLOCK was not significant ($F(4,64)<1$). Bonferroni-corrected post-hoc tests were performed to further investigate the main effect of MOVEMENT_TYPE. These tests indicated that the reaction time was significantly shorter for the dominant gestures as compared to grooming ($F(1,16)=37.35$; $p_{Bon}<0.0001$) and also significantly shorter for the subordinate gestures as compared to grooming ($F(1,16)=44.75$; $p_{Bon}<0.0001$). The difference between dominant and subordinate gestures was not significant ($F(1,16)=5.31$; $p_{Bon}=0.10$).

### *Imaging results*

#### *Dominant gestures vs. grooming*

The processing of dominant gestures vs. grooming elicited greater levels of activation in the left temporo-occipital cortex. The two local maxima of this activation were found in the posterior STS (see Table 2, Fig. 3) and the lateral part of the middle occipital gyrus.

Increased levels of activation for dominant gestures as compared to grooming were also found in the inferior parietal lobule (BA 40) bilaterally and in the precentral sulcus bilaterally. Additionally, activations in the medial part of the left middle occipital cortex, the medial part of the left middle frontal gyrus (BA9), the right intraparietal sulcus and in the right fusiform gyrus were observed.

In the reverse contrast (grooming>dominant gestures), greater levels of activations were observed in the putamen bilaterally.

#### *Subordinate gestures vs. grooming*

The processing of subordinate gestures as compared to grooming was associated with increased activation in the left temporo-occipital cortex (see Table 2, Fig. 4). The two local maxima of this activation were located in the posterior STS and the temporo-occipital junction. Additionally, increased activation was observed in the inferior parietal lobule (BA 40) bilaterally and the left fusiform gyrus. Upon reducing the activation threshold minimally ($Z>2.58$; $p<0.005$), it immediately became apparent that differences in the precentral sulcus bilaterally as well as the right fusiform gyrus were present in this contrast as well.

In the reverse contrast (grooming>subordinate gestures), increased levels of activations were found in the putamen bilaterally, the right middle frontal gyrus and the left anterior cingulate gyrus.

#### *Subordinate gestures vs. dominant gestures*

There was no increased activation for subordinate gestures vs. dominant gestures (Sub>Dom). However, in the reverse contrast (Dom>Sub) we observed a significant activation difference in the

Table 2
List of significantly activated regions

| Contrast | Region | $Z_{max}$ | Extent ($mm^3$) | $x$ | $y$ | $z$ |
|---|---|---|---|---|---|---|
| D>G | Left medial middle frontal gyrus | 3.95 | 1026 | −8 | 45 | 36 |
| | Right precentral sulcus | 3.98 | 1458 | 49 | 3 | 36 |
| | Left precentral sulcus | 3.93 | 972 | −47 | 3 | 33 |
| | Right inferior parietal lobule | 4.35 | 1728 | 58 | −36 | 30 |
| | Left inferior parietal lobule | 4.16 | 1215 | −59 | −36 | 33 |
| | Right intraparietal sulcus | 4.03 | 351 | 34 | −39 | 42 |
| | Right fusiform gyrus | 4.03 | 864 | 37 | −48 | −6 |
| | Left temporo-occipital cortex | | 3699 | | | |
| | Posterior STS | 4.08 | | −50 | −54 | 15 |
| | Lateral middle occipital gyrus | 4.2 | | −38 | −75 | 24 |
| | Left medial middle occipital gyrus | 3.51 | 1107 | −8 | −96 | 9 |
| G>D | Left putamen | −3.93 | 648 | −20 | 9 | 3 |
| | Right putamen | −4.25 | 972 | 19 | 9 | 3 |
| S>G | Left precentral sulcus* | 3.17 | 594 | −47 | 6 | 27 |
| | Right precentral sulcus* | 3.19 | 594 | 43 | 0 | 27 |
| | Right inferior parietal lobule | 4.54 | 594 | 55 | −24 | 39 |
| | Left inferior parietal lobule | 4.38 | 837 | −56 | −36 | 33 |
| | Right fusiform gyrus* | 3.29 | 513 | 40 | −48 | −9 |
| | Left fusiform gyrus | 3.96 | 378 | −41 | −51 | −9 |
| | Left temporo-occipital cortex | | 3861 | | | |
| | Posterior STS | 4.03 | | −47 | −57 | 12 |
| | Occipito-temporal junction | 4.66 | | −53 | −72 | 12 |
| G>S | Right middle/inferior frontal gyrus | −3.59 | 486 | 46 | 39 | −9 |
| | Left cingulate gyrus | −4.51 | 783 | −11 | 18 | 30 |
| | Left putamen | −4.65 | 3888 | −23 | 6 | 0 |
| | Right putamen | −4.12 | 3834 | 19 | 3 | −3 |
| S>D | No significantly activated regions | | | | | |
| D>S | Left lateral middle frontal gyrus↓ | −3.45 | 378 | −35 | 24 | 36 |
| | White matter | −5.03 | 459 | −20 | 42 | 12 |
| | White matter | −4.28 | 2268 | 28 | −60 | 24 |
| Parametric Effect of Hand Motion | | | | | | |
| | Left inferior temporal sulcus | 3.96 | 459 | −38 | −66 | 15 |
| | Right temporo-occipital junction | 3.91 | 1188 | 37 | −69 | 6 |
| | Left lingual gyrus | 4.28 | 1539 | 16 | −75 | 0 |
| | Right cuneus/precuneus | 4.67 | 4536 | 10 | −84 | 45 |
| | Left cuneus | 4.24 | 3159 | −11 | −99 | 15 |

Results of fMRI experiment. Abbreviations: D=Dominant gesture; S=Subordinate gesture; G=Grooming; STS=Superior temporal sulcus. Significance threshold $p<0.001$ (uncorrected); cluster size threshold 324 $mm^3$. Activations marked by * are significant at a $p<0.005$ (uncorrected). The activation marked by ↓ was due to a negative BOLD response (see also online supplementary Figures), all other activations were found to be due to positive BOLD responses (see also Figs. 3 and 4).

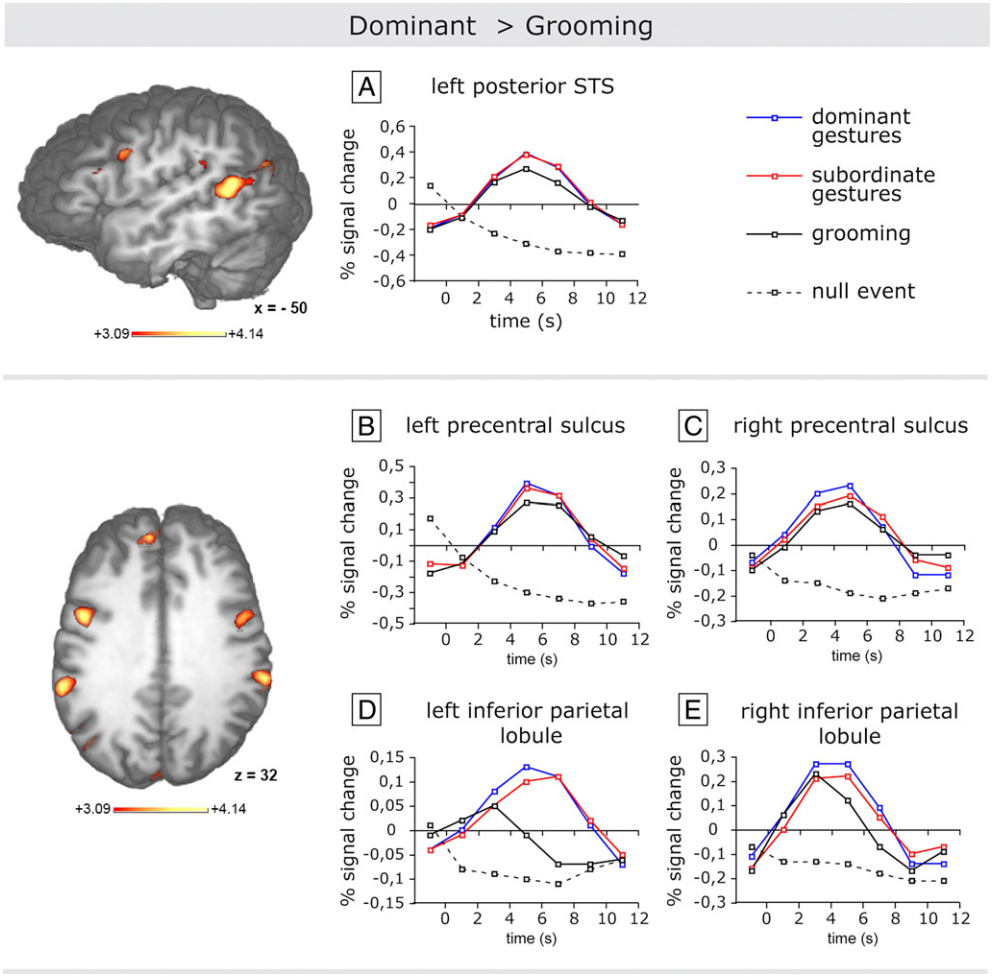



Fig. 3. Illustration of brain regions showing an increased BOLD response to dominant gestures as compared to grooming. Time-courses are given for the most significant voxel of each cluster (for the Talairach coordinates of the voxels, see Table 2).

left lateral middle frontal gyrus (BA9). The corresponding time-course analysis revealed that this difference was due to a negative BOLD response in the time range from 4 to 8 s which was stronger in the case of the subordinate gesture (see online supplementary material).

Because the hypothesis for the present study specifically targeted the left IFG and the posterior STS, we additionally checked whether there were activation differences at a reduced significance threshold present in these brain areas ($Z>2.58$, $p<0.005$). No differences were observed when directly contrasting the two gesture types (neither for Sub>Dom nor for Dom>Sub).

Additionally, we checked whether the processing of gestures as compared to grooming yielded significant activation differences in the left anterior inferior IFG, because this brain area has been suggested to support the global integration of gesture and speech (Willems et al., 2006). The time course of MR signal intensity was extracted from a spherical ROI of 10 mm diameter around the center coordinate of the IFG activations reported in the Willems study (Talairach coordinates: −43, 11, 26 Willems et al., 2006, their Fig. 3b). The mean percent signal change between 4 and 8 s was analyzed as dependent variable in a repeated-measures ANOVA with the factor MOVEMENT_TYPE (dominant, subordinate, grooming). The main effect of MOVEMENT_TYPE was not significant ($F(2,32)<1$).

*Effect of hand motion*

Because some of the reported activations fall within areas that are associated with motion processing (Culham et al., 2001), we performed a post-hoc analysis of the amount of hand motion in the video clips. In a first step, we wanted to know whether the measure based on the pixel coordinates (see Materials and methods) is a valid indicator of brain activity related to hand motion in the video sequences. To this end, the mean amount of hand motion for each video was modeled as an epoch in the design matrix. As can be seen from the results (see Table 2, Fig. 5), this contrast yielded reliable activations in areas tightly associated with motion processing, including the lingual gyrus, cuneus and precuneus as well the right temporo-occipital junction, probably corresponding to the human homologue of the monkey MT complex (hMT+, see Culham et al., 2001). Thus, the variable seems to be a valid indicator for brain activity related to motion in video sequences (Dupont et al., 1997; Grill-Spector and Malach, 2004).

In a next step, we tested for differences between conditions by subjecting the mean amount of hand motion across both hands for each video to a repeated measures one-way ANOVA with the factor MOVEMENT_TYPE (dominant gesture, subordinate gesture, grooming). A significant main effect of MOVEMENT_TYPE was found ($F(2,82)=1.12$; $p<0.0001$; $\varepsilon=0.95$). Bonferroni-corrected post-hoc tests indicated that the dominant gesture videos

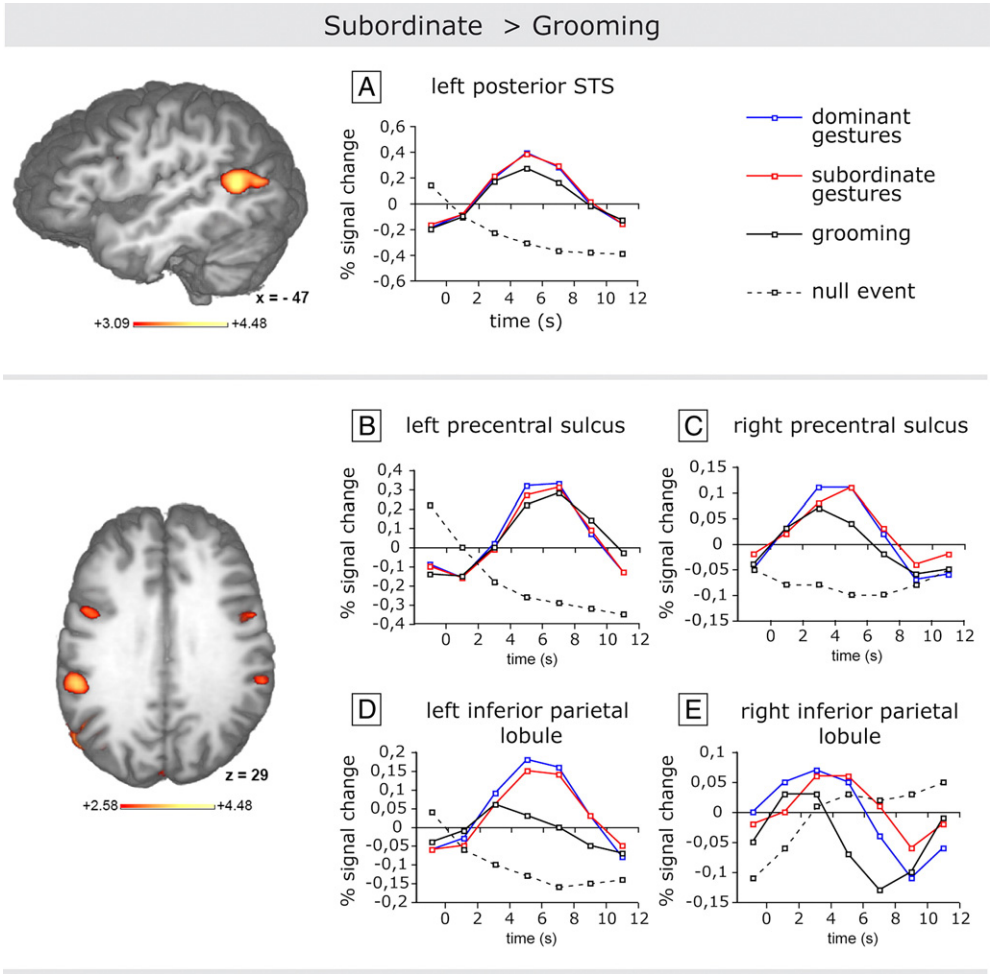



Fig. 4. Illustration of brain regions showing an increased BOLD response to subordinate gestures as compared to grooming.

contained more hand motion than the grooming videos ($F(9,41)=9.04$, $p_{Bon}<0.05$). Similarly, the subordinate gestures contained more hand motion than grooming ($F(1,41)=21.92$, $p_{Bon}<0.0003$. The difference between dominant and subordinate gestures was not significant ($F(1,41)=2.25$, $p_{Bon}>0.42$).

Upon this discovery, the fMRI data was analyzed again, using only a subset of 99 videos (33 per condition), that on average did not differ significantly between conditions in the amount of hand motion, the video length and the position of the gating point (all $F(2,96)<1.89$, all $p>.16$). All of the activations discussed below were replicated in this matched subset of items (For a table comparing the fMRI results for both the complete as well as the matched subset, please see the online supplementary material).

## Discussion

The present study investigated the neural correlates of the processing of co-speech gestures. Sentences containing an unbalanced

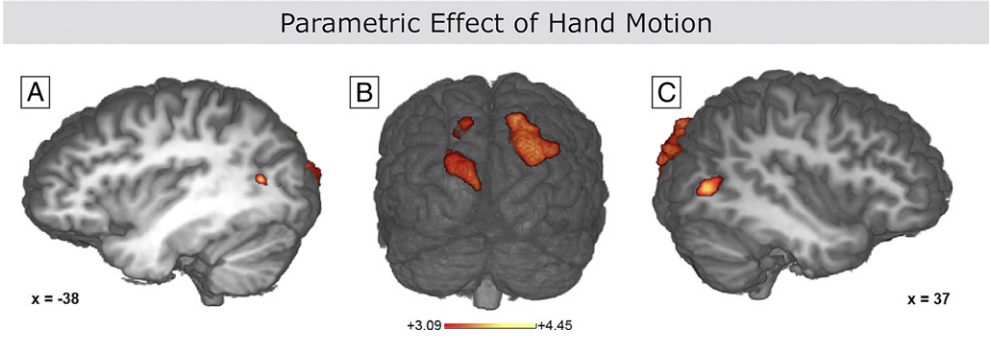



Fig. 5. Illustration of brain regions showing a BOLD response that parametrically varied as a function of the amount of hand motion in the videos. (A) Left inferior temporal sulcus (B) cuneus/precuneus bilaterally (C) right temporo-occipital junction (hMT+).

ambiguous word were accompanied by either a meaningless grooming movement, a gesture supporting the more frequent dominant meaning or a gesture supporting the less frequent subordinate meaning. We had two specific hypotheses in mind when designing this experiment. First, we expected that the STS would be more involved in the processing of co-speech gestures than in the processing of co-speech grooming movements, because only in the case of gesture-supported sentences, there was a local interaction between auditory and visual information. Second, it was hypothesized that the processing of subordinate gestures would recruit the left IFG to a stronger degree than dominant gestures. We found support for the first hypothesis, but the second hypothesis was not supported by our data. The main results are that when contrasted with grooming, both types of gestures (dominant and subordinate) activated an array of brain regions consisting of the left posterior STS, the inferior parietal lobule bilaterally and the ventral precentral sulcus bilaterally.

### *Behavioral data*

Before discussing the fMRI results, the behavioral data merit attention. Participants reliably selected the intended target word after the gesture trials, suggesting that gesture was able to bias the interpretation of the sentences. When the hand movements offered no helpful cue for the interpretation of the sentence (i.e., in the grooming condition), word meaning frequency had a significant (albeit small) influence on target word selection. These results are in line with our previous findings in showing that listeners use the information provided by iconic gestures to disambiguate speech (Holle and Gunter, 2007). In the absence of a cue for meaning selection, word meaning frequency influences which meaning of the ambiguous sentence is selected (Holle and Gunter, 2007, Exp. 3).

### *Imaging data*

#### *Gesture vs. grooming*

*STS.* When contrasted with grooming, the processing of both gesture types (dominant and subordinate) elicited greater levels of activation in the left posterior STS (see Table 2, Figs. 3 and 4).

The human STS is known to be an important audiovisual integration site (Beauchamp, 2005). For example, the McGurk-Effect is associated with increased levels of activation in the left posterior STS (Sekiyama et al., 2003). The STS was also found to be crucial for the integration of letters and speech sounds (van Atteveldt et al., 2004), pictures and sounds (Beauchamp et al., 2004b) as well as videos of tool actions and their corresponding sounds (Beauchamp et al., 2004a). This suggests a rather broad spectrum of audiovisual integration processes that recruit this brain area. In the present study, the local maxima in the posterior STS for dominant and subordinate gestures are in close proximity to those coordinates reported for the integration of lip movements and speech (Calvert et al., 2000; Sekiyama et al., 2003). Given the interactive nature of iconic gestures (i.e. their dependency on a co-speech context in order to become distinctly meaningful), the increased activation for gestures vs. grooming observed in the left posterior STS is suggested to reflect the interaction of gesture and speech in comprehension. Because a gesture has to be interpreted to some extent before it can be associated with its co-speech unit, the interaction has to occur on a semantic level. In the present stimulus set, most local gesture–speech interactions occurred between the gesture and the verb of the ambiguous sentence. As has been argued previously, a likely way in which the gestures biased the interpretation of the sentence locally is by imposing selectional restrictions on the co-expressive verb. The combined information of verb and gesture enabled later on the disambiguation of the homonym. The multimodal matching of co-expressive gesture and speech is suggested to yield increased activation in the posterior STS. In contrast, grooming did not interact in a meaningful way with the ambiguous sentence, hence the signal increase in the posterior STS is less pronounced.

Because the contrast is based on the comparison of different stimuli (gesture vs. grooming), it is in principle possible that the posterior STS activation primarily reflects differences in the stimuli kinematics (e.g. amount of motion). We have some reasons to believe that this is not the case. First, the average length of the videos did not differ between the three experimental conditions. Second, although the posterior STS has been found to be involved in the processing of biological motion, these activations have been characterized as being markedly right-lateralized (Pelphrey et al., 2003). In contrast, in the present study, we found greater levels of activation in the left posterior STS for gesture as compared to grooming, suggesting that the activation was not primarily driven by biological motion. Third, activation in the left posterior STS is not modified when analyzing a subset of items matched for the amount of hand motion (see Results). All in all, it seems therefore rather unlikely that this activation is driven by kinematic differences between gesture and grooming.

Another possible explanation of the posterior STS activation is that it reflects the difference of meaningful vs. meaningless hand movements (cf. Allison et al., 2000). However, as has been repeatedly stated in the literature, iconic gestures only become distinctly meaningful when accompanied by their co-speech context. There is a large variability in the meaning listeners attribute to decontextualized iconic gestures (Feyereisen et al., 1988; Hadar and Pinchas-Zamir, 2004; Krauss et al., 1991), therefore it is rather unlikely that the STS activation reflects the processing of gesture meaning *per se*.

Finally, the greater levels of activation for gesture vs. grooming might partially reflect a less attentive processing of the grooming videos. Participants may have, as soon as they realized that the sentence was accompanied by a grooming movement, put less effort on processing the stimulus and prepared themselves to respond at random. Such a strategy would result in shorter reaction times for grooming as compared to gesture. However, the reaction time after grooming was actually longer than after the gesture videos (see Results) suggesting that grooming videos were also processed attentively.

The processing of iconic gestures as compared to grooming did not elicit activation in the left anterior inferior IFG. Willems et al. (2006) have suggested that this brain area is involved in global integration processes at the sentence level. Although negative findings can occur for a variety of reasons, one possible explanation for the lack of activation in anterior inferior IFG in the present study is that the local integration of gesture and speech is anatomically distinct from global integration processes. The local integration of gesture and speech, presumably housed in the posterior STS, may be followed by an integration at the global level in the left IFG, where a supramodal representation of the sentence meaning is assembled from the individual meaningful parts of the sentence. Of course, other factors like the employed design (mismatch vs. disambiguation) or the type of analysis (epoch-related vs. event-

related) might also be a reason for the divergent findings between the present study and the study by Willems and co-workers. Clearly, further research is needed to determine the interplay of these two brain regions in the processing of co-speech gestures.

*Frontal and parietal activations.* When contrasted with grooming, both types of gestures elicited increased activation in the inferior parietal lobule (IPL, BA 40). Only dominant gestures additionally elicited greater levels of activation in the precentral sulcus (BA 6), extending anteriorly into BA 44. However upon reducing the activation threshold minimally ($Z>2.58$; $p<0.005$), it immediately became apparent that differences in this area were present for subordinate gestures bilaterally as well (see Table 2). Please note also that there were no significant differences in the right fusiform gyrus and the precentral sulcus observed when the two types of gesture were directly compared (dominant vs. subordinate) suggesting that the pattern of activation in the precentral sulcus and the right fusiform gyrus is not qualitatively different between dominant and subordinate gestures. Because they fall within the specified area, the activation peaks in the precentral sulcus are henceforth referred to as ventral premotor cortex (vPMC).[5]

The frontal and parietal brain regions in which the processing of co-speech gestures elicited increased levels of activation have been described in the literature as core components of the putative human mirror neuron system (Rizzolatti and Craighero, 2004). It has been demonstrated previously that planning as well as execution of transitive gestures (i.e. gestured movements involving an object) activates the left premotor cortex and left BA 40 (Fridman et al., 2006). Lipreading, another instance where speech-related visual information has to be analyzed, has been found to be correlated with increased activation in the left IFG (Paulesu et al., 2003). Given that the majority of iconic gestures in the present study re-enacted the actions described in the sentence, we interpret this system of fronto-parietal activations as an involvement of the mirror neuron system in co-speech gesture comprehension. However, in which way might the mirror neuron system support the integration of gesture and speech? One recent theoretical suggestion is that the mirror neuron system determines the goal of observed actions through an observation–execution matching process (Iacoboni, 2005; Iacoboni and Wilson, 2006). Translated to the context of the present experiment, determining the goal is equivalent to finding the answer to the following question: "Why did the speaker just produce this hand movement?". In the case of grooming, the answer would be "because she wanted to scratch herself". In the case of gesture (e.g. the clicking mouse gesture), the answer would be "because she wanted to show how the touching was done". In both cases, there is a goal that can be attributed to the observed hand movement. However, the process leading to goal attribution might be more costly in the case of gesture. According the action–observation matching model (Iacoboni, 2005), the goal of an observed action has been determined when the predicted sensory consequences of the internal motor simulation matches the observed visual input. When there is no match, because the initial goal hypothesis was incorrect, a new goal has to be generated which entails a new simulation cycle. Iconic gestures are undeniably more complex hand movements than grooming and the meaning of these gestures is inherently vague. Because of this, the goal initially attributed to a gesture probably not always turns out to be the correct one, therefore the total number of simulation cycles needed for gesture is presumably larger than for grooming. Thus, the greater levels of activation in vPMC and IPL for gesture vs. grooming might reflect greater "simulation costs" for the processing of gestures.

An alternative explanation for the activation in the precentral sulcus might be that participants used a verbalization strategy for the gestures but not for grooming. The observed activation extended anteriorly into BA 44, an area known to be involved in verbalization processes (e.g. Nixon et al., 2004). However, we think this explanation is rather unlikely for two reasons: First, it is probably difficult to employ a verbalization strategy when the gesture is embedded in a co-speech context, because the phonological loop is already busy with the processing of the sentence (Baddeley, 2002). Second, a verbalization account would actually predict increased left IFG activation for grooming, because the meaning of an iconic gesture is often difficult to name (Feyereisen et al., 1988) and it is probably easier to use a verbalizing strategy for the grooming movements (e.g. "scratch").

*Fusiform gyrus.* We found increased levels of activation in the right fusiform gyrus for dominant gestures vs. grooming. For subordinate gestures vs. grooming, significant activation in the fusiform gyrus was restricted to the left fusiform gyrus, however, at a lower significance threshold ($Z>2.58$, $p<0.005$), differences in the right fusiform gyrus were present for the subordinate gestures as well. It has been suggested that the fusiform gyrus supports the processing of complex visual stimuli for which visual expertise has been developed (Gauthier and Bukach, 2007; Tarr and Gauthier, 2000). In this view, the activation of the fusiform gyrus during face observation (Kanwisher et al., 1997) indicates that we are all experts in face processing. Participants who are experts in other domains, such as recognition of types of birds or cars, exhibited increased levels of activation in the fusiform region for those stimuli (Gauthier et al., 2000). Grooming movements tend to be very repetitive and most of them go unnoticed (Goldin-Meadow, 2003). The higher levels of activation for gestures vs. grooming in the right fusiform gyrus may therefore be due to the fact that participants have more expertise in the processing of gestures than in the processing of grooming movements.

*Subordinate>dominant.* We hypothesized that the processing of subordinate gestures would elicit increased levels of activation in the left IFG than dominant gestures because this brain area is known to be sensitive to semantic processing difficulties like word frequency (Fiebach et al., 2002; Fiez et al., 1999; Joubert et al., 2004). However, no significant differences in this brain area were observed. In light of the rather low amount of dominant target word selections after grooming videos (just above chance level), it is a possibility that word meaning frequency was not effectively varied in this study. Note, however, that in a number of previous experiments, the same set of homonyms elicited strong effects of word meaning frequency (Gunter et al., 2003; Holle and Gunter, 2007). For example, Holle and Gunter (2007) used sentences containing a homonym that were disambiguated at a target word later in the sentence (e.g. *She touched the mouse, which the cat / computer...*). Coincident with the initial part of the sentence, the speaker

[5] The anatomical border between ventral and dorsal premotor cortex is still a matter of debate (see, for instance, Schubotz, 2004.). Here, we follow the suggestion from Rizzolatti et al. (2002), who locate the border at the upper limit of the frontal eye field, corresponding to z=51 in Talairach space.

produced either a disambiguating gesture or a grooming movement. Following a grooming movement, the N400 time-locked to the target words was significantly larger at subordinate target words as compared to the dominant targets. This suggests that at the position of the target word, the subordinate word meaning was less active in working memory than the dominant meaning. Thus, in the absence of a gestural cue for meaning selection, word meaning frequency governed the selection process. Why did we not observe a similar effect of word meaning frequency in the present experiment, although the gestures and the homonyms were identical and the sentence structure was highly similar? One explanation might be the nature of the task in the present experiment (two-alternative forced-choice) which contrasts with the much more subtle measure of N400 amplitude in the experiment of Holle and Gunter (2007). Frequency effects are generally considered to influence the lexical access of a word. However, in the case of the present study, both word meanings of the homonym are explicitly presented to the participants (via the display of the two related target words during the response phase). Thus, there was no need for the participants to search their mental lexicon for the possible word meanings of the homonym and therefore little range for an effect of word meaning frequency to occur. Thus, we have some confidence in assuming that word meaning frequency was effectively manipulated in the present study, although the behavioral data suggest the opposite. Please note also that the statistical modeling of the fMRI data was performed at the gating point during the gesture video (see Methods) and not during the delayed response.[6]

## Conclusion

The present study investigated the neural correlates of co-speech gesture processing. The processing of speech accompanied by meaningful hand movements reliably activated the left posterior STS, possibly reflecting the multimodal semantic interaction between a gesture and its co-expressive speech unit. The processing of co-speech gestures additionally elicited a fronto-parietal system of activations in classical human mirror neuron system brain areas. The mirror neuron system is suggested to be involved in the decoding of the goal of observed hand movements through an observation–execution matching process.

## Acknowledgments

We are grateful to Angela Friederici, who kindly supported the research described here, Christian Obermeier, who has carried out large parts of the gating study and Karsten Müller, who was of great help during the analysis of the data.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2007.10.055.

[6] Another potential explanation for the lack of a frequency effect in inferior frontal areas might be that the fMRI data was modeled to the gating point, but the frequency of word frequency could only occur later on at the homonym. Note, however, that there was also no evidence for a frequency effect when modeling the data to the onset of the homonym.

## References

Alibali, M.W., Flevares, L.M., Goldin-Meadow, S., 1997. Assessing knowledge conveyed in gesture — do teachers have the upper hand. J. Educ. Psychol. 89, 183–193.

Allison, T., Puce, A., McCarthy, G., 2000. Social perception from visual cues: role of the STS region. Trends Cogn. Sci. 4, 267–278.

Baddeley, A.D., 2002. Is working memory still working? Eur. Psychol. 7, 85–97.

Beattie, G., Shovelton, H., 1999. Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. Semiotica 123, 1–30.

Beattie, G., Shovelton, H., 2002. An experimental investigation of some properties of individual iconic gestures that mediate their communicative power. Br. J. Psychol. 93, 179–192.

Beauchamp, M.S., 2005. See me, hear me, touch me: multisensory integration in lateral occipital–temporal cortex. Curr. Opin. Neurobiol. 15, 145–153.

Beauchamp, M.S., Argall, B.D., Bodurka, J., Duyn, J.H., Martin, A., 2004a. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat. Neurosci. 7, 1190–1192.

Beauchamp, M.S., Lee, K.E., Argall, B.D., Martin, A., 2004b. Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41, 809–823.

Binkofski, F., Buccino, G., 2006. The role of ventral premotor cortex in action execution and action understanding. J. Physiol. (Paris) 99, 396–405.

Calvert, G.A., Thesen, T., 2004. Multisensory integration: methodological approaches and emerging principles in the human brain. J. Physiol. (Paris) 98, 191–205.

Calvert, G.A., Campbell, R., Brammer, M.J., 2000. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr. Biol. 10, 649–657.

Cassell, J., McNeill, D., McCullough, K.-E., 1999. Speech–gesture mismatches: evidence for one underlying representation of linguistic and nonlinguistic information. Pragmat. Cogn. 7, 1–33.

Chomsky, N., 1965. Aspects of the theory of syntax. MIT Press, London.

Culham, J., He, S., Dukelow, S., Verstraten, F.A., 2001. Visual motion and the human brain: what has neuroimaging told us? Acta Psychol. (Amst.) 107, 69–94.

Dupont, P., De Bruyn, B., Vandenberghe, R., Rosier, A.M., Michiels, J., Marchal, G., Mortelmans, L., Orban, G.A., 1997. The kinetic occipital region in human visual cortex. Cereb. Cortex 7, 283–292.

Feyereisen, P., Van de Wiele, M., Dubois, F., 1988. The meaning of gestures: what can be understood without speech? Cah. Psychol. Cogn./Curr. Psychol. Cogn. 8, 3–25.

Fiebach, C.J., Friederici, A.D., Muller, K., von Cramon, D.Y., 2002. fMRI Evidence for dual routes to the Mental Lexicon in visual word recognition. J. Cogn. Neurosci. 14, 11–23.

Fiez, J.A., Balota, D.A., Raichle, M.E., Petersen, S.E., 1999. Effects of lexicality, frequency, and spelling-to-sound consistency on the functional anatomy of reading. Neuron 24, 205–218.

Fridman, E.A., Immisch, I., Hanakawa, T., Bohlhalter, S., Waldvogel, D., Kansaku, K., Wheaton, L., Wu, T., Hallett, M., 2006. The role of the dorsal stream for gesture production. NeuroImage 29, 417–428.

Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., Turner, R., 1998. Event-related fMRI: characterizing differential responses. NeuroImage 7, 30–40.

Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G., 1996. Action recognition in the premotor cortex. Brain 119, 593–609.

Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G., 2002. Action representation and the inferior parietal lobule. In: Prinz, W., Hommel, B. (Eds.), Common mechanisms in perception and action. Oxford Univ. Press, New York.

Gauthier, I., Bukach, C., 2007. Should we reject the expertise hypothesis? Cognition 103, 322–330.

Gauthier, I., Skudlarski, P., Gore, J.C., Anderson, A.W., 2000. Expertise for

cars and birds recruits brain areas involved in face recognition. Nat. Neurosci. 3, 191–197.
Goldin-Meadow, S., 2003. Hearing Gesture — How Our Hands Help Us Think. The Belknap Press of Harvard Univ. Press, Cambridge.
Grill-Spector, K., Malach, R., 2004. The human visual cortex. Annu. Rev. Neurosci. 27, 649–677.
Grodzinsky, Y., Friederici, A.D., 2006. Neuroimaging of syntax and syntactic processing. Curr. Opin. Neurobiol. 16, 240–246.
Grosjean, F., 1996. Gating. Lang. Cogn. Processes 11, 597–604.
Gunter, T.C., Bach, P., 2004. Communicating hands: ERPs elicited by meaningful symbolic hand postures. Neurosci. Lett. 372, 52–56.
Gunter, T.C., Wagner, S., Friederici, A.D., 2003. Working memory and lexical ambiguity resolution as revealed by ERPs: a difficult case for activation theories. J. Cogn. Neurosci. 15, 643–657.
Hadar, U., Pinchas-Zamir, L., 2004. The semantic specificity of gesture — Implications for gesture classification and function. J. Lang. Soc. Psychol. 23, 204–214.
Holle, H., Gunter, T.C., 2007. The role of iconic gestures in speech disambiguation: ERP evidence. J. Cogn. Neurosci. 19, 1175–1192.
Holler, J., Beattie, G., 2003. Pragmatic aspects of representational gestures: do speakers use them to clarify verbal ambiguity for the listener? Gesture 3, 127–154.
Iacoboni, M., 2005. Neural mechanisms of imitation. Curr. Opin. Neurobiol. 15, 632–637.
Iacoboni, M., Dapretto, M., 2006. The mirror neuron system and the consequences of its dysfunction. Nat. Rev., Neurosci. 7, 942–951.
Iacoboni, M., Wilson, S.M., 2006. Beyond a single area: motor control and language within a neural architecture encompassing Broca's area. Cortex 42, 503–506.
Josephs, O., Turner, R., Friston, K., 1997. Event-related fMRI. Hum. Brain Mapp. 5, 243–248.
Joubert, S., Beauregard, M., Walter, N., Bourgouin, P., Beaudoin, G., Leroux, J.M., Karama, S., Lecours, A.R., 2004. Neural correlates of lexical and sublexical processes in reading. Brain Lang. 89, 9–20.
Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311.
Kelly, S.D., Kravitz, C., Hopkins, M., 2004. Neural correlates of bimodal speech and gesture comprehension. Brain Lang. 89, 253–260.
Kita, S., Özyürek, A., 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. J. Mem. Lang. 48, 16–32.
Krauss, R.M., Morrel-Samuels, P., Colasante, C., 1991. Do conversational hand gestures communicate? J. Pers. Soc. Psychol. 61, 743–754 (November).
Lee, J.H., Garwood, M., Menon, R., Adriany, G., Andersen, P., Truwit, C.L., Ugurbil, K., 1995. High contrast and fast three-dimensional magnetic resonance imaging at high fields. Magn. Reson. Med. 34, 308–312.
Levelt, W.J.M., Richardson, G., la Heij, W., 1985. Pointing and voicing in deictic expressions. J. Mem. Lang. 24, 133–164.
Lohmann, G., Muller, K., Bosch, V., Mentzel, H., Hessler, S., Chen, L., Zysset, S., von Cramon, D.Y., 2001. LIPSIA — a new software system for the evaluation of functional magnetic resonance images of the human brain. Comput. Med. Imaging Graph. 25, 449–457.
McNeill, D., 1992. Hand and Mind — What Gestures Reveal about Thought. The University of Chicago Press, Chicago.
McNeill, D. (Ed.), 2000. Language and Gesture. Cambridge Univ. Press, Cambridge.
McNeill, D., 2005. Gesture and Thought. University of Chicago Press, Chicago and London.
Molnar-Szakacs, I., Kaplan, J., Greenfield, P.M., Iacoboni, M., 2006. Observing complex action sequences: the role of the fronto-parietal mirror neuron system. NeuroImage 33, 923–935.
Morrel-Samuels, P., Krauss, R.M., 1992. Word familiarity predicts temporal asynchrony of hand gestures and speech. Journal of Experimental Psychology: Learning. Mem. Cogn. 18, 615–622.
Nixon, P., Lazarova, J., Hodinott-Hill, I., Gough, P., Passingham, R., 2004. The inferior frontal gyrus and phonological processing: an investigation using rTMS. J. Cogn. Neurosci. 16, 289–300.
Nobe, S., 2000. Where do most spontaneous representational gestures actually occur with respect to speech? In: McNeill, D. (Ed.), Language and Gesture. Cambridge Univ. Press, Cambridge, pp. 186–198.
Norris, D.G., 2000. Reduced power multislice MDEFT imaging. JMRI—J. Magn. Reson. Imaging 11, 445–451.
Oldfield, R.C., 1971. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia 9, 97–113.
Özyürek, A., Willems, R.M., Kita, S., Hagoort, P., 2007. On-line integration of semantic information from speech and gesture: insights from event-related brain potentials. J. Cogn. Neurosci. 19, 605–616.
Paulesu, E., Perani, D., Blasi, V., Silani, G., Borghese, N.A., De Giovanni, U., Sensolo, S., Fazio, F., 2003. A functional–anatomical model for lipreading. J. Neurophysiol. 90, 2005–2013.
Pelphrey, K.A., Mitchell, T.V., McKeown, M.J., Goldstein, J., Allison, T., McCarthy, G., 2003. Brain activity evoked by the perception of human walking: controlling for meaningful coherent motion. J. Neurosci. 23, 6819–6825.
Rizzolatti, G., Craighero, L., 2004. The mirror-neuron system. Annu. Rev. Neurosci. 27, 169–192.
Rizzolatti, G., Fogassi, L., Gallese, V., 2001. Neurophysiological mechanisms underlying the understanding and imitation of action. Nat. Rev., Neurosci. 2, 661–670.
Rizzolatti, G., Fogassi, L., Gallese, V., 2002. Motor and cognitive functions of the ventral premotor cortex. Curr. Opin. Neurobiol. 12, 149–154.
Rose, M.L., 2006. The utility of arm and hand gestures in the treatment of aphasia. Adv. Speech–Lang. Pathol. 8, 92–109.
Saygin, A.P., Dick, F., Wilson, S.M., Dronkers, N.F., Bates, E., 2003. Neural resources for processing language and environmental sounds: evidence from aphasia. Brain 126, 928–945.
Schubotz, R.I., 2004. Human Premotor Cortex: Beyond Motor Performance. Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig.
Sekiyama, K., Kanno, I., Miura, S., Sugita, Y., 2003. Auditory–visual speech perception examined by fMRI and PET. Neurosci. Res. 47, 277–287.
Skipper, J.I., Nusbaum, H.C., Small, S.L., 2005. Listening to talking faces: motor cortical activation during speech perception. NeuroImage 25, 76–89.
Tanenhaus, M.K., Trueswell, J.C., 1995. Sentence Comprehension. In: Miller, J., Eimas, P. (Eds.), Speech, language, and communication. Handbook of perception and cognition. Academic Press, San Diego, pp. 217–262.
Tarr, M.J., Gauthier, I., 2000. FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. Nat. Neurosci. 3, 764–769.
Ugurbil, K., Garwood, M., Ellermann, J., Hendrich, K., Hinke, R., Hu, X., Kim, S.G., Menon, R., Merkle, H., Ogawa, S., et al., 1993. Imaging at high magnetic fields: initial experiences at 4 T. Magn. Reson. Q. 9, 259–277.
Umiltà, M.A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., Rizzolatti, G., 2001. I know what you are doing. a neurophysiological study. Neuron 31, 155–165.
van Atteveldt, N., Formisano, E., Goebel, R., Blomert, L., 2004. Integration of letters and speech sounds in the human brain. Neuron 43, 271–282.
Wagner, S., 2003. Verbales Arbeitsgedächtnis und die Verarbeitung lexikalisch ambiger Wörter in Wort- und Satzkontexten (PhD thesis). Max-Planck-Institute for Cognitive Neuroscience, Leipzig.
Ward, B.D., 2000. AlphaSim — Simultaneous Inference for FMRI Data. Biophysics Research Institute, Medical College of Wisconsin, Milwaukee, WI.
Willems, R.M., Özyürek, A., Hagoort, P., 2006. When Language Meets

Action: The Neural Integration of Gesture and Speech. Cerebral Cortex, bhl141.
Worsley, K.J., Liao, C.H., Aston, J., Petre, V., Duncan, G.H., Morales, F., Evans, A.C., 2002. A general statistical analysis for fMRI data. NeuroImage 15, 1–15.
Wright, T.M., Pelphrey, K.A., Allison, T., McKeown, M.J., McCarthy, G., 2003. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cereb. Cortex 13, 1034–1043.
Wu, Y.C., Coulson, S., 2007. How iconic gestures enhance communication: an ERP study. Brain Lang. 101, 234–245.