On-line Integration of Semantic Information from Speech and Gesture: Insights from Event-related Brain Potentials

Aslı Özyürek[1,2], Roel M. Willems[1], Sotaro Kita[3], and Peter Hagoort[1,2]

Abstract

■ During language comprehension, listeners use the global semantic representation from previous sentence or discourse context to immediately integrate the meaning of each upcoming word into the unfolding message-level representation. Here we investigate whether communicative gestures that often spontaneously co-occur with speech are processed in a similar fashion and integrated to previous sentence context in the same way as lexical meaning. Event-related potentials were measured while subjects listened to spoken sentences with a critical verb (e.g., *knock*), which was accompanied by an iconic co-speech gesture (i.e., KNOCK). Verbal and/or gestural semantic content matched or mismatched the content of the preceding part of the sentence. Despite the difference in the modality and in the specificity of meaning conveyed by spoken words and gestures, the latency, amplitude, and topographical distribution of both word and gesture mismatches are found to be similar, indicating that the brain integrates both types of information simultaneously. This provides evidence for the claim that neural processing in language comprehension involves the simultaneous incorporation of information coming from a broader domain of cognition than only verbal semantics. The neural evidence for similar integration of information from speech and gesture emphasizes the tight interconnection between speech and co-speech gestures. ■

INTRODUCTION

In ordinary face-to-face conversation, language users not only hear speech but also see the speaker's hand, mouth, and body movements. The listener's brain therefore continuously integrates spoken language information with several streams of visual information, including information from the lips, the eyes and, crucially, semantic information from the hand gestures that accompany speech (McNeill, 1992). For example, when talking about drinking a glass of milk, speakers often perform a concomitant drink gesture (i.e., C-shaped hand moved toward the mouth) as they say ''drink'' in their spoken utterance. Yet, whether and how listeners integrate the semantic information from co-speech gestures on-line into the previous sentence context, and how this compares to the integration of spoken words, has not been addressed.

So far, most studies on language comprehension have focused on the on-line processing of the acoustic and written input in isolation (but see Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995, for visual world related processing). Studies of on-line language comprehension using the event-related potential (ERP) technique have shown that spoken words are integrated into a context representation in an incremental way. That is, listeners use the global semantic representation from the sentence or discourse context to integrate the meaning of each upcoming word immediately into an overall message representation (e.g., Hagoort, 2003a, 2003b; Van Berkum, Zwitserlood, Brown, & Hagoort, 2003; Van Berkum, Hagoort, & Brown, 1999; Osterhout, McLaughlin, & Bersick, 1997; Kutas & Hillyard, 1980).

Previous studies on multimodal processing during language comprehension have often investigated the relationship between speech and lip movements by exploiting the McGurk effect (e.g., acoustic /pa/ combined with visual /ka/ perceived as /ta/, McGurk & Mac Donald, 1976; e.g., Colin et al., 2002; Mottonnen, Krause, Tiippana, & Sams, 2002; Sams et al., 1991). These studies using electrophysiological recordings have shown that visual information from articulation interacts with the auditory information quite early, that is, within 200 msec during audio/visual speech observation. However, little is known about how other types of visual information, such as gestures, are processed in relation to speech. The relationship between lip movements and syllables is based on form matching, whereas the

[1]F. C. Donders Centre for Cognitive Neuroimaging, Nijmegen, [2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, [3]University of Birmingham, UK

relation between speech and gestures is based on meaning. Thus, the latter might be processed in a different way, that is, at a higher, semantic level.

The Role of Co-speech Gestures in Communication

Here we focus on a ubiquitous form of communication that speakers use along with speech, namely, meaningful hand movements, usually referred to as co-speech gestures (Kendon, 2004; Goldin-Meadow, 2003; McNeill, 1992, 2000). During face-to-face conversation, speakers spontaneously use different types of gestures. These can be classified as either iconic (e.g., hands represent a climbing action), deictic (e.g., pointing), or emblematic (e.g., thumbs-up, OK). In this study, we focus on iconic gestures, which convey information about the shape, size, motion, and action characteristics of the events described in the spoken utterance. These gestures are meaningful within the speech context but do not have conventional or unambiguous meanings in the absence of speech (Krauss, Morrel-Samuels, & Colasante, 1991; Feyereisen, van de Wiele, & Dubois, 1988).

Iconic gestures have different representational properties than speech in terms of the meaning they convey. Consider, for example, an upward hand movement in a climbing manner when a speaker says: ''the cat climbed up the tree.'' Here, the gesture depicts the event as a whole, describing manner (''climb'') and direction (''up'') simultaneously, whereas in speech the message unfolds over time, broken up into smaller meaningful segments (i.e., different words for manner and direction). However, despite these differences in representational format, the information expressed in the two modalities is systematically related to each other (Bernardis & Gentilucci, 2006; Kendon, 2004; McNeill, 1992, 2000), and conveys the speaker's overall meaning during conversation as a ''composite signal'' (Clark, 1996).

The systematic relationship between speech and gestures exists at three levels. First, there is semantic overlap between the representation in gestures and the meaning expressed in the concurrent speech, as in the ''climb up'' example above (e.g., Kita & Özyürek, 2003; McNeill, 1992). Speech and gesture convey related and similar information. Second, speech and gesture are temporally aligned to each other. A gesture phrase has three phases: the preparation, the stroke (semantically the most meaningful part of the gesture), and the retraction or hold (McNeill, 1992). Studies have shown that the onset of the gesture phrase (i.e., preparation) usually precedes the onset of the relevant speech segment by less than a second (Morrel-Samuels & Krauss, 1992; Butterworth & Beattie, 1978). More importantly, in most speech–gesture pairs, the stroke coincides with the relevant speech segment (McNeill, 1992). Finally, it has been shown that the spontaneous use of gestures has a similar function as speech (e.g., Kendon, 2004; Melinger & Levelt, 2004; Özyürek, 2002), namely, to communicate the intended message to the addressee.

A vast amount of behavioral studies on speech and gesture comprehension has shown that listeners/viewers pay attention to iconic gestures and pick up the information that they encode. For example, Graham and Argyle (1975) had speakers describe abstract line drawings with and without gestures, and required listeners to make drawings on the basis of the speakers' input. Listeners were more accurate in their drawings in the speech-and-gesture condition than in the speech-alone condition. In another study, Beattie and Shovelton (1999) showed that listeners answer questions about the size and relative position of objects in a speaker's message more accurately when gestures are part of the description than when gestures are absent.

Another set of studies has investigated whether listeners pick up the information in gesture when gesture conveys different information than speech. McNeill, Cassell, and McCullough (1999) presented listeners with a videotaped narrative in which the semantic relationship between speech and gesture was manipulated. It was found that listeners/viewers incorporated information from the gestures in their retellings of the narratives, and attended to the information conveyed in gesture when that information complemented or even contradicted the information conveyed in speech (see also Singer & Goldin-Meadow, 2005; Goldin-Meadow & Sanhofer, 1999; Kelly & Church, 1998).

Despite firm evidence that co-speech gestures contribute to comprehending the speaker's message, not much is known about the nature of the on-line cognitive processes underlying the comprehension of co-occurring multimodal semantic information from speech and gesture. The present study investigates the integration of speech and gesture as they naturally occur, that is, simultaneously, and embedded into a sentence context. For this purpose, we exploited an ERP paradigm that is often used for studying the nature of on-line semantic integration in sentence and discourse contexts.

ERP Studies on Semantic Integration during Comprehension

Event-related potentials are voltage deflections generated by the brain and recorded from electrodes placed on the scalp. One important characteristic of ERPs is their high temporal resolution, which is in the order of milliseconds. The processing of semantic information has been found to influence the amplitude of a negative-going ERP component between 250 and 550 msec. This amplitude modulation is referred to as the N400 effect and is usually larger over posterior electrodes than over frontal sites (Kutas & Hillyard, 1980).

N400 studies have typically employed a paradigm in which the semantic integration load of a word in relation to the preceding sentence context is manipulated. Kutas

and Hillyard (1980) were the first to observe that, relative to a semantically acceptable control word, a sentence-final word that is semantically anomalous in the sentence context, as in ''He spread the warm bread with *socks*,'' elicits an N400 effect. Additional studies have shown that it does not require a semantic violation to elicit an N400 effect. In general, N400 effects are triggered by more or less subtle differences in the semantic fit between the meaning of a word and its context, where the context can be a single word, a sentence, or a discourse (e.g., Van Berkum et al., 1999, 2003; Osterhout et al., 1997; Hagoort & Brown, 1994; Kutas & Hillyard, 1984).

More recent studies on semantic processing have investigated how extralinguistic information, such as world knowledge or pictorial information, is integrated into previous context. Hagoort, Hald, Bastiaansen, and Petersson (2004) showed their subjects sentences that contained either a semantically anomalous word (e.g., Dutch trains are *sour* and very crowded) or a world knowledge violation (e.g., Dutch trains are *white* and very crowded). The N400 effects to the semantic and to the world knowledge violations were identical in their latency and topography. These results indicate that even in the case of extralinguistic information such as world knowledge, the brain integrates this information immediately, that is, with the same temporal profile as lexical–semantic information.

Processing of extralinguistic information has also been investigated in terms of integrating information from pictures to previous context (West & Holcomb, 2002; Federmeyer & Kutas, 2001; McPherson & Holcomb, 1999; Ganis, Kutas, & Sereno, 1996; Barret & Rugg, 1990). In picture priming studies, an N300 has been reported that is more negative for unrelated than for related pictures (McPherson & Holcomb, 1999; Holcomb & McPherson, 1994; Barret & Rugg, 1990). This N300 has a frontal distribution and is not reported in ERP studies that used only linguistic stimuli. The N300 was followed by a more widely distributed N400 effect. However, in other studies in which either anomalous words or pictures were presented in a sentence context, only N400 effects were found (Ganis et al., 1996; Nigam, Hoffman, & Simons, 1992). In these studies, the pictures elicited an N400 effect with a more frontal distribution than is usually observed for language stimuli. Finally, studies investigating the semantic integration of pictures to a scene or event without using any linguistic context sometimes (West & Holcomb, 2002), but not always (Ganis & Kutas, 2003; Sitnikova, Kuperberg, & Holcomb, 2003), found a frontal N300 preceding an N400. The partial differences in distribution of ERP effects for words versus pictures have led researchers to suggest that they have both overlapping as well as nonoverlapping semantic representations (e.g., West & Holcomb, 2002; Federmeyer & Kutas, 2001). In the light of these findings, it is especially interesting to see how iconic gestures compare to semantic integration of pictures. Gestures can be claimed to share certain visual characteristics with pictures. However, they do not have the exact semantic specificity of pictures because, unlike pictures, the full interpretation of gestures depends on the semantic content of the accompanying speech.

Finally, two recent priming studies have investigated the modulation of ERPs to words preceded by gestures, or to gestures preceded by cartoon images. Kelly, Kravitz, and Hopkins (2004) found that ERPs to spoken words (targets) are modulated when these words are preceded by gestures (primes) that contained information about the size and shape of objects that the target words referred to. Compared to matching target words, mismatching words evoked an early P1/N2 effect, followed by an N400 effect. On the basis of these findings, Kelly et al. (2004) claimed that the gesture primes influenced word comprehension, first at the level of ''sensory/phonological'' processing and later at the level of semantic processing. In a recent study by Wu and Coulson (2005), it was found that congruous and incongruous gestures shown without speech and following cartoon images elicit a negative-going ERP effect around 450 msec. In addition, it was observed that congruous or incongruous words following the cartoon–gesture pairs elicited an N400 effect. However, neither of these studies has investigated speech and gesture comprehension within sentence context and when they occur simultaneously as in everyday conversations.

The Present Study

The present study investigates the integration of speech and gesture as they naturally occur, that is, simultaneously and embedded into a sentence context. Using a similar ERP paradigm as for investigating the semantic integration of words, we aim to compare the latency, the amplitude, and the topography of gesture integration to sentence context with the integration of spoken words. Our main focus is on understanding how the integration of conceptual information from gestures into a previous sentence context (i.e., global integration) compares to integration of semantic information from spoken words. Secondly, we also investigate how listeners/viewers comprehend and integrate the information from the temporally overlapping speech and gesture segments (i.e., local integration). For example, when a listener hears ''the cat climbed up the tree and caught the bird'' and sees a CATCH gesture as he/she hears the word ''caught,'' the comprehension of gesture in relation to previous sentence would be ''global integration'' and its relation to the verb ''catch'' is referred to as ''local integration.'' Thus, we aim to reveal the underlying nature and time course of these two types of multimodal integration processes.

The particular questions that we investigated are: (i) Are gestures and speech integrated to previous sentence

Table 1. An Example of the Stimulus Materials

A) Language gesture match (Correct condition): L+G+	
He slips on the roof and rolls down [roll down]	
B) Language mismatch: G+**L**−	Local mismatch
He slips on the roof and **walks** to the other side [roll down]	Local mismatch
C) Gesture mismatch: **G**−L+	Local mismatch
He slips on the roof and rolls down **[walk across]**	Local mismatch
D) Double mismatch: **G**−**L**−	Local match
He slips on the roof and **walks** to the other side **[walk across]**	Local match

A verbal description of the iconic gesture is presented in brackets [ ]. Gestures were time-locked to the onset of the critical verb (underlined). ERPs were time-locked to the beginning of the critical word and the gesture in each sentence. The condition coding (G+L+, G+L−, etc.) refers to the match/mismatch of either the verb (Language = L) or the gesture (Gesture = G) to the preceding sentence context, with a minus sign indicating a mismatch. Mismatches to the preceding context are indicated in **bold**. Conditions B and C also contain local mismatches where the concurrent speech and gesture are different. All stimuli were in Dutch.

context simultaneously, or is speech integrated first and gesture later? (ii) How does the integration of gesture information to previous sentence context (i.e., global integration) compare to integration of gesture information to the temporally overlapping word (i.e., local integration)?

In order to determine the nature of the integration of verbal and gestural semantic information, we manipulated the semantic fit of speech (i.e., a critical verb) and/or gesture in relation to the preceding part of the sentence (global integration) as well as the semantic relations between the temporally overlapping gesture and speech (local integration) (see Table 1).

Movie clips of iconic gestures were temporally aligned to the critical verbs in the sentences. This manipulation resulted in four conditions (see Table 1, Figure 1): correct condition [Gesture (G)+, Language (L)+]; language mismatch condition [G+L−]; gesture mismatch condition [G−L+]; double mismatch condition [G−L−]. In the language mismatch, the critical verb was harder to fit semantically to the preceding context, whereas the co-occurring gesture matched the sentence context. In the gesture mismatch condition, the gesture was harder to integrate to previous context, whereas the critical verb matched the spoken sentence context. In the double mismatch condition, both the gesture and the word were difficult to integrate to previous sentence context. Note that in the language and gesture mismatch conditions, the critical verb and the overlapping gesture locally mismatched (i.e., speech: roll; gesture: walk, and vice-versa), whereas in the double mismatch condition, they locally matched (i.e., both ''walk''). This extra

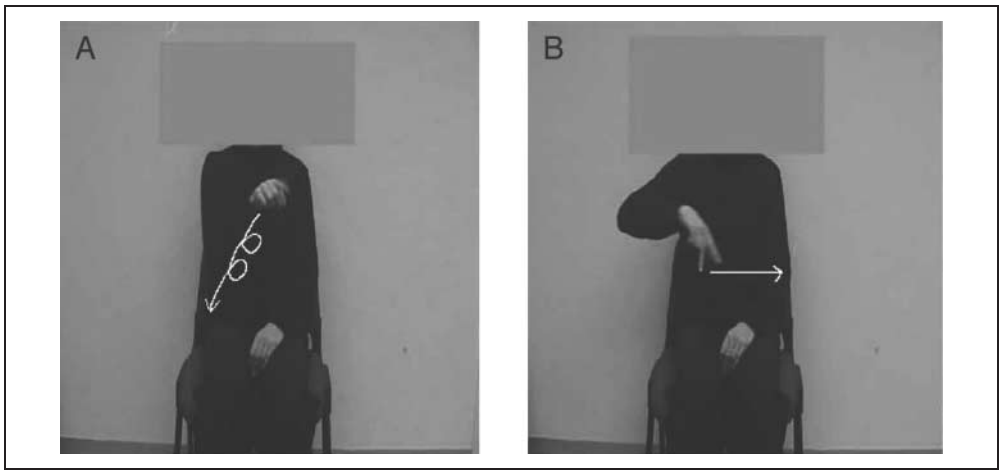

Figure 1. Examples of the gesture movies. Stills from two gestures that were used as stimuli: (A) Roll down; (B) Walk across.

manipulation allowed us to investigate and compare the effects of local and global integration of speech and gesture in sentence context.

In our materials, an increase of semantic integration load does not necessarily involve a semantic *violation*. The meaning of the critical verb in the mismatch condition, however, always fits the previous sentence context less well than the meaning of its counterpart in the correct condition. ERP studies in language processing have found that semantically less expected critical words elicit an increase in the amplitude of the N400, just as semantic violations do (Hagoort & Brown, 1994; Kutas & Hillyard, 1984). For reasons of simplicity, we will refer to conditions in which speech and/or gesture are harder to integrate as ''mismatches.''

If the brain uses an incremental and parallel processing of linguistic and extralinguistic information as found in previous studies (Hagoort et al., 2004), we expect a similar latency and amplitude of the N400 effect for all types of mismatches (i.e., language, gesture, and double), revealing that the brain integrates information from both speech and gesture at the same time. These results would also be in line with the claims that speech and gestures are tightly linked systems of communication (Kendon, 2004; Goldin-Meadow, 2003; Özyürek, 2002; McNeill, 1992, 2000; Clark, 1996). However, if the latency of the N400 effect was found to be later for the gesture mismatch than for the language mismatch, this would support a speech-first-gesture-later model of comprehension. This model is compatible with the view that the semantic interpretation of sentences precedes the integration of pragmatic, extralinguistic information (Forster, 1979). It would be also in line with the view of Krauss et al. (1991) that the meaning we assign to gestures is mostly constructed from the meaning of concurrent speech, and that gestures do not add any information to what the listener picks up from the concurrent speech. Accordingly, gestural information will have to be integrated after the relevant speech segment has been interpreted (if it is integrated at all).

Furthermore, according to the incremental processing principle, we do not expect differences across conditions with local mismatches (language and gesture mismatches) and the condition with the local match (double mismatch) because integration takes place immediately in relation to a discourse model and not in multiple steps from lower to higher levels of semantic organization (Van Berkum et al., 1999, 2003). According to this view, the gesture and the concurrent speech segment (i.e., the verb) are integrated in parallel into the preceding context, and not after they first formed a common semantic object. Alternatively, it might be argued that the local conflict between speech and gesture has to be resolved first, before the global integration can take place in the local mismatch conditions. In this case, the double mismatch effect should precede the effects for the single language and gesture mismatches because in this condition a local integration problem is absent.

METHODS

Subjects

Sixteen healthy subjects (12 women; mean age = 22.4, range = 19–27), with normal or corrected-to-normal vision and no hearing complaints, took part in the study. All subjects were right-handed and had Dutch as their mother tongue. Subjects gave written informed consent and were paid for participation.

Materials

The materials consisted of 320 spoken Dutch sentences. The sentences were spoken by a female native speaker of Dutch and digitized at a sample frequency of 44.1 kHz. The sentences formed 160 sentence pairs. The members of the pair were identical up until the critical verb. Half of the sentences contained a critical verb that matched the preceding context. In the other half, the critical verb was semantically anomalous in relation to the prior sentence context. Overall, 12 different critical verbs were used (see Appendix). For each sentence, the onset of the critical verb was determined by using the speech analysis software package Praat (version 4.0; www.praat.org). The sentences had an average duration of 3720 msec (*SD* = 81), and the critical verbs had an average duration of 322 msec (*SD* = 85 msec).

The spoken sentences were combined with 12 iconic gestures (see Appendix). Iconic gestures are a class of gestures that speakers spontaneously use as they talk about spatial and activity-related aspects of events (e.g., using wiggling fingers moving horizontally while talking about someone walking). The iconic gestures used in this study were based on a larger database collected to investigate speakers' natural and spontaneous use of speech and gestures in narratives of spatial events (Kita & Özyürek, 2003; Özyürek, 2002). For the purposes of this study, 12 of these gestures were selected and modeled by a native Dutch speaker with the requirement that they resembled spontaneous gestures in this database. Modeled gestures were preferred over natural ones from the database to make each gesture comparable across the conditions in terms of gesture space that was used, the handedness, and the gesturing person. In order to match the speed and length of the gestures as closely as possible to naturally occurring ones, we asked our model to produce concurrent sentences as she was performing the gestures. The gestures were filmed by using a digital camera (Sony, TCR-TRV950, PAL). During editing, the audio was removed from the movie. Movies were edited using Adobe Premier (version 6.0; Adobe Systems, San Jose, CA; www.adobe.com). The preparation

and the retraction phases of each gesture were removed, leaving the stroke. Previous research has shown that especially the stroke phase conveys the meaning of a gesture (McNeill, 1992). By isolating the gesture stroke phase, we eliminated differences among gestures that were due to the fact that, for some gestures, hand shape might reveal information before the stroke began, and/or that some gestures might have longer preparation time than others. The average length of the strokes was 767 msec (*SD* = 284 msec). Finally, the face of the model was blocked to eliminate the contribution of information coming from the lips.

The gestures corresponded to the meaning of the critical verbs. They were combined with the sentence pairs in such a way that in half of the items the gesture matched the preceding sentence context, and in the other half it mismatched the preceding sentence context. This resulted in a total of 160 stimulus quartets (see Table 1).

The gesture movies and the sentence files were combined using the Adobe Premier (version 6.0) and After Effects software (version 5.5; Adobe Systems, www.adobe.com). For each movie file, the onset of the gesture stroke was temporally aligned with the onset of the critical verb because in 90% of natural speech–gesture pairs the stroke coincides with the relevant speech segment (McNeill, 1992). For verbs with a separable prefix, the alignment point was not word onset, but the body of the verb following the prefix. The latter was the case for 44 sentences. Additional still frames with the hand resting on the lap were added to the part of the sentence before the critical verb, and the last frame of the stroke was elongated until the end of the sentence.

Four different stimulus lists were created to distribute the four versions of each item equally over the four lists (see Table 1). This was done in such a way that all four lists contained an equal number of items (n = 40) per condition. Each list was presented to one quarter of the participants. As a result, none of the participants were presented with more than one item of a stimulus quartet as in Table 1.

Procedure

The stimuli were presented using the Nijmegen Experiment Setup program (NESU, MPI for Psycholinguistics). The visual content of the movies was presented via a computer screen. The subjects watched the movies at a distance of 80 cm from the screen. The size of the movie frame was 10 cm in height and 11.8 cm in width. The movies were presented at 25 frames per second. Speech was presented to the subjects through headphones.

Subjects were instructed to carefully listen to the sentences and watch the movies without a specific task. They were given the instruction that they could blink or move their eyes only during the interstimulus intervals when a fixation cross was shown. The fixation cross was presented between the movies for a duration of 3600 msec. Finally, they were told that they would receive general questions about the items after the experiment to make sure that they would attend to the items.

The test session started with a practice block of 30 practice items to familiarize the subjects with the procedure. The whole test session lasted approximately 40 min.

EEG Recording and Analysis

The electroencephalogram (EEG) was recorded from 26 electrode sites across the scalp using an Electrocap with 26 Ag/AgCl electrodes, each referred to the left mastoid and off-line re-referenced to average mastoids. Electrodes were placed on midline (Fz, FCz, Cz, Pz), frontal and fronto-central (F3, F4, F8, F7, FC5, FC1, FC2, FC6), temporal (T7, T8), central (C3, C4), centro-parietal (CP5, CP1, CP2, CP6), parietal (P7, P3, P4, P8), and occipital (O1, O2) sites. Vertical and horizontal eye movements were monitored via a supra- to suborbital bipolar montage and a right-to-left canthal bipolar montage, respectively. Activity over the right mastoid bone was recorded on an additional channel to determine if there were additional contributions of the experimental variables to the two presumably neutral mastoid sites. No such differences were observed.

The EEG and the electrooculogram (EOG) recordings were amplified with BrainAmp DC amplifiers. A bandpass filter was applied from 0.01 to 70 Hz. Impedances were kept below 5 kΩ for all channels. The EEG and EOG signals were recorded and digitized using Brain Vision Recorder Software (version 1.03), with a sampling frequency of 500 Hz.

Prior to off-line averaging, all single-trial waveforms were screened for eye movements, electrode drifting, amplifier blocking, and muscle (EMG) artifacts in a critical window that ranged from 150 msec before to 1000 msec after the onset of the critical verb and the gesture stroke. Trials containing such artifacts were rejected (7.7%). Rejected trials were equally distributed across conditions.

Event-related potentials time-locked to the onset of the critical verb and the gesture were averaged after baseline correction by subtracting mean amplitude in the −150 to 0 msec prestimulus interval, for each condition (correct, gesture mismatch, language mismatch, double mismatch) for each subject at each electrode site. Repeated measures analyses of variance (ANOVAs) with the factors match (correct, gesture mismatch, language mismatch, and double mismatch) and quadrant (left anterior: F3, F7, FC1, FC5, C3; right anterior: F4, F8, FC2, FC6, C4; left posterior: CP1, CP5, P3, P7, O1; and right posterior: CP2, CP6, P4, P8, O2)

were conducted for three time windows. Separate ANOVAs were conducted for the midline electrodes. Huynh–Feldt correction for violation of sphericity was applied when appropriate.

RESULTS

Figure 2 displays the grand-average waveforms time-locked to the onset of critical verbs and gesture strokes. A visual inspection of the waveforms (see Figure 2) shows an N1 followed by a P2, and a negativity with a bimodal morphology peaking at about 380 and 480 msec, respectively. Apart from a slightly smaller N1 in the correct condition, the waveforms suggest that the mismatch conditions started to deviate from the correct condition in the latency window of the P2 component around 225–275 msec. However, this effect seems especially strong for the correct condition, which could be a carryover from the reduced N1 amplitude in this condition. Next, at around 350 msec, the mismatch conditions deviate from the correct condition. This effect is followed by a similar modulation between 410 and 550 msec, with a peak latency that is slightly later than is usually seen for the N400.

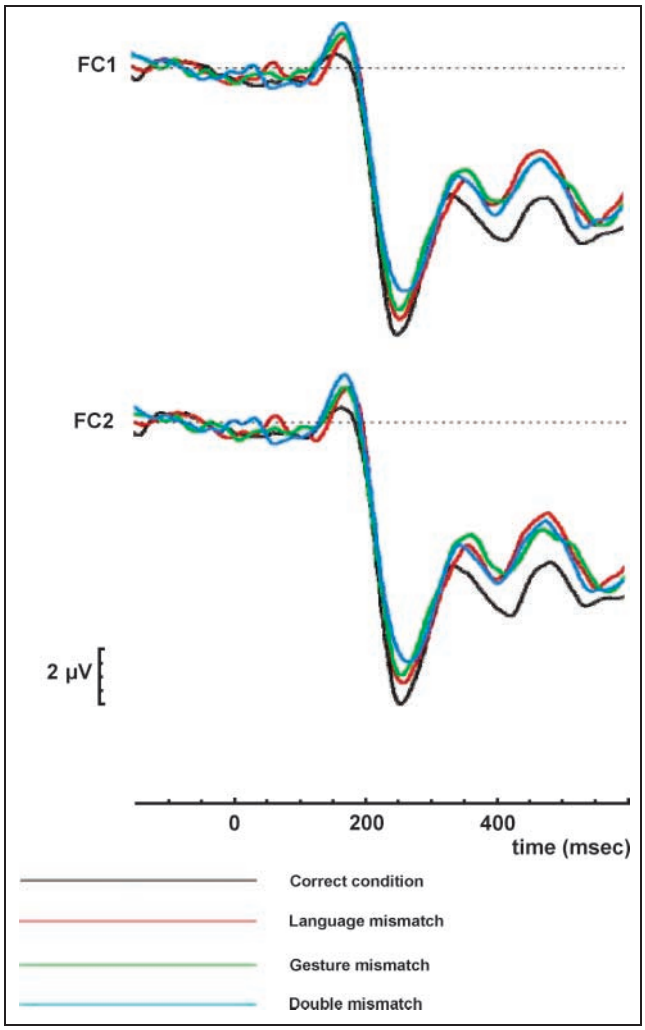

Figure 2. Grand-average waveforms for ERPs elicited in the three mismatch conditions and the correct condition at two representative electrode sites (FC1 and FC2). Negativity is plotted upward. Waveforms are time-locked to the onset of spoken verb and gesture (0 msec).

For the P2, the double mismatch seemed to show a reduced amplitude. A repeated measures ANOVA on the mean amplitudes in the 225–275 msec latency range, with the factors match and quadrant, failed to show a significant main effect of match [$F(3,45) = 1.90$; $MSE = 22.1$; $p = .14$]. There was also no significant Match × Quadrant interaction ($F < 1$). In addition, the ANOVA over the midline sites failed to reach significance [$F(3,45) = 2.29$; $MSE = 10.29$; $p = .09$]. However, in a planned comparison, a significant difference between the double mismatch and the correct condition was found [$F(1,15) = 4.69$; $MSE = 5.63$; $p < .05$]. Also over the midline electrodes, this planned comparison showed that ERPs to the double mismatch were less positive than those to the correct condition [$F(1,15) = 5.28$; $MSE = 26.17$; $p < .05$]. In addition, in a planned comparison, it was found that over the midline sites the language mismatch was significantly less positive than the correct condition [$F(1,15) = 4.7$; $MSE = 11.4$; $p < .05$]. However, these effects are qualified by the fact that for the N1, the correct condition shows a smaller amplitude than the other conditions. If we take this unexplained early difference into account by using another baseline (100–200 msec), no significant differences remain. In short, the P2 effect observed for the double mismatch does not seem to be a stable effect. The conclusion that there is an earlier effect for the double mismatch than for the language and gesture mismatches would therefore be premature.

The next window in which effects were tested was in the latency range of 350–410 msec. This is the window around the first negative peak (approximately at 380 msec) in the waveforms following the P2. For this latency window, repeated measures ANOVAs, with the factors match and quadrant, did not show a significant main effect for match [$F(3,45) = 1.22$; $MSE = 18.1$; $p = .32$], or a significant Match × Quadrant interaction [$F(9,135) = 1.40$; $MSE = 5.43$; $p = .23$]. Additional planned comparisons resulted in significant differences between the gesture mismatch and the correct condition in the left anterior quadrant [$F(1,15) = 7.92$; $MSE = 20.49$; $p < .05$], the right anterior quadrant [$F(1,15) = 14.77$; $MSE = 12.19$; $p < .005$], and over the midline sites [$F(1,15) = 6.18$; $MSE = 12.5$; $p < .05$]. In addition, the language mismatch condition showed a marginally significant effect in the left anterior quadrant [$F(1,15) = 4.48$; $MSE = 13.18$; $p = .051$], and just failed to reach significance over the midline sites [$F(1,15) = 3.40$; $MSE = 21.64$; $p = .085$]. For the double mismatch,

no significant effects were found in this latency window. However, in contrasts testing the differences between the three mismatching conditions, no significant effects were obtained.

Finally, the average waveforms were tested in the time window of 410–550 msec. As can be seen in Figure 2, all three types of mismatch elicit a clear negative deflection that all peak around the same time. Moreover, the topographic distribution shows that for all three mismatches, the effects are maximal over anterior sites, without a clear hemispheric dominance (see Figure 3).

The results of repeated measures ANOVAs in this latency window are summarized in Table 2. The main effect of match is modulated by an interaction between match and quadrant due to the clear anterior distribution of the condition effects. Planned comparisons conducted in separate quadrants revealed significant differences between all three mismatch conditions, and the correct condition for the anterior electrode sites (both left and right hemisphere sites), as well as for the midline sites (with the exception of the double mismatch). Further planned comparisons between the three mismatch conditions did not reveal any significant differences. No significant effects were obtained over posterior quadrants.

Thus, the results show that language, gesture, and double mismatch conditions modulated the N400 in a similar way, in terms of N400 latency and amplitude. In all conditions, the N400 component reached its peak around 480 msec. Furthermore, all conditions showed a similar topographical distribution.

DISCUSSION

This study investigated the semantic integration of words and iconic gestures into a sentence context when they both occur simultaneously as in natural speech and gesture production. The most important finding of the study is that co-occurring speech and gestures are integrated simultaneously into a preceding sentence context. That is, semantic information provided by both spoken words and visual gestures is integrated within 350–550 msec after word and gesture onset. The time course of the observed N400 effects testifies to the immediacy of contextual integration because, in many

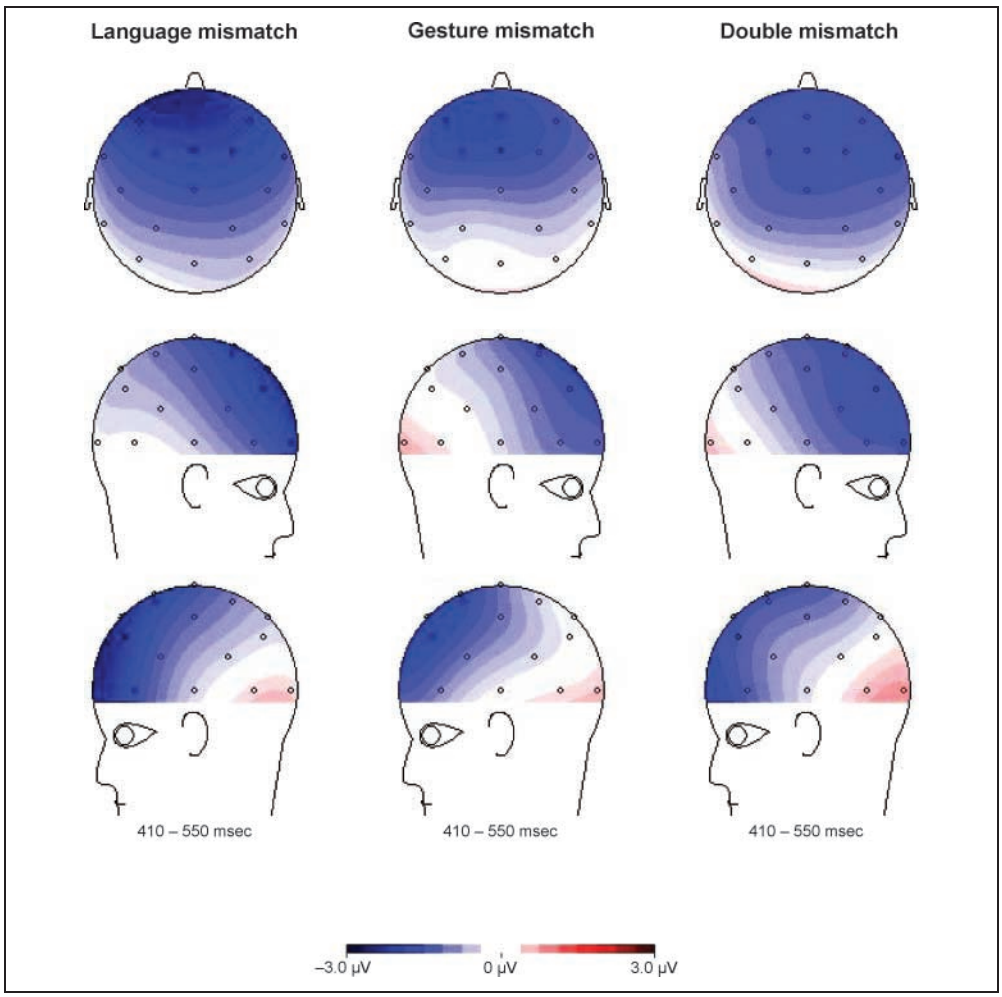

Figure 3. Spline-interpolated isovoltage maps displaying the topographic distributions of the mean differences from 410 to 550 msec between the three types of mismatches and the correct condition.

Table 2. Repeated Measures ANOVAs on Mean ERP Amplitudes for the Four Experimental Conditions in the 410–550 msec Latency Range

Source	*df*	*F*	*MSE*	*p*
ANOVA: Match (4 levels), Quadrants (4 levels) (20 electrodes)				
Match	3,45	2.84	14.90	.048*
Match × Quadrant	9,135	3.12	4.91	.013*
Left Anterior Quadrant (Electrodes: F3, F7, FC1, FC5, C3)				
Match	3,45	4.01	8.02	.013*
Planned comparisons:				
L−G+	1,15	16.8	10.08	.001**
L+G−	1,15	9.38	11.54	.008**
L−G−	1,15	4.46	19.07	.052
Right Anterior Quadrant (Electrodes: F4, F8, FC2, FC6, C4)				
Match	3,45	6.03	5.31	.002**
Planned comparisons:				
L−G+	1,15	16.77	9.38	.001**
L+G−	1,15	8.60	10.46	.01*
L−G−	1,15	9.35	13.14	.008**
Midline Sites (Electrodes: Fz, FCz Cz, Pz)				
Match	3,45	3.15	8.10	.034*
Planned comparisons:				
L−G+	1,15	7.92	18.03	.013*
L+G−	1,15	4.37	16.19	.054
L−G−	1,15	2.73	23.39	.12

Huynh–Feldt correction is applied when appropriate. The original degrees of freedom are reported. Planned comparisons are always against the correct condition. L−G+ = language mismatch; L+G− = gesture mismatch; L−G− = double mismatch.

*$p < .05$.

**$p < .01$.

cases, they occur well before the end of the acoustic word token or the visual gesture. As the topographic distributions of the gesture and word integration effects are identical, it is most parsimonious to assume that the nature of the semantic integration process is very similar in both cases.

No solid evidence was obtained that the effect for the double mismatch came earlier than the single mismatch effects (i.e., gesture and language mismatches). In the double mismatch condition, the co-occurring critical verb and the gesture provided compatible semantic information (i.e., local match). This was different in the language and gesture mismatch conditions. In these conditions, the co-occurring verb and gesture were mutually inconsistent (i.e., local mismatch). This local mismatch, however, did not seem to modulate the global mismatch effect, that is, the effect triggered by the mismatch in relation to the preceding sentence context. More in particular, no evidence was obtained that the effect for the double mismatch (i.e., the local match) preceded the effects of the language and gesture mismatches (i.e., the local mismatch). This suggests that verb and gesture are not first integrated together to form a common semantic object, before integration into the preceding context takes place. Instead, verb and gesture seem to be integrated in parallel. This is in line with the view supported by N400 data in Van Berkum et al. (1999, 2003) that semantic integration takes place immediately in relation to a discourse model rather than in a series of sequential steps from lower to higher levels of semantic organization.

In terms of their latency and amplitude characteristics, the effects are similar to the well-known N400 effect that is observed if word meaning violates the semantic context (Kutas & Hillyard, 1980). However, the waveforms show a clearly biphasic morphology, and the effects have a more anterior distribution than is reported for the classical N400 effect. The first negative peak in the biphasic negativity is reminiscent of the N300 that has been reported before for visual materials, and which has been found to be more negative for unrelated than for related pictures (McPherson & Holcomb, 1999; Holcomb & McPherson, 1994; Barret & Rugg, 1990). The N300 effect might be related to the presence of the visual–gestural information.

For the N400, an anterior distribution has been observed previously for visual information such as pictures (e.g., West & Holcomb, 2002; Federmeyer & Kutas, 2001; Ganis et al., 1996). In the current study, the visual characteristics of the gestures might have elicited a frontal distribution. It is interesting here to note that even the language mismatch condition elicited an anterior effect, which suggests that the mere presence of a simultaneous gesture is responsible for the anterior distribution, even when the integration problem is located in the speech channel. The finding that all mismatch conditions have similar topographic distributions suggests that semantic integration of information from both modalities might be instantiated by overlapping neuronal sources. Interestingly, it suggests that with respect to contextual integration, there is no reason to distinguish between visual semantics and verbal semantics.

As a cautionary note, we want to point out that the N300 and N400 effects are descriptive labels. There is no evidence that both effects are independently modulated or generated by nonoverlapping neural generators. Earlier studies involving visual materials have reported both N300 and N400 effects. We have chosen our descriptive terms here in connection to these earlier studies.

However, our main conclusions do not depend on the question whether N300 and N400 effects are one and the same extended negativity.

The present study, together with the studies by Wu and Coulson (2005) and Kelly et al. (2004), point to the fact that iconic gestures trigger semantic processing, as is indicated by the presence of N400 effects. However, the current study differs from these earlier studies in crucial ways. In these studies, words and gestures were presented sequentially, and ERPs were measured to either word targets preceded by gestures or to gesture targets preceded by cartoon images. In the present study, the gestures and the relevant speech segments were presented simultaneously as they naturally occur, and furthermore in a sentence context by which the integration of gestural information to speech context beyond single word and gesture levels could be investigated.

It is also important to note that we found an N400 effect instead of the earlier negativities normally reported to speech–lip movement mismatches in the McGurk effect (e.g., Colin et al., 2002; Mottonnen et al., 2002; Sams et al., 1991). This provides evidence that speech and gesture integration occurs at a higher semantic level than the integration of information from lip movements and speech sounds; that is, different types of multimodal information are processed in different ways in the brain, even though both concern processing relations between speech and visual movements.

Finally, our results parallel those of a recent functional magnetic resonance imaging (fMRI) study (Willems, Özyürek, & Hagoort, in press), using the same stimuli in a design with the same conditions. In the fMRI study, it was found that all mismatch conditions activated a common area, namely, the left inferior frontal context. This area has been claimed to be crucial for the integration of semantic information into previous context (Hagoort, 2003b, 2005; Hagoort et al., 2004). Together with the ERP results of the current study, the fMRI data suggest that the semantic integration of both speech and gesture semantics to sentence context involves very similar processes.

In conclusion, when understanding an utterance, the brain does not restrict itself to language information alone, but also integrates semantic information conveyed through other modalities, such as co-speech gestures. Furthermore, the neuronal sources and the time course of the integration processes seem to be similar across gesture and language semantics. Both constrain the interpretation domain simultaneously during on-line processing. This opens the interesting possibility that language comprehension involves the incorporation of information in a ''single unification space'' (Hagoort, 2003b, 2005; Hagoort et al., 2004), coming from a broader range of cognitive domains than is usually thought. The neural evidence for the tight link between speech and gesture that we observed underscores the fact that, in natural conversation, speech and gesture are often tightly interconnected (Bernardis & Gentilucci, 2006; Willems et al., in press; Kelly et al., 2004; Kendon, 2004; Goldin-Meadow, 2003; Kita & Özyürek, 2003; Özyürek, 2002; McNeill, 1992, 2000; Clark, 1996). Further research has to reveal if, in this sense, co-speech gestures are special or representative of a broad domain of visual information constraining on-line sentence interpretation (Tanenhaus et al., 1995).

APPENDIX

List of Critical Verbs (Originals in Dutch) and Gestures Used as Stimuli within Sentence Context

Critical Verb	*Gesture*	*Gesture Description*
Break	BREAK	Fist hands make a break motion from the middle to the sides and down
Give	GIVE	Hand opens up as it moves forward
Knock	KNOCK	Fist hand moves back and forth
Punch	PUNCH	Fist hand make a punching motion away from body
Push	PUSH	Both flat hands move away from body
Roll away	ROLL_AWAY	Index finger pointing to the right makes circles as it moves away from the body
Roll down	ROLL_DOWN	Index finger pointing away from body makes circles as it moves down and left
Swing across	SWING_ACROSS	Index finger pointing away from body moves left making an arc
Swing away	SWING_AWAY	Index finger pointing towards right moves away from body making an arc
Walk away	WALK_AWAY	V hand shape with wiggling fingers moves forward away from self
Walk across	WALK_ACROSS	V hand shape with wiggling fingers moves left horizontally
Write	WRITE	One hand makes a writing gesture moving to the right

Acknowledgments

This research was supported by a grant from the Netherlands Organization for Scientific Research (NWO), 051.02.040, and Biotechnology and Biological Sciences Research Council, UK (BBS/B/08906). We thank Femke Deckers, Miriam Kos, Nienke Weder, and Tineke Snijders for their assistance during the running of this experiment, and Jos van Berkum for his comments on an earlier version of this article.

Reprint requests should be sent to Aslı Özyürek, PhD, F. C. Donders Centre for Cognitive Neuroimaging, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands, or via e-mail: asliozu@mpi.nl.

REFERENCES

Barret, S. E., & Rugg, M. D. (1990). Event-related potentials and the semantic matching of pictures. *Brain and Cognition, 14,* 201–212.

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.

Bernardis, P., & Gentilucci, M. (2006). Speech and gesture share the same communication system. *Neuropsychologia, 44,* 178–190.

Butterworth, B., & Beattie, G. (1978). Gesture and silence as indicators of planning in speech. In N. Campbell & P. T. Smith (Eds.), *Recent advances in the psychology of language: Formal and experimental approaches* (pp. 347–360). New York: Plenum.

Clark, H. H. (1996). *Using language.* Cambridge: Cambridge University Press.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: A phonetic representation within short term memory. *Clinical Neurophysiology, 113,* 495–506.

Federmeyer, K. D., & Kutas, M. (2001). Meaning and modality: Influences of context, semantic memory, organization, and perceptual predictability on picture processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 202–204.

Feyereisen, P., Van de Wiele, M., & Dubois, F. (1988). The meaning of gestures: What can be understood without speech? *Cahiers de Psychologie Cognitive, 8,* 3–25.

Forster, K. I. (1979). Levels of processing and the structure of the language processor. In W. E. Cooper & C. T. Walker (Eds.), *Sentence processing: Psycholinguistic essays presented to Merrill Garrett* (pp. 27–85). Hillsdale, NJ: Erlbaum.

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research, 16,* 123–144.

Ganis, G., Kutas, M., & Sereno, M. (1996). The search for common sense: An electrophysiological study of the comprehension of words and pictures for reading. *Journal of Cognitive Neuroscience, 8,* 89–106.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think.* Cambridge: Harvard University Press.

Goldin-Meadow, S., & Sanhofer, C. M. (1999). Gesture conveys substantive information about a child's thoughts to ordinary listeners. *Developmental Science, 2,* 64–74.

Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology, 10,* 57–67.

Hagoort, P. (2003a). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *Journal of Cognitive Neuroscience, 15,* 883–899.

Hagoort, P. (2003b). How the brain solves the binding problem for language: A neurocomputational model of syntactic processing. *Neuroimage, 20,* S18–S29.

Hagoort, P. (2005). On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences, 9,* 416–423.

Hagoort, P., & Brown, C. (1994). Brain responses to lexical ambiguity resolution and parsing. In L. Frazier, C. J. Clifton, & K. Rayner (Eds.), *Perspectives in sentence processing* (pp. 45–80). Hillsdale, NJ: Erlbaum.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304,* 438–441.

Holcomb, P. J., & McPherson, W. B. (1994). Event related potentials reflect semantic priming in an object decision task. *Brain and Cognition, 24,* 259–276.

Kelly, S. D., & Church, R. B. (1998). A comparison between children's and adults' ability to detect children's representational gestures. *Child Development, 69,* 85–93.

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language, 89,* 253–260.

Kendon, A. (2004). *Gesture: Visible action as utterance.* Cambridge: Cambridge University Press.

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. *Journal of Memory and Language, 48,* 16–32.

Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology, 61,* 743–754.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207,* 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307,* 161–163.

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

McNeill, D. (1992). *Hand and mind.* Chicago: University of Chicago Press.

McNeill, D. (2000). *Language and gesture.* Cambridge: Cambridge University Press.

McNeill, D., Cassell, J., & McCullough, K.-E. (1999). Communicative effects of speech-mismatched gestures. *Research on Language and Social Interaction, 27,* 223–238.

McPherson, W. B., & Holcomb, P. J. (1999). An electrophysiological investigation of semantic priming with pictures of real objects. *Psychophysiology, 36,* 53–65.

Melinger, A., & Levelt, W. J. M. (2004). Gesture and the communicative intention of the speaker. *Gesture, 4,* 119–141.

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts the temporal asynchrony of hand gesture and

speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 615–623.
Mottonnen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in human auditory cortex. *Cognitive Brain Research, 13,* 417–425.
Nigam, A., Hoffman, J. E., & Simons, R. F. (1992). N400 to semantically anomalous pictures and words. *Journal of Cognitive Neuroscience, 4,* 15–22.
Osterhout, L., McLaughlin, J., & Bersick, M. (1997). Event-related brain potentials and human language. *Trends in Cognitive Sciences, 1,* 203–209.
Özyürek, A. (2002). Do speakers design their co-speech gestures for their addressees? The effects of addressee location on representational gestures. *Journal of Memory and Language, 46,* 688–704.
Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., et al. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters, 127,* 141–145.
Singer, M. A., & Goldin-Meadow, S. (2005). Children learn when their teachers' speech and gestures differ. *Psychological Science, 16,* 85–89.
Sitnikova, T., Kuperberg, G., & Holcomb, P. J. (2003). Semantic integration in videos of real world events: An electrophysiological investigation. *Psychophysiology, 40,* 160–164.
Tanenhaus, M., Spivey-Knowlton, J. M., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268,* 1632–1634.
Van Berkum, J., Hagoort, P., & Brown, C. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of Cognitive Neuroscience, 11,* 657–671.
Van Berkum, J., Zwitserlood, P., Brown, C., & Hagoort, P. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research, 17,* 701–718.
West, W. C., & Holcomb, P. J. (2002). Event-related potentials during discourse-level semantic integration of complex pictures. *Cognitive Brain Research, 13,* 363–375.
Willems, R. M., Özyürek, A., & Hagoort, P. (in press). When language meets action: The neural integration of speech and gesture. *Cerebral Cortex*.
Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology, 42,* 654–667.