Brain and Language 185 (2018) 1–8

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l

Listening beyond seeing: Event-related potentials to audiovisual processing in visual narrative

Mirella Manfredi[a,*], Neil Cohn[b], Mariana De Araújo Andreoli[a], Paulo Sergio Boggio[a]

[a] Social and Cognitive Neuroscience Laboratory, Center for Biological Science and Health, Mackenzie Presbyterian University, São Paulo, Brazil
[b] Tilburg Center for Cognition and Communication, Tilburg University, Tilburg, Netherlands

ARTICLE INFO

Keywords:
Semantic memory
Cross-modal processing
Visual narrative
Audiovisual processing
N400

ABSTRACT

Every day we integrate meaningful information coming from different sensory modalities, and previous work has debated whether conceptual knowledge is represented in modality-specific neural stores specialized for specific types of information, and/or in an amodal, shared system. In the current study, we investigated semantic processing through a cross-modal paradigm which asked whether auditory semantic processing could be modulated by the constraints of context built up across a meaningful visual narrative sequence. We recorded event-related brain potentials (ERPs) to auditory words and sounds associated to events in visual narratives—i.e., seeing images of someone spitting while hearing either a word (*Spitting!*) or a sound (the sound of spitting)—which were either semantically congruent or incongruent with the climactic visual event. Our results showed that both incongruent sounds and words evoked an N400 effect, however, the distribution of the N400 effect to words (centro-parietal) differed from that of sounds (frontal). In addition, words had an earlier latency N400 than sounds. Despite these differences, a sustained late frontal negativity followed the N400s and did not differ between modalities. These results support the idea that semantic memory balances a distributed cortical network accessible from multiple modalities, yet also engages amodal processing insensitive to specific modalities.

1. Introduction

We live in a multisensory world and constantly integrate information from numerous sources of meaning. Researchers have long questioned whether semantic information associated to different types of stimuli is stored in a single "amodal" store or in a modality-specific knowledge system. Previous studies of event-related brain potentials (ERPs) suggested the existence of a single semantic store shared by pictures and words (Nigam, Hoffman, & Simons, 1992) while others showed slight variation to brain responses evoked by semantic information presented in different modalities (Ganis, Kutas, & Sereno, 1996; Nigam et al., 1992; Olivares, Iglesias, & Bobes, 1999; Van Petten & Rheinfeldert, 1995). However, the majority of these studies have investigated semantic processing of stimuli presented in the same sensory modality (Van Petten & Rheinfeldert, 1995) or by substituting an element from one modality for an element in another modality (Nigam et al., 1992; Ganis et al., 1996, Manfredi, Cohn, & Kutas, 2017). Only recently have researchers begun to examine the semantic system in the context of simultaneous presentation of meaningful stimuli presented in different sensory modalities (Hendrickson, Walenski, Friend, & Love, 2015; Liu, Wang, Wu, & Meng, 2011). To this aim, the present study investigates the simultaneous interaction of auditory and visual-graphic information, in the context of a visual narrative sequence.

Debates about the distribution of conceptual knowledge in the brain have focused on how semantic memory is represented in modality-specific stores specialized for specific types of information, and/or in an amodal, shared system. Modality-specific theories argue that conceptual knowledge is divided into anatomically distinct sensory and functional stores (Yee, Chrysikou, & Thompson-schill, 2013), organized in the brain by modality (visual, olfactory, motor/functional, etc.). In contrast, amodal accounts suggest a unitary system of conceptual organization (Caramazza, Hillis, Rapp, & Romani, 1990; Hillis & Caramazza, 1995; Rapp, Hillis, & Caramazza, 1993), where defining properties of an object are highly intercorrelated and members of a superordinate category share many common features (1975; Gelman & Coley, 1990; Keil, 1989; Markman, 1989; Rosch, 1973; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976).

Such views are mediated by the recent hub-and-spoke theory (Ralph, Jefferies, Patterson, & Rogers, 2016), which posits that concepts are built from multisensory knowledge encoded in modality-specific

* Corresponding author at: Social and Cognitive Neuroscience Laboratory, Center for Health and Biological Sciences, Mackenzie Presbyterian University, Rua Piaui, 181, São Paulo 01241-001, Brazil.
E-mail address: mirella.manfredi@gmail.com (M. Manfredi).

https://doi.org/10.1016/j.bandl.2018.06.008
Received 8 December 2017; Received in revised form 28 June 2018; Accepted 28 June 2018

cortices, which are distributed across the brain. These multisensory "spokes" come together in a single transmodal "hub" situated bilaterally in the anterior temporal lobes (ATLs), which mediates the cross-modal interactions between modality-specific sources of information.

A substantial body of literature has investigated cross-modal and multimodal semantic processing by analyzing the N400, an electrophysiological event-related brain potential (ERP) that peaks roughly 400 ms after the onset of a stimulus (Kutas & Federmeier, 2011). The N400 is thought to index the spreading activation in the access of semantic information by a stimulus in relation to its preceding context (Kutas & Federmeier, 2011). Within a single modality, the N400 has been observed in meaningful contexts to various levels of linguistic structure (e.g., Kutas & Hillyard, 1980, 1984; Camblin, Ledoux, Boudewy, Gordon, & Swaab, 2007; Bentin, McCarthy, & Wood, 1985), individual visual images (Van Berkum, Zwitserlood, Hagoort, & Brown, 2003; Van Berkum, Hagoort, & Brown, 1999; Ganis et al., 1996; Olivares et al., 1999; Proverbio & Riva, 2009; Bach, Gunter, Knoblich, Prinz, & Friederici, 2009), or sequences of images in visual narratives or events (Cohn, Paczynski, Jackendoff, Holcomb, & Kuperberg, 2012; Sitnikova, Holcomb, Kiyonaga, & Kuperberg, 2008; Sitnikova, Kuperberg, & Holcomb, 2003; West and Holcomb, 2002). In general, semantic processing is made easier by greater context, as indicated by attentuation of the N400 across ordinal sequence position for sentences (Van Petten & Kutas, 1991) or visual narrative sequences (Cohn et al., 2012).

The N400 may also differ between types of information presented in the same sensory system. For example, Van Petten and Rheinfeldert (1995) carried out a study which compared auditory (un)related pairs of sounds and words. Words and meaningful environmental sounds elicited similar N400 effects. However, they observed an asymmetric hemispheric laterality: right dominant for words and left dominant for environmental sounds. Similarly, scalp distributions also differ in the N400s to visual information, with images eliciting a widespread frontal distribution (Barrett & Rugg, 1990; Ganis et al., 1996; McPherson & Holcomb, 1999) and written words eliciting a more right posterior distribution (Kutas and Hillyard, 1984).

Studies analyzing cross-modal processing by using simultaneous presentation have suggested more complexity in semantic processing than monomodal studies. For example, gestures combined with inconsistent verbal information also elicit N400s (Cornejo et al., 2009; Habets, Kita, Shao, Ozyurek, & Hagoort, 2011; Özyürek, Willems, Kita, & Hagoort, 2007; Proverbio, Calbi, Manfredi, & Zani, 2014a; Wu and Coulson, 2005, 2007a, 2007b). Similar observations arise when speech and/or natural sounds are combined with semantically inconsistent pictures or video frames (Cummings, Ceponiene, Dick, Saygin, & Townsend, 2008; Liu et al., 2011; Plante, Petten, & Senkfor, 2000; Puce, Epling, Thompson, & Carrick, 2007). However, different types of inconsistent information modulate the N400 responses. For example, Liu et al. (2011) observed a larger magnitude and later latency N400 effect for videos with semantically inconsistent speech than those with semantically inconsistent natural sound.

Another method of comparing meaningful information across modalities has substituted elements of one modality into the structure of another to examine the extent that they access a common semantic system, particularly words and pictures. For example, N400s are elicited by pictures of objects that replace words in sentences (Ganis et al., 1996; Nigam et al., 1992). However, variation in the scalp distribution between images and words suggests processing by similar, albeit not identical brain regions.

In a recent study (Manfredi et al., 2017), we used the reverse method, by replacing words for a unit within a visual narrative sequence. Recent works have demonstrated that structural constraints govern visual narratives analogous to those in found in sentences (Cohn et al., 2012; Cohn, Jackendoff, Holcomb, & Kuperberg, 2014), and elicit electrophysiological responses similar to manipulations of linguistic syntax (Cohn and Kutas, 2015; Cohn et al., 2014). We thus took advantage of this structure by substituting a word for the climactic event in a visual narrative. We compared onomatopoeic words, which ostensibly imitate the sounds of actions (*Pow!*), with descriptive words, which describe an action (*Punch!*). Across two experiments, larger N400s appeared to onomatopoeic or descriptive words that were incongruent to their sequential context than to their congruent counterparts. However, despite the context in a visual narrative, these N400 effects appeared to have a more right posterior scalp distribution reminiscent of words, not pictures (Kutas and Hillyard, 1984). This work suggested that event comprehension in a visual narrative can be accessed across different domains, and in line with previous work, these results indicated that cross-model integration of semantics engage domain-independent integration/interpretation mechanisms.

In our prior work, we took advantage of multimodal aspects of semantic processing within a visual narrative by substituting a visual word for an omitted visual event. Building on this, the current study sought to investigate cross-modal processing when a preceding context builds an expectation of particular semantics, but in separate sensory modalities (auditory, visual). While some work has investigated simultaneous cross-modal information (sounds with images), and others have examined the build-up of context through a meaningful sequence (sentences, visual narratives), no works have yet combined these features. Therefore, we tested whether auditory semantic processing could be modulated by the constraints of context built up across a meaningful visual narrative sequence.

Here, we recorded ERPs to auditory words and sounds associated to events in visual narratives—i.e., seeing images of someone spitting while hearing either a word (*Spitting!*) or a sound (the sound of spitting), as in Fig. 1. Auditory words and sounds were either semantically congruent or incongruent with the climactic visual event. We expected attenuated N400s to congruent alignment between the word and sound with its visual event, compared to large N400s to incongruent words or sounds. Such results would further suggest that cross-modal information relies on an integrated system of semantic processing.

In addition, if the semantic system is affected by stimulus types, we could expect different processing based on the nature of the incoming stimuli: sounds of events may result in different meaningful connections than words of events. Our previous work found no difference between the N400s elicited by onomatopoetic and descriptive words (Manfredi et al., 2017), while hemispheric differences in scalp distribution have been observed between words and meaningful environmental sounds (Van Petten & Rheinfeldert, 1995). In light of these findings, we tested the possibility that the type of auditory stimulus may modulate the cross-modal processing of visual events. If such differences occur, we might expect to observe a different distribution of the N400 effect—but not a different ERP response altogether—suggesting the existence of a distributed cortical network accessible from multiple modalities (Kutas & Federmeier, 2000).

2. Methods

2.1. Stimuli

We designed 100 novel 4 panel long visual narrative sequences using black and white panels from the *Complete Peanuts* volumes 1 through 6 (1950–1962) by Charles Schulz (Fantagraphics Books, 2004–2006), modified from sequences created for prior studies (Cohn & Kutas, 2015; Cohn & Kutas, 2015; Cohn & Wittenberg, 2015; Cohn et al., 2012; Manfredi et al., 2017). To eliminate the influence of written language, we used panels either without text or with text deleted. All panels were adjusted to a single uniform size. All sequences had a coherent narrative structure, as defined by the theory of Visual Narrative Grammar (Cohn, 2014), and confirmed by behavioral ratings (Cohn et al., 2012).

We combined these base sequences with a sound or an auditory word timelocked to the climactic image (Peak) of each strip. Congruent

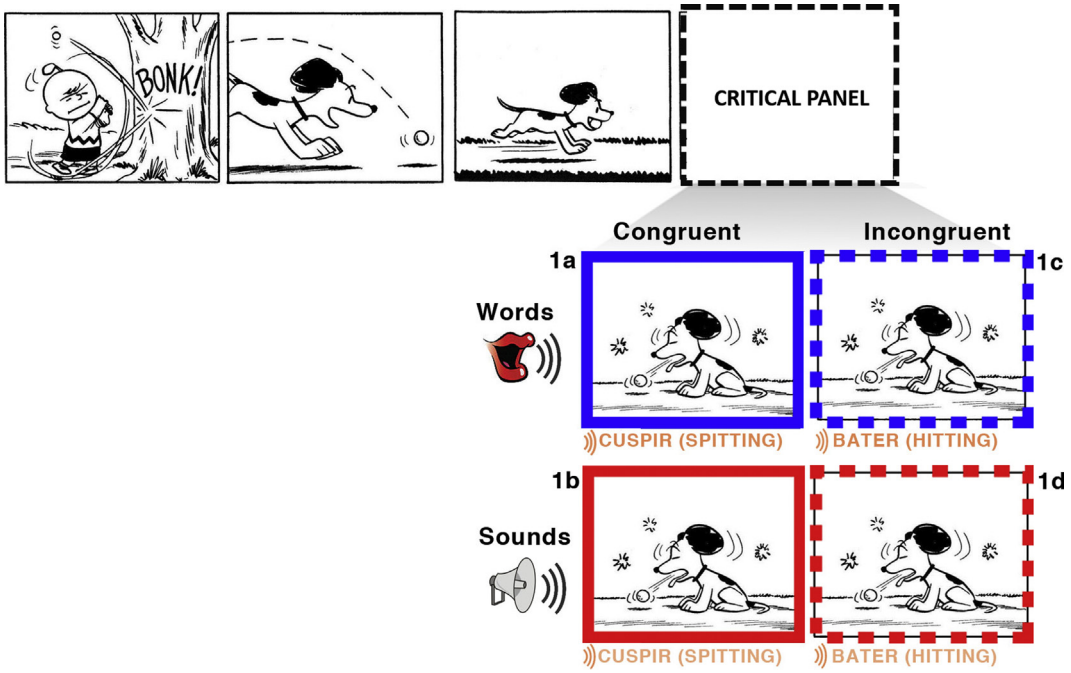

Fig. 1. Example of visual sequences used as experimental stimuli. We manipulated these base sequences by adding a congruent word, an incongruent word, a congruent sound and incongruent sound.

sequences matched climatic pictures with auditory Portuguese words that described familiar actions, or with sounds that corresponded to the same actions. We presented 27 auditory Portuguese words and 23 sounds. The latter included environmental (n = 11), human (n = 9) and animal sounds (n = 3). We then created four sequence types by modulating the type of stimulus (words vs. sounds) and their congruence with the visual images (congruent vs. incongruent). As in Fig. 1, *congruent word* panels (1a) contained an auditory word coherent with the contents of the image, *congruent sound* panels (1b) used a sound coherent with the image, *incongruent word* panels (1c) contained an auditory word incoherent with the image, and *incongruent sound* panels (1d) used a sound incoherent with the image. The critical panels appeared in the second to the fourth panel positions, with equal numbers at each position. Some words and sounds were repeated across the sequences. The average values (number of repetitions) were not significantly different across the conditions ($t(47) = 1.10$, $p > 0.05$) (Words = 4.43, SD = 1.40; Sounds = 3.92, SD = 1.78).

A female native Portuguese speaker produced the word stimuli (mean duration = 996 ms, SD = 72.01 ms), which were recorded in a single session in a sound attenuating booth. Environmental sound stimuli (mean duration = 985 ms, SD = 42.31 ms) were obtained from several online sources. Word stimuli and environmental sounds were standardized for sound quality (44.1 kHz, 16 bit, stereo). A *t*-test revealed no differences ($p = 0.5$) between the duration of words (996 ms, SD = 72.01 ms) and environmental sounds (985 ms, SD = 42.31 ms).

Pre-assessment of stimuli were made by a group of 8 judges of similar age (mean age = 22.25, SE = 2.25) and educational level as the experimental subjects. Congruent sequences rated as incoherent by at least 80–99% of judges were discarded, as were incongruent sequences evaluated as coherent. Our final stimulus set included 100 experimental sequences (~25 per condition). A total of four lists (each consisting of 100 strips in random order) were created, with the four conditions counterbalanced using a Latin Square Design such that participants viewed each sequence only once in a list.

2.2. Participants

Twenty-four undergraduate students (9 males) were recruited from the Mackenzie Presbyterian University, São Paulo, Brazil. Participants were native Portuguese speakers (mean age = 22.33, SE = 4.6), had normal or corrected-to-normal vision, and reported no history of neurological illness or drug abuse. Handedness was assessed by the Portuguese version of the Edinburgh Handedness Inventory (Oldfield, 1971), a laterality preference questionnaire reported right-handedness dominance for all participants. The study adhered to the Declaration of Helsinki guidelines and was approved by the institutional ethics committee of Mackenzie Presbyterian University, Brazil, and registered with the National Ethics Committee. All the participants provided written, informed consent. All participants knew *Peanuts*.

2.3. Procedure

Participants sat in front of a monitor in a sound-proof, electrically-shielded recording chamber. Before each strip, a fixation cross appeared for a duration of 1500 ms. The strips were presented panel-by-panel in the center of the monitor screen and the sound or the auditory word was timelocked to the climactic image (Peak) of each strip. Panels stayed on screen for 1200 ms, separated by an ISI of 300 ms (e.g., Cohn & Kutas, 2015). When the strip concluded, a question mark appeared on the screen and participants indicated whether the strip was easy or hard to understand by pressing one of the two hand-held buttons. Response hand was counterbalanced across participants and lists.

Participants were instructed not to blink or move during the experimental session. The experiment had five sections separated by breaks. Experimental trials were preceded by a short practice to familiarize participants with the procedures.

2.4. Electroencephalographic recording parameters

The electroencephalogram (EEG) was recorded from 128 electrodes at a sampling rate of 250 Hz (bandpass 0.01–100 Hz). The EEG was recorded and analyzed using the Net station software (*Geodesic EEG Net Station, EGI*, Eugene, OR). The impedance of all electrodes was kept below 50 kΩ over the experiment. All recordings were referenced to Cz electrode during data acquisition. This solution allowed us to analyze the mastoid-temporal lobe activity in addition to all other important

sites for the linguistic processing. EEG epochs were synchronized with the onset of stimuli presentation.

2.5. Statistical analysis of ERPs

Trials contaminated by blinks, muscle tension (EMG), channel drift, and/or amplifier blocking were discarded before averaging. Approximately 9% of critical panel epochs were rejected due to such artifacts, with losses distributed approximately evenly across the four conditions. Each participant's EEG was time-locked to the onset of critical panels and their accompanying auditory stimuli, and ERPs were computed for epochs extending from 100 ms before stimulus onset to 1500 ms after stimulus onset.

Our analysis focused on three epochs of interest. We investigated the mean amplitude voltage and latency of the N1 in the 100–200 ms epoch and of the N400 in the 350–550 ms epoch. A subsequent time window of 550–750 ms was examined to investigate any later or sustained effects. These responses were measured at frontal (18, 16, 10, 19, 11, 4, 22, 23, 9, 3, 124, 24), central (13, 6, 112, 7, 106, 31, 80, 55, 30, 105, 87, 37) and posterior (61, 62, 78, 67, 72, 77, 71, 76, 75, 81, 70, 83) electrode sites.

Mean amplitude of each component was analyzed using repeated-measures ANOVAs with factors of Congruency (2 levels: Congruent, Incongruent), Modality (2 levels: words, sounds) and Region (anterior, central, posterior). Multiple comparisons of means were performed with post-hoc Fisher's tests.

3. Results

3.1. Behavioral results

Overall, a 2 × 2 ANOVA computed on ratings revealed a significant main effect of Congruency ($F(1, 21) = 252.95$, $p < 0.01$, partial Eta squared = 0.92), arising because congruent sounds/words were rated as more coherent that the incongruent ones. There was no main effect of Modality (p = n.s.). In addition, a Modality × Congruency interaction ($F(1, 21) = 15.519$, $p < 0.01$, partial Eta squared = 0.42) showed that sequences with incongruent words (16%, SE = 3.08), were considered significantly less coherent ($p < 0.01$) than those with incongruent sounds (24%, SE = 3.41), which were both less coherent than congruent sounds (79%, SE = 2.00) and congruent words (83%, SE = 2.18) ($p < 0.05$). No differences were found between sequences with congruent sounds and words ($p > 0.05$).

3.2. Electrophysiological results

3.2.1. N1 (100–200 ms)

Mean amplitude of the N1 component revealed a significant Modality × Region interaction ($F(2, 46) = 6.56$, $p < 0.05$, partial Eta squared = 0.22), showing a greater amplitude negativity to words than sounds ($p < 0.01$) in the frontal areas compared to the central and the parietal ones. No differences were observed in the latency of the ERP responses to words and sounds (p = n.s). In addition, we observed no main effects or interactions with Congruency.

3.2.2. N400 (350–550 ms)

Mean amplitude of the N400 component showed a main effect of Congruency ($F(1, 23) = 30.60$, $p < 0.05$, partial Eta squared = 0.57), revealing a greater amplitude negativity to incongruent than congruent stimuli ($p < 0.01$). In addition, a near significant main effect of Modality ($F(1, 23) = 3.98$, $p = 0.05$, partial Eta squared = 0.14) suggested that words were more negative compared to sounds (Fig. 2).

A Congruency × Modality × Region interaction ($F(2, 46) = 3.15$; $p < 0.05$, partial Eta squared = 0.12) showed that the N400 amplitude was more negative in response to incongruent words than congruent word panels only in the centro-parietal areas ($p < 0.01$). No differences were found between the N400 response to congruent and incongruent word panels in the frontal sites (p = n.s.). Conversely, the N400 response to incongruent sound panels was greater than the N400 response to congruent sound panels only in the front-central sites ($p < 0.05$) (Figs. 3 and 4).

Analysis of mean latency of the N400 component revealed a main effect of Modality ($F(1, 23) = 6.41$, $p < 0.01$, partial Eta squared = 0.21), with a later N400 evoked in response to sounds as compared as words. In addition, a main effect of Congruency ($F(1, 23) = 15.37$, $p < 0.01$, partial Eta squared = 0.40), suggested a later N400 in response to incongruent than congruent stimuli. Moreover, the Modality x Region interaction ($F(2, 46 = 7.39$; $p < 0.01$, partial Eta squared = 0.24) showed a later N400 response to sounds than congruent words and incongruent words in the frontal and central sites. Finally, the Congruency x Region interaction ($F(2, 46) = 8.30$, $p < 0.01$, partial Eta squared = 0.26) revealed a later N400 response to incongruent than congruent stimuli just in the centro-parietal areas.

3.2.3. Frontal negativity (550–750 ms)

A frontal negativity was suggested by a main effect of Congruency ($F(1, 23 = 25.88$, $p < 0.01$; partial Eta squared = 0.53) and was greater in response to incongruent critical panels compared to congruent ones. The Congruency x Region interaction ($F(2, 46 = 3.63$; $p < 0.01$; partial Eta squared = 0.13) revealed that the Incongruent panels were more negative than the congruent ones over the frontal sites than central and posterior ones.

4. Discussion

In this study, we investigated how semantic auditory information interacts in a simultaneous multimodal presentation with the context of a visual narrative sequence. To this aim we recorded ERPs to auditory words and sounds that were either semantically congruent or incongruent with a cross-modally presented climactic event in a visual narrative sequence. Our results showed that both incongruent sounds and words elicited an N400 effect, however, we observed a different distribution of the N400 effect between modalities. This result aligns with the idea that processing differs based on the nature of the incoming auditory stimuli, and suggests the existence of a distributed cortical network accessible from multiple modalities (Kutas & Federmeier, 2000; Ralph et al., 2016). Below, we elaborate on this interpretation.

First, a greater N1 amplitude appeared to words than sounds, but no differences were observed regarding congruency. Since the N1 component reflects sensory processing (Näätänen & Winkler, 1999), this result confirmed that the sensory processing is sensitive only to the modality difference, and not the congruency. This suggested that the semantic processing of congruency, as in the subsequent N400 which was also sensitive to modality (below), occurred later than the basic sensory processing of words versus sounds.

Primary evidence of a distributed semantic network appeared in the 350–550 ms time window, where we observed two effects: a frontal N400 effect (sounds) and a centro-parietal N400 effect (words). The larger N400s to incongruent auditory stimulus—regardless of type—-with visual information is consistent with the idea of a more difficult retrieval process for semantic information that is discordant or unexpected (Kutas & Federmeier, 2011). Such results are consistent with findings of N400s to anomalies monomodally to words in sentences (e.g., Bentin, et al., 1985; e.g., Kutas & Hillyard, 1980, 1984; Camblin, et al., 2007), to words in discourse (van Berkum et al., 1999, 2003), to visual images (Bach et al., 2009; Proverbio & Riva, 2009), to visual event sequences (Cohn et al., 2012; Sitnikova et al., 2003; Sitnikova, et al., 2008; West & Holcomb, 2002), and cross-modally between speech and gesture (Cornejo et al., 2009; Habets et al., 2011; Özyürek et al., 2007; Proverbio et al., 2014a; Wu and Coulson, 2005, 2007a, 2007b) and between speech or natural sounds with pictures or videos (Cummings et al., 2008; Liu et al., 2011; Plante et al., 2000; Puce et al.,

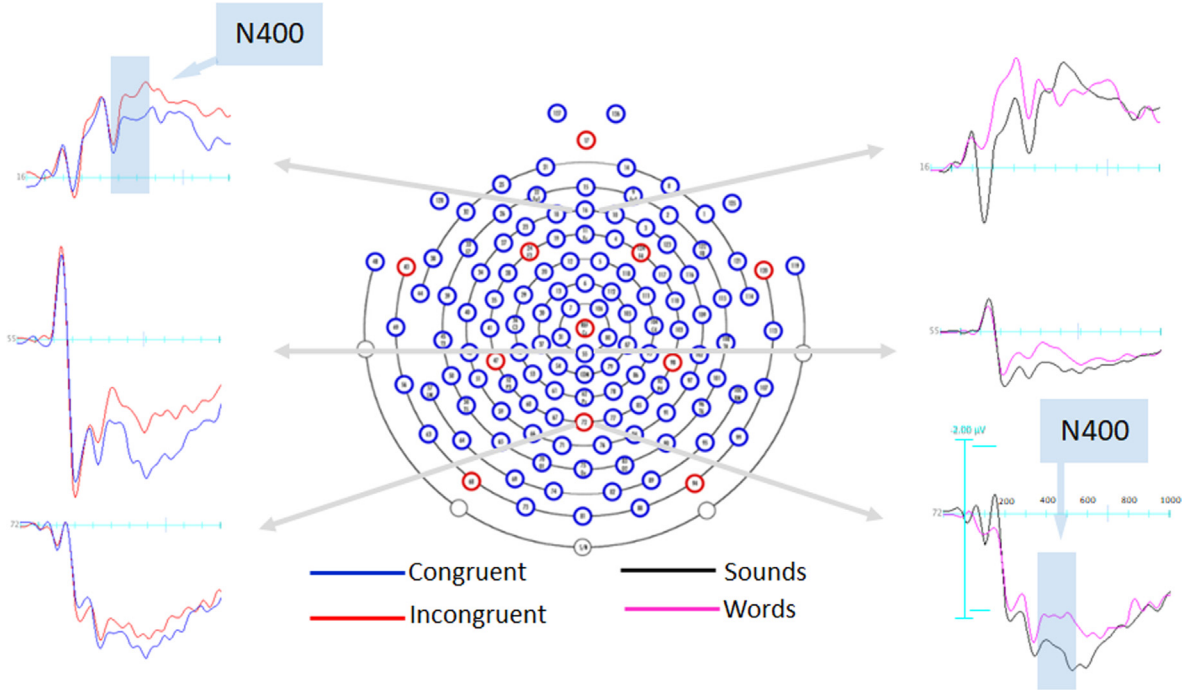

Fig. 2. Grand-average ERP waveforms recorded at frontal, central and posterior midline sites in response to Congruent (blue) and Incongruent (red) critical panels and to Words (violet) and Sounds (black) critical panels. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

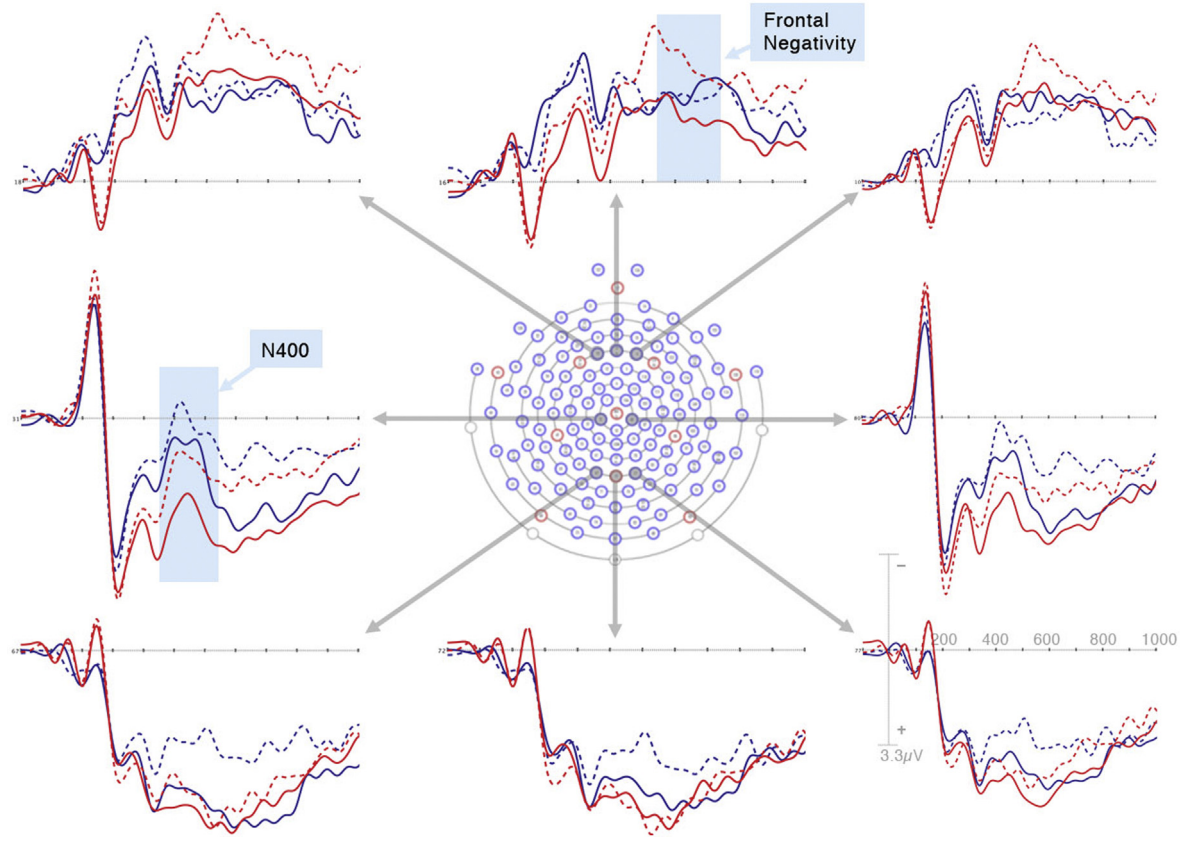

Fig. 3. Grand-average ERP waveforms recorded at frontal, central and posterior sites in response to Congruent word (solid blue line), Incongruent word (dotted blue line), Congruent sound (solid red line) and Incongruent sound (dotted red line) critical panels. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

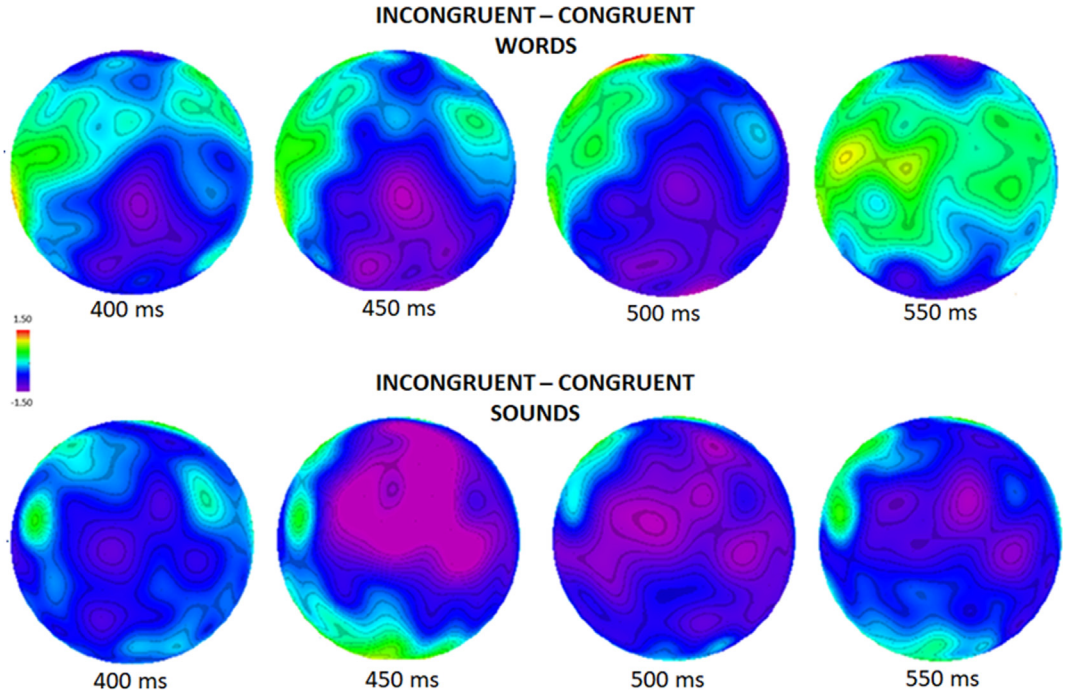

Fig. 4. Voltage of scalp distribution of the N400 in response to the difference between congruent and incongruent words or sounds.

2007).

Nevertheless, these N400 effects differed based on the type of information presented (words/sounds): the incongruency effect to *words* paired with images had a centro-parietal scalp distribution, while that to *sounds* had a more fronto-central distribution. Despite presenting words and sounds simultaneously with images in a cross-modal design, the relative differences between N400s to words and sounds appear similar to what have been observed in monomodal presentations. When words appear in isolation, the N400-effect is typically distributed over centro-parietal sites, both for words in sentences and presented individually in priming (Kutas & Federmeier, 2011). However, the N400-effect for auditory words has been observed to have a slightly more frontal distribution than their written counterparts (Domalski, Smith, & Halgren, 1991; Holcomb and Neville, 1990; Kutas & Van Petten 1994; Kutas and Federmeier, 2011; McCallum, Farmer, & Pocock, 1984). This centro-parietal distribution is consistent with our findings when words are presented along with visual images in a sequence.

Our findings also align with findings that the N400 differs between words and sounds. For example, as in previous work (Ganis et al., 1996; Nigam et al., 1992; Cummings et al., 2006; Cummings et al., 2008), we observed a slightly more posterior N400 effect in response to words than environmental sounds. However, other work has observed N400s to words evoking larger responses in the right hemisphere and environmental sounds eliciting larger responses in the left hemisphere (e.g., Van Petten & Rheinfeldert, 1995; Plante et al., 2000). We found no such laterality differences. These differences in scalp distribution are consistent with variance between other modalities, such as the responses to visual words, which have a centro-parietal distribution, and pictures, which have a fronto-central distribution (Ganis et al., 1996; Holcomb & McPherson, 1994; McPherson & Holcomb, 1999).

Overall, this variation in scalp distribution, despite the consistency of the N400 across domains, suggest the involvement of partially non-overlapping neural structures underlying the processing of different information (Cummings et al., 2006, 2008; Plante et al., 2000; Van Petten and Rheinfeldert, 1995). It is important to underline that the above-mentioned previous comparisons between N400s of different modalities did not present different types of stimuli simultaneously, but rather analyzed semantic processing of stimuli within the same sensory modality (i.e. vision), albeit from different systems of communication (i.e., words/sounds, words/images). Our study compared such auditory information while also presenting participants with visual information from an image sequence (see also Liu et al., 2011). Yet, since in our study the visual stimuli were held constant, the semantic processing differed based on the words/sound contrast. In fact, as discussed above, our findings did not highlight substantial differences in the interaction between either words/images or sound/images in terms of scalp distribution from N400s presented without images (Ganis et al., 1996; Nigam et al., 1992; Cummings et al., 2006, 2008). This may suggest that the features activated by the visual sequence in semantic memory would not differentially affect the ones for words or sounds.

Interestingly, these scalp distributions are similar to those observed in our previous studies (Manfredi et al., 2017) in which we substituted a word for the climactic event in a visual narrative. There, the distribution of the N400 to the words looked more similar to word-N400s than image-N400s, despite being in the context of a pictorial visual sequence. Nevertheless, as observed there, the semantics of the sequence here did influence the processing of these stimuli, given that the observed congruency effect only arises due to the cross-modal relationship of the auditory and visual information. This implies that certain unique features of semantic memory are activated based on specific features of the stimuli (words vs. sounds), but they also share features across modalities, such that they can be modulated by a visual context.

In addition to scalp distributions, we also observed differences in the latency of the ERP responses. In particular, we observed that the N400 amplitudes to sounds peaked at a later latecy (550 ms) compared to words (400 ms), albeit not modulated by congruency. This might suggest that sound processing would require the recovery of the meaning of the sound and its association with the visual event. We could speculate that during a word/visual-event presentation, the meaning of the visual event is matched with the meaning of the word that is punctual, specific, and unambiguous because it describes a specific action. In contrast, a sound always needs to be associated to something else like a visual/hidden action, i.e., the meaning of the visual event is integrated with the meaning of the sound. In fact, since an action usually produces a natural sound that is semantically congruent with the associated visual event, it should be easy to understand.

On the other hand, when the sound is not congruent with the visual event, more effort would be needed to match it with the visual context, possibly taking more time. In a previous work, Liu et al. (2011) found that word/video combinations were later than sound/video ones. However, the visual stimuli presented in their work were real-word events without a narrative structure. In this context, the word/video combination could be considered a less common representation of cross-modal interactions than the sound/video one, and so it might have required more cognitive processing to be comprehended. According to this interpretation, in their study the N400 latency delay may reflect this additional effort. Therefore, it is possible that the nature of visual stimuli (static drawings vs. dynamic video) would differently affect the comprehension of multisensory information, but would require further study.

Following the N400, we observed a sustained fronto-central negativity in the 550–750 ms time window that was larger to incongruous than congruous stimuli, as in the preceding N400. Similar sustained negativities have been observed following N400s in response to visual anomalies in sequences of visual narratives or visual events (West & Holcomb, 2002; Cohn et al., 2012). The frontal negativity observed here to incongruency did not differ in distribution between sounds/words. Thus, this sustained frontal negativity could reflect a general cost of further processing the inconsistent interaction between auditory and visual information, regardless of the word/sound distinction.

To summarize our results, this study showed that sounds and words combined with visual events elicit similar processes of semantic processing (N400) but with different characteristics. In particular, the different latency and distribution of the N400 effects might reflect the different nature of these stimuli. This is in line with previous not-simultaneous cross-modal studies (Ganis et al., 1996; Nigam et al., 1994; Cummings et al., 2006, 2008) that revealed differences in the scalp distributions of the congruity effects for different stimuli. Yet, we also observed a late frontal negativity effect that did not differ in scalp distribution across modalities, implying a more general processing mechanism. Our findings might indicate that different sensory information converge into an amodal store across a time-course of processing (Ralph et al., 2016). Under this interpretation, the N400 epochs could reflect the access of semantic memory for different cortical systems, while the late negativity could represent amodal processing integrating these cross-modal signals into a single understanding. Nevertheless, research would be needed to further investigate such interpretations.

Researchers have long questioned whether conceptual knowledge is represented in modality-specific semantic stores (Caramazza & Shelton, 2009), or in an amodal, shared system (Caramazza et al., 1990; Hillis & Caramazza, 1995; Rapp et al., 1993). Previous ERP studies suggested the existence of a single semantic store shared by pictures and words (Nigam et al., 1992). Indeed, it has been observed that interactions between different modalities may modulate the N400 response. For example, a reduced N400 occurs from integrating different types of stimulus with a related prime (Barrett & Rugg, 1990; Bentin et al., 1985; Holcomb & McPherson, 1994) or a congruent sentence context (Ganis et al., 1996; Kutas & Hillyard, 1980, 1984; Nigam et al., 1992). On the other hand, the N400 responses to words and pictures differed in scalp distribution, implicating non-identical neural generators (Ganis et al., 1996; Holcomb & McPherson, 1994). In addition, Federmeier et al. (2007), in light of work comparing N400s between pictures and words, concluded that semantic processing is not amodal.

Our findings further support that the nature of the semantic information is able to affect the semantic processing across different modalities. In addition, when different stimuli are presented simultaneously, they maintained a relatively separate distribution in semantic memory, rather than linking up with the visual semantics in some particularly varying way. Therefore, the sound/visual-event and word/visual-event stimuli may index different processing in the semantic system, but without some emergent added value. This might suggest that different stimuli modalities involve separate (but connected) distributions within semantic memory. Such results open questions about how semantic memory manages additional meanings created by the union of different stimuli, particularly those where emergent multimodal inferences go beyond the contributions of each individual modality's meaning (e.g., Forceville and Urios-Aparisi, 2009; Özyürek et al., 2007; Habets et al., 2011; Wu and Coulson, 2005; 2007a; 2007b; Cornejo et al., 2009; Proverbio et al., 2014a).

In conclusion, our results showed that both incongruent sounds and words evoked an N400 effect, however we observed a different latency and distribution of the N400 effect according to the types of sensory information. Nevertheless, a sustained late frontal negativity followed the N400s and did not differ between modalities. These results support the idea semantic memory balances a distributed cortical network accessible from multiple modalities, yet also involves amodal system insensitive to specific modalities.

Acknowledgments

Paulo S. Boggio is supported by a CNPq research grant (311641/2015-6). Mirella Manfredi is supported by a FAPESP post-doctoral researcher grant (2015/00553-5).

References

Bach, P., Gunter, T. C., Knoblich, G., Prinz, W., & Friederici, A. D. (2009). N400-like negativities in action perception reflect the activation of two components of an action representation. *Social Neuroscience, 4*, 212–232.
Barrett, S. E., & Rugg, M. D. (1990). Event-related potentials and the semantic matching of pictures. *Brain and Cognition, 2*, 201–212.
Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision, and semantic priming. *Electroencephalography & Clinical Neurophysiology, 60*, 353–355.
Camblin, C. C., Ledoux, K., Boudewy, M., Gordon, P. C., & Swaab, T. Y. (2007). Processing new and repeated names: Effects of coreference on repetition priming with speech and fast RSVP. *Brain Research, 1146*, 172–184.
Caramazza, A., Hillis, A. E., Rapp, B. C., & Romani, C. (1990). The multiple semantics hypothesis: Multiple confusions? *Cognitive Neuropsychology, 7*, 161–189.
Caramazza, A., & Shelton, J. R. (2009). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience, 10*, 1–34.
Cohn, N. (2014). You're a good structure, Charlie Brown: The distribution of narrative categories in comic strips. *Cognitive Science, 38*(7), 1317–1359.
Cohn, N., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2014). The grammar of visual narrative: Neural evidence for constituent structure in sequential image comprehension. *Neuropsychologia, 64*, 63–70.
Cohn, N., & Kutas, M. (2015). Getting a cue before getting a clue: Event-related potentials to inference in visual narrative comprehension. *Neuropsychologia, 77*, 267–278.
Cohn, N., Paczynski, M., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2012). (Pea) nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive Psychology, 65*(1), 1–38.
Cohn, N., & Wittenberg, E. (2015). Action starring narratives and events: Structure and inference in visual narrative comprehension. *Journal of Cognitive Psychology, 27*(7), 812–828.
Cornejo, C., Simonetti, F., Ibáñez, A., Aldunate, N., Ceric, F., López, V., & Núñez, R. E. (2009). Gesture and metaphor comprehension: Electrophysiological evidence of cross-modal coordination by audiovisual stimulation. *Brain and Cognition, 70*(1), 42–52.
Cummings, A., Ceponiene, R., Dick, F., Saygin, A. P., & Townsend, J. (2008). A developmental ERP study of verbal and non-verbal semantic processing. *Brain Research, 1208*, 137–149.
Cummings, A., Ceponiene, R., Koyama, A., Saygin, A. P., Townsend, J., & Dick, F. (2006). Auditory semantic networks for words and natural sounds. *Brain Research, 1115*(1), 92–107.
Domalski, P., Smith, M. E., & Halgren, E. (1991). Cross-modal repetition effects on the N4. *Psychological Science, 2*, 173–178.
Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research, 1146*, 75–84.
Forceville, C., & Urios-Aparisi, E. (2009). *Multimodal metaphor*. New York: Mouton De Gruyter.
Ganis, G., Kutas, M., & Sereno, M. I. (1996). The search for "common sense": An electrophysiological study of the comprehension of words and pictures in reading. *Journal of Cognitive Neuroscience, 8*(2), 89–106.
Gelman, S., & Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology, 26*, 796–804.
Habets, B., Kita, S., Shao, Z., Ozyurek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech-gesture integration during comprehension. *Journal of Cognitive Neuroscience, 23*(8), 1845–1854.
Hendrickson, K., Walenski, M., Friend, M., & Love, T. (2015). The organization of words

and environmental sounds in memory. *Neuropsychologia, 69*, 67–76.
Hillis, A., & Caramazza, A. (1995). Cognitive and neural mechanisms underlying visual and semantic processing: Implication from "optic aphasia". *Journal of Cognitive Neuroscience, 7*, 457–478.
Holcomb, P. J., & McPherson, W. B. (1994). Event related potentials reflect semantic priming in an object decision task. *Brain and Cognition, 24*, 259–276.
Holcomb, P. J., & Neville, H. J. (1990). Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language and Cognitive Processes, 5*, 281–312.
Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.
Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Science, 4*, 463–470.
Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*, 621–647.
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*, 203–208.
Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*, 161–163.
Kutas, M., & Van Petten, C. (1994). Psycholinguistics electrified: Event-related potential investigations. In M. A. Gernsbacher (Ed.). *Handbook of Psycholinguistics* (pp. 83–143). Academic Press.
Liu, B., Wang, Z., Wu, G., & Meng, X. (2011). Cognitive integration of asynchronous natural or non-natural auditory and visual information in videos of real-world events: An event-related potential study. *Neuroscience, 180*, 181–190.
Manfredi, M., Cohn, N., & Kutas, M. (2017). When a hit sounds like a kiss: An electrophysiological exploration of semantic processing in visual narrative. *Brain and Language, 169*, 28–38.
Markman, E. M. (1989). *Categorization and naming in children: Problems of induction.* Cambridge, MA: MIT Press.
McCallum, W. C., Farmer, S. F., & Pocock, P. V. (1984). The effects of physical and semantic incongruities on auditory event-related potentials. *Electroencephalograpby and Clinical Neurophysiology, 59*, 477–488.
McPherson, W. B., & Holcomb, P. J. (1999). An electrophysiological investigation of semantic priming with pictures of real objects. *Psychophysiology, 36*(1), 53–65.
Näätänen, R., & Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychological Bulletin, 125*(6), 826–859.
Nigam, A., Hoffman, J. E., & Simons, R. F. (1992). N400 to semantically anomalous pictures and words. *Journal of Cognitive Neuroscience, 4*(1), 15–22.
Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*, 97–113.
Olivares, E. I., Iglesias, J., & Bobes, M. A. (1999). Searching for face-specific long latency ERPs: A topographic study of effects associated with mismatching features. *Cognitive Brain Research, 7*, 343–356.
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. *Journal of Cognitive Neuroscience, 19*, 605–616.
Plante, E., Petten, C. V., & Senkfor, J. (2000). Electrophysiological dissociation between verbal and nonverbal semantic processing in learning disabled adults. *Neuropsychologia, 38*(13), 1669–1684.
Proverbio, A. M., Calbi, M., Manfredi, M., & Zani, A. (2014a). Comprehending body language and mimics: An ERP and neuroimaging study on italian actors and viewers. *PLoS One, 9*(3), e91294.
Proverbio, A. M., & Riva, F. (2009). RP and N400 ERP components reflect semantic violations in visual processing of human actions. *Neuroscience Letters, 459*, 142–146.
Puce, A., Epling, J. A., Thompson, J. C., & Carrick, O. K. (2007). Neural responses elicited to face motion and vocalization pairings. *Neuropsychologia, 45*(1), 93–106.
Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2016). *Nature Reviews Neuroscience, 18*, 42–55.
Rapp, B. C., Hillis, A. E., & Caramazza, A. C. (1993). The role of representations in cognitive theory: More on multiple semantics and the agnosias. *Cognitive Neuropsychology, 10*, 235–249.
Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.). *Cognitive development and the acquisition of language* (pp. 111–144). New York: Academic Press.
Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192–233.
Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes- Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382–439.
Sitnikova, T., Holcomb, P. J., Kiyonaga, K. A., & Kuperberg, G. R. (2008). Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. *Journal of Cognitive Neuroscience, 20*, 1–21.
Sitnikova, T., Kuperberg, G., & Holcomb, P. J. (2003). Semantic integration in videos of real-world events: An electrophysiological investigation. *Psychophysiology, 40*(1), 160–164.
Van Berkum, J. J., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of Cognitive Neuroscience, 11*(6), 657–671.
Van Berkum, J. J., Zwitserlood, P., Hagoort, P., & Brown, C. M. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research, 17*(3), 701–718.
Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open and closed class words. *Memory and Cognition, 19*, 95–112.
Van Petten, C., & Rheinfeldert, H. (1995). Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia, 33*(4), 485–508.
West, W. C., & Holcomb, P. J. (2002). Event-related potentials during discourse-level semantic integration of complex pictures. *Cognitive Brain Research, 13*, 363–375.
Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology, 42*(6), 654–667.
Wu, Y. C., & Coulson, S. (2007b). How iconic gestures enhance communication: An ERP study. *Brain & Language, 101*, 234–245.
Wu, Y. C., & Coulson, S. (2007a). Iconic gestures prime related concepts: An ERP study. *Psychonomic Bulletin & Review, 14*, 57–63.
Yee, E., Chrysikou, E. G., Thompson-schill, S. L. (2013). The cognitive neuroscience of semantic memory, 1–16. In Stephen Kosslyn, Kevin Ochsner (Eds.), Oxford Handbook of Cognitive Neuroscience. Oxford University Press.

Further reading

Proverbio, A. M., Calbi, M., Manfredi, M., & Zani, A. (2014b). Audio-visuomotor processing in the Musician's brain: An ERP study on professional violinists and clarinetists. *Scientific Reports, 4*, 5866.