Monitoring in Language Perception: Mild and Strong Conflicts Elicit Different ERP Patterns

Nan van de Meerendonk[1], Herman H. J. Kolk[1], Constance Th. W. M. Vissers[2], and Dorothee J. Chwilla[1]

Abstract

■ In the language domain, most studies of error monitoring have been devoted to language production. However, in language perception, errors are made as well and we are able to detect them. According to the monitoring theory of language perception, a strong conflict between what is expected and what is observed triggers reanalysis to check for possible perceptual errors, a process reflected by the P600. This is at variance with the dominant view that the P600 reflects syntactic reanalysis or repair, after syntactic violations or ambiguity. In the present study, the prediction of the monitoring theory of language perception was tested, that only a strong conflict between expectancies triggers reanalysis to check for possible perceptual errors, reflected by the P600. Therefore, we manipulated plausibility, and hypothesized that when a critical noun is mildly implausible in the given sentence (e.g., "The eye consisting of among other things a pupil, iris, and eyebrow . . ."), a mild conflict arises between the expected and unexpected event; integration difficulties arise due to the unexpectedness but they are resolved successfully, thereby eliciting an N400 effect. When the noun is deeply implausible however (e.g., "The eye consisting of among other things a pupil, iris, and sticker . . ."), a strong conflict arises; integration fails and reanalysis is triggered, eliciting a P600 effect. Our hypothesis was confirmed; only when the conflict between the expected and unexpected event is strong enough, reanalysis is triggered. ■

INTRODUCTION

Monitoring refers to the process of watching over the quality of ones behavior (e.g., McGuire, Silbersweig, & Frith, 1996; Stuss & Benson, 1986). Monitoring is an aspect of executive control, which is a process that becomes necessary when different responses compete to be selected in a certain situation (e.g., Gazzaniga, Ivry, & Mangun, 2002). Monitoring entails the triggering of corrective actions whenever there is a conflict between what is planned and what is observed.

An example of the effect of monitoring in the language domain can be seen in the case of speech errors. When people produce a speech error this often leads to overt self-repairs, such as "Go from left again to uh. . ., from pink again to blue" (Levelt, 1983). This example shows that we are able to detect the discrepancy, in this case, between the speech element that we intended to produce and the actual speech element that was produced, and are able to correct for our mistake. According to Levelt (1983), self-monitoring has two important functions: first, to match the intended and produced message, and second, to create instructions to adjust the message.

Many studies of monitoring in the language domain have been devoted to language production. However, in language perception, errors are made as well and we are able to detect them. Cutler and Butterfield (1992) call these perception errors "slips of the ear." For example, a speaker may produce "Into opposing camps," which the listener may erroneously perceive as "Into a posing camp." Cutler and Butterfield looked at these errors in the light of how language (in particular the English language) is perceived and misperceived, but the question remains of how these errors are detected. To the best of our knowledge, perception errors have been documented, but there has been no research on how these errors are monitored for. Because the listener does not know the intentions of the speaker, perception errors cannot be observed directly by comparing the intended with the produced message. How is it then that perceptual errors are detected? In the upcoming paragraphs, we will describe different studies that led to the following proposal: A strong conflict between what is expected in the current context and what is perceived triggers reanalysis of the input to check for possible processing errors. This monitoring process is reflected in a P600 effect. In order to explain how this hypothesis was developed, a brief summary of the recent event-related potential (ERP) literature is in order.

In the language research field, there used to be a clear distinction between two ERP components. On the one hand, there was the N400 component, reflecting semantic processes. The N400 is a negative-going component

[1]Radboud University Nijmegen, Donders Institute for Brain, Cognition, and Behavior, Netherlands, [2]UMCN St. Radboud, Netherlands

that peaks around 400 msec after critical stimulus onset. The scalp distribution of the N400 is widespread but usually larger over central and parietal electrode sites with a right-hemisphere preponderance (Kutas & Van Petten, 1994). It was first discovered by Kutas and Hillyard (1980c), who found that the N400 was more negative in response to words that were semantically incongruous. For example, in the sentence "He spread the warm bread with socks," the word *socks* elicited a larger N400 amplitude than the word *butter* in the same sentence. This difference in amplitude between congruous and incongruous words regarding their previous context has been referred to as the N400 effect. Generally, the N400 component is assumed to reflect semantic processing, more specifically, its amplitude reflects how easily a word can be integrated in the current context (e.g., Van Berkum, Hagoort, & Brown, 1999; Chwilla, Hagoort, & Brown, 1998; Holcomb, 1993). On the other hand, there was the P600 component, reflecting syntactic processes. This is a positive component starting at about 500 msec and generally lasting till at least 800 msec after critical stimulus onset. It usually has a central–posterior scalp distribution and was first discovered by Osterhout and Holcomb (1992), who found a P600 after a syntactic anomaly. The P600 has been related to syntactic reanalysis or repair processes. Different kinds of syntactic violations, such as violations of case inflection (Münte, Heinze, Matzke, Wieringa, & Johannes, 1998), verb–noun number agreement violations (Hagoort, Brown, & Groothusen, 1993), phrase structure violations (Hagoort et al., 1993), and verb inflection violations (Friederici, Pfeifer, & Hahne, 1993) induce an increase in P600 amplitude. Syntactic complexity also has an influence on the P600 component. Kaan, Harris, Gibson, and Holcomb (2000) found a P600 effect to unambiguous, syntactically correct sentences that had a relatively complex structure as compared to control sentences. Furthermore, garden-path sentences, which are sentences that are locally ambiguous and have a preferred parse (e.g., "The broker persuaded to sell the stock …"), show an increase in P600 amplitude as well (Osterhout & Holcomb, 1992). This difference in amplitude between ungrammatical or more complex/garden-path sentences and grammatical or unambiguous sentences has been referred to as the P600 effect.

This clear distinction between semantic processes reflected by the N400 component and syntactic processes reflected by the P600 component, however, was challenged when different studies found P600 effects without N400 effects to semantic anomalies. Kim and Osterhout (2005) found a P600 effect to semantic verb-argument violations in sentences such as "The hearty meal was devouring" (see also Hoeks, Stowe, & Doedens, 2004). In these sentences, syntactic cues signal that *meal* is the agent of *devouring*, but the meaning of the individual words signals that *meal* is the theme of *devouring*. According to the authors, the relation be-

tween the individual words is so strong that the role assignment signaled by semantic cues is pursued and overrules the role assignment signaled by syntactic cues. Due to this so-called semantic attraction between the individual words, participants perceive the grammatical correct sentence as ungrammatical (*devouring* instead of *devoured*), causing the P600 effect. Kuperberg, Sitnikova, Caplan, and Holcomb (2003) also found a P600 effect to semantic verb-argument violations in which the subject inanimate noun phrase (NP) could have had a more plausible alternative thematic role. An example sentence of such a thematic role animacy violation is the following: "For breakfast the eggs would only eat toast and jam." In this example, the verb *eat* elicited a P600 effect. The authors proposed that the inanimate subject NP (*the eggs*) violated the thematic structure of the verb (*eat*), causing an attempt to reassign the thematic role of the subject NP from agent to theme, thereby eliciting a P600 effect. However, this view of reassignment of thematic roles causing the P600 effect in semantic anomalies is not without problems. Why would reassignment take place in sentences that are syntactically unambiguous? The sentences allow only one interpretation or role assignment and, when trying to reassign thematic roles, this veridical sentence interpretation is lost (e.g., Kolk & Chwilla, 2007; Van Herten, Chwilla, & Kolk, 2006).[1]

Kolk, Chwilla, Van Herten, and Oor (2003) found a P600 effect to semantic reversal anomalies such as: *De kat die voor de muizen vluchtte rende door de kamer* (literal translation: "The cat that from the mice fled$_{[sg]}$ ran through the room"). These sentences did not involve verb-argument violations, as in the Kim and Osterhout (2005) and in the Kuperberg et al. (2003) studies because both cats and mice can flee. However, what is violated is general world knowledge; mice flee from cats and not the other way around. Van Herten, Kolk, and Chwilla (2005) tested whether the P600 effect for these semantic reversal anomalies was caused by a mismatch between the observed and predicted number inflection on the verb. Based on general world knowledge, the sentence could be interpreted as if the mice were fleeing from the cat. In this case, the verb should carry a plural inflection, but the perceived singular inflection of the verb violates this expectation. However, this syntactic prediction hypothesis was ruled out when it turned out that a P600 effect was also present when subject and object NP had the same number (e.g., *De kat die voor de muis vluchtte* …, literal translation: "The cat that from the mouse fled$_{[sg]}$ …"). The authors proposed that the syntactic reprocessing account for the P600 effect should be extended to a more general process of reanalysis, more specifically, a monitoring process (e.g., Vissers, Kolk, Van de Meerendonk, & Chwilla, 2008; Kolk & Chwilla, 2007; Vissers, Chwilla, & Kolk, 2006, 2007; Van Herten et al., 2005, 2006; Kolk et al., 2003). According to this monitoring hypothesis, as in language production, there is monitoring in lan-

guage perception. A conflict arises when the reader is highly expecting a certain linguistic event (e.g., based on world knowledge, the interpretation that the mice flee from the cat) but encounters another unexpected linguistic event (e.g., based on the sentence parse, the interpretation that the cat flees from the mice). This conflict brings the language system into a state of indecision and triggers reanalysis. The function of the reanalysis is to check the input for possible processing errors which gives rise to the P600 effect. The entire process entailing a response conflict (that exceeds a certain threshold; for details, see below), reanalysis and resolution, is referred to as monitoring.

This monitoring process is similar to the one by which we discover perceptual errors such as "Into a posing camp" (produced utterance: "Into opposing camps"), in which the context will render it likely to the listener that she or he misunderstood the speaker and that she or he will ask the speaker to repeat again what was said. In case of a reading error, it is likely that the reader will go back to check whether the reading was correct. A conflict between what is expected in the current context and what is perceived triggers these checking behaviors. The idea that *conflict* is an essential part of the monitoring process is not new; Yeung, Botvinick, and Cohen (2004) proposed this for the error-related negativity (ERN) component. The ERN is a negative deflection in the EEG that is seen when participants make an error in a wide variety of psychological tasks (e.g., Holroyd & Coles, 2002). According to Yeung et al. (2004), the ERN signals that a response conflict has occurred between two incompatible interpretations and this brings the system to correct the error.

The monitoring hypothesis can account for the fact that syntactic violations, garden-path sentences, semantic verb-argument violations, and semantic reversal anomalies trigger a P600 effect. In all instances, a certain linguistic event (e.g., a certain morpheme) is expected, but an unexpected one is encountered, causing a conflict and a monitoring response: "Did I read that correctly?" Moreover, a monitoring process might explain the fact that a P600 effect was found for sentences with a more complex structure (Kaan et al., 2000). These sentences are more difficult and the chance of conflicting representations is higher. Therefore, monitoring for errors may underlie the P600 effect in these sentences (Van Herten et al., 2006).

Since the first proposal of the monitoring theory for language perception (Kolk et al., 2003), different studies have been conducted to test this theory. In a study by Van Herten et al. (2006), it was found that implausible sentences that contained plausible sentence parts (e.g., *Jan zag dat de eekhoorn het brood bakte* ..., literal translation: "John saw that the squirrel the bread baked ...") elicited a P600 and not an N400 effect at the verb (*baked*). In these sentences, a conflict exists between the plausible sentence part (e.g., of baking bread) and the implausibility of the sentence as a whole. To check if the subject NP (the squirrel) was not misread, reanalyzing the input for possible processing errors would be meaningful. Furthermore, Vissers et al. (2006) showed that pseudohomophones in high-cloze sentences elicited a P600 (e.g., *In die bibliotheek lenen de leerlingen boekun* ..., translation: "In that library the pupils borrow bouks ..."), whereas pseudohomophones in low-cloze sentences did not (e.g., *De kussens zijn volgestopt met boekun* ..., translation: "The pillows are stuffed with bouks ..."). In this study, a conflict was present at the word level in the high-cloze sentences, between the highly expected word "books" and the observed pseudohomophone "bouks." Again, reprocessing the input for possible processing errors would be meaningful to be sure the word was not misread. A recent study by Vissers et al. (2008) studied the effect of picture–sentence mismatches. In this study, participants were shown pictures depicting locative relations (e.g., of a circle in front of a square) followed by a sentence that correctly or incorrectly described the picture. It was hypothesized that, in case of a mismatch between picture and sentence, there would be a conflict at the conceptual level between the representation based on the picture and the representation based on the sentence. As predicted, this conflict gave rise to a P600 effect for the mismatching trials.

The abovementioned studies showed that, when there is a conflict between expectancies, error monitoring in language perception can occur at a number of linguistic levels: the sentence, word, and conceptual levels (see Vissers et al., 2008). In the present study, we zoom in on the size of the conflict and test the prediction that only a strong conflict between expectancies will trigger reanalysis and a P600 effect, as it would not be efficient for a monitoring process to be elicited by every conflict (e.g., Vissers et al., 2006). We hypothesized that, if this prediction is correct, a P600 should be triggered by strong semantic violations as well. In daily conversation, we often encounter mildly unexpected units corresponding to new, important information. This information should be integrated, otherwise no learning will take place. However, sometimes we encounter linguistic information that is highly unexpected and impossible to integrate in the current context. To avoid the risk of integrating wrong information, it would be useful to mistrust what was heard or read, and reanalyze the input for possible errors. We hypothesized that the degree of the unexpected event, and therefore, also of the resulting conflict, could be varied by manipulating plausibility. This led to the prediction that only when the implausibility of the sentence, and therefore, the conflict between the expected and unexpected linguistic event, is strong enough will reprocessing take place and a P600 effect should be elicited. In cases where this conflict is not strong enough, such as for mildly implausible sentences, an attempt at integration is made and an N400 effect should occur in the absence of a P600.

In the present study, we tested this prediction of the monitoring theory by presenting participants with sentences in which a category exemplar was highly expected given the previous context. Three experimental conditions were compared: plausible, mildly implausible, and deeply implausible sentences. The plausible sentences (e.g., ''The eye consisting of among other things a pupil, iris, and retina …'') were used as the control sentences. We hypothesized that the mildly implausible sentences (e.g., ''The eye consisting of among other things a pupil, iris, and eyebrow …'') would trigger a mild conflict between the exemplar from the expected category and the unexpected critical noun. In this case, integration difficulties arise due to the unexpectedness but they are resolved successfully, thereby eliciting an N400 effect. For the deeply implausible sentences (e.g., ''The eye consisting of among other things a pupil, iris, and sticker …''), however, we hypothesized that a strong conflict between the exemplar from the expected category and the unexpected critical noun would occur, causing integration failure and triggering reanalysis and thereby eliciting a P600 effect.

METHODS

Participants

Thirty students (26 women; mean age = 21.0 years; age range = 18 to 28 years) participated. All participants were native speakers of Dutch, had no language disability, had no neurological or psychological impairment, had normal or corrected-to-normal vision, and were right-handed. Handedness was assessed with an abridged Dutch version of The Edinburgh Inventory (Oldfield, 1971). Eight participants reported the presence of left-handedness in their immediate family. Participants were paid, or received course credit for their participation.

Materials

One hundred seventy-six sentences were constructed in which a word from a certain category was highly expected given the previous context. The expectancy was created by giving two examples of a certain category (e.g., ''The eye consisting of among other things a pupil, iris, and …''), which made it highly likely that a word from this category would follow and not a word from a different category. In the sentences, the verb always came after the critical noun.

The sentences were used in a pilot study with nine participants (6 women; mean age = 22.3 years; age range = 19 to 25 years), in which participants were asked to give a set of possible one-word completions, completions that were related but did not fit the sentence, and completions that did not fit at all. To be sure participants were not influenced by the part of the sentence after the critical word, the sentence was only shown up to the criti-

cal position. The words that were obtained in this pilot study were then used in a plausibility judgment task with 20 participants (16 women; mean age = 19.9 years; age range: 18 to 27 years): 10 participants got one half of all the sentences and 10 participants got the other half. In the plausibility test, participants had to rate how plausible the words were in the sentences by giving a number ranging from 1 (*very implausible*) to 5 (*very plausible*). Participants rated various words for each sentence, and therefore, saw all three different completions for each sentence that were constructed hereafter.

From the plausibility judgment task, 99 experimental sentences were constructed with each three versions differing in plausibility: (1) plausible; (2) mildly implausible; and (3) deeply implausible (see Table 1). To obtain these three conditions, words with a mean plausibility rating of 5, 3, and 1, respectively, were used and the three conditions of a sentence only differed in their critical word. The critical words could be maximally three syllables long, and for each word the frequency was searched using the CELEX lemma database. The experimental set was constructed in such a way that, overall, across the three conditions, the critical words did not differ in mean length and frequency (both $Fs < 1$). Sentences differed in total length (mean length = 12 words) and the position of the critical word varied across the different sentences (for some examples of the experimental sentences, see Appendix A).

To obtain a more objective measurement of the plausibility of the critical materials, a post hoc latent seman-

Table 1. Example of the Three Sentence Types: (1) Plausible, (2) Mildly Implausible, and (3) Deeply Implausible

Condition	Sentence
(1) Plausible	Het oog bestaande uit onder andere een pupil, iris en *netvlies* is erg gevoelig.
	(The eye consisting of among other things a pupil, iris, and *retina* is very sensitive.)
(2) Mildly implausible	Het oog bestaande uit onder andere een pupil, iris en *wenkbrauw* is erg gevoelig.
	(The eye consisting of among other things a pupil, iris, and *eyebrow* is very sensitive.)
(3) Deeply implausible	Het oog bestaande uit onder andere een pupil, iris en *sticker* is erg gevoelig.
	(The eye consisting of among other things a pupil, iris, and *sticker* is very sensitive.)

The translation is given in parentheses and the critical word is given in *italics*.

tic analysis (LSA) of the stimuli was carried out.[2] To this aim, the two category exemplars for each sentence (e.g., *pupil* and *iris*, referred to as A and B) and the critical nouns (e.g., *retina*, *eyebrow*, and *sticker*, referred to as C, D, and E) were translated for all stimuli. Six trials were excluded because no translation could be found for a portion of the nouns. Pairwise comparisons were carried out using the ''General reading up to 1st year college'' space. For the plausible sentences, pairwise comparisons were conducted between AB, AC, and BC. Likewise, for the mildly and deeply implausible sentences, pairwise comparisons were conducted between AB, AD, BD, and AB, AE, BE, respectively. An estimate of the semantic similarity value (SSV) between the three words was obtained by computing the mean of the pairwise comparisons per sentence type (e.g., AB + AC + BC/3 for the plausible sentences; see Chwilla & Kolk, 2005 for a similar approach). The mean SSV for the 93 trials was 0.36 for the plausible, 0.28 for the mildly implausible, and 0.16 for the deeply implausible sentences. An ANOVA indicated that these differences in mean SSV were reliable [$F(2, 278) = 40.001$, $p < .001$]. Follow-up LSD pairwise comparisons between the three plausibility conditions revealed that the plausible sentences differed significantly from the mildly implausible and from the deeply implausible sentences (both $ps < .001$). Most important for the present purposes, these analyses revealed a significant difference in SSV between the mildly and deeply implausible sentences ($p < .001$).

Ninety-nine filler sentences were created in such a way that, overall, there was an equal number of sentences of the same length and an equal number of correct and incorrect sentences. Sixty-six of the filler sentences did not contain any violation. The remaining 33 filler sentences were low-cloze sentences adapted from Vissers et al. (2006).

Three experimental lists were created on the basis of these materials, which were presented to an equal number of participants. The three versions of each sentence were counterbalanced across lists, in such a way that each participant saw only one version of a sentence. Therefore, each list contained 33 plausible, 33 mildly implausible, and 33 deeply implausible sentences. To each list the 99 fillers were added. Each list consisted of four blocks with pauses in between. Within each block the trials were pseudorandomized using the following constraints: each block began with two filler trials, a filler or experimental trial never occurred more than three times in a row, each sentence type condition never occurred more than three times in a row, and a violation (yes/no) never occurred more than three times in a row.

Procedure

Participants were tested individually, seated in front of a computer screen in a dimly lit Faraday cage. The sentences were presented at the center of the computer screen in serial visual presentation mode. The words were presented in black capitals on a white background in a 10 cm by 2 cm window and viewing distance was approximately 1 m.

Trials began with a fixation cross (duration = 510 msec) followed by a 500-msec blank screen. Then the sentence was presented; word duration was 345 msec and the stimulus onset asynchrony was 645 msec. Sentence-final words were indicated with a full stop and intertrial intervals lasted 2000 msec. Participants were instructed to attentively read the sentence. Furthermore, they were instructed to make eye movements, for instance, eye blinks, in between sentences.

Each block lasted about 10 min and participants were given short breaks in between. During these short breaks, participants were given a recognition task to ensure they were attentively reading the sentences. The task consisted of a couple of sentences (5 per block) for which participants had to indicate whether they had been presented in the previous block.

EEG Recording

With 27 tin electrodes mounted in an elastic electrode cap, the continuous EEG was recorded (Electro-Cap International, Eaton, OH). The electrode positions included 12 electrodes placed at locations from the standard International 10–20 system, namely, at the frontal (F3, F4, F7, and F8), temporal (T5 and T6), parietal (P3 and P4), and midline (Fz, Cz, Pz, and Oz) locations. Furthermore, seven electrodes were placed at anterior frontal (F3A, F4A, F7A, and F8A), parietal (P3P and P4P), and midline (FzA) locations. Another eight electrodes were placed at locations that have been reported to be sensitive to language manipulations (e.g., Holcomb & Neville, 1990) which included left and right anterior temporal sites (LAT and RAT: 50% of the distance between T3/4 and F7/8), left and right temporal sites (LT and RT: 33% of the interaural distance lateral to Cz), left and right temporo-parietal sites (LTP and RTP: corresponding to Wernicke's area and its right hemisphere homologue, 30% of the interaural distance lateral to a point 13% of the nasion–inion distance posterior to Cz), and left and right occipital sites (OL and OR: 50% of the distance between T5/6 and O1/2). This electrode montage has been used in earlier studies (e.g., Van Herten et al., 2006; Vissers et al., 2006; Figure 1 shows the position of the electrodes).

Both the left and right mastoids were recorded and the right mastoid served as reference. The signal was re-referenced to the average of the left and right mastoids before the analysis. Eye blinks and eye movements were recorded by horizontal EOG electrodes next to both eyes and vertical EOG electrodes placed below and above the right eye. The ground was placed on the forehead, in between both eyes. For the EOG electrodes, the impedance was smaller than 5 kΩ, and for all the other electrodes, impedance was smaller than 3 kΩ. The

Figure 1. Electrode montage used in the present experiment.
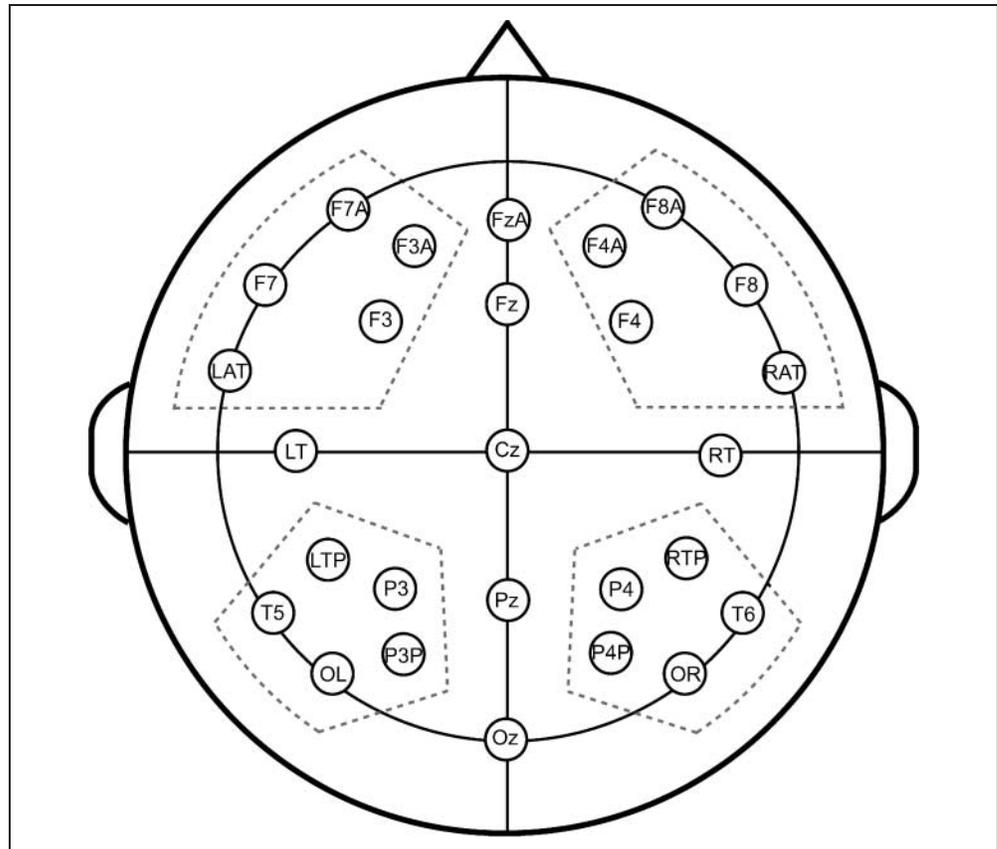
EEG and EOG signals were amplified (time constant = 8 sec, band pass = 0.02–30 Hz) and digitized on-line with a sampling frequency of 200 Hz.

EEG Data Analysis

EEG and EOG records were examined for artifacts and excessive EOG amplitude (>100 μV) extending from 100 msec before the onset of the critical noun up to 1000 msec following its onset, and contaminated trials were removed. Averages were aligned to a 100-msec baseline period preceding the critical noun.

The ERPs were analyzed in the following way. Mean amplitudes were calculated in an early window (i.e., 300–500 msec) and a late window (i.e., 500–800 msec), capturing N400 and P600 effects, respectively. These windows were based upon visual analysis and corresponded to the time intervals in which maximal differences between conditions were obtained. Repeated measures analyses of variance (MANOVAs) were conducted with plausibility (plausible, mildly implausible, deeply implausible) as factor. The multivariate approach to repeated measurements was used to avoid problems concerning sphericity (e.g., Vasey & Thayer, 1987). To examine laterality effects, ERPs at the midline and lateral sites were analyzed in separate MANOVAs. The midline analysis included the additional factor site (FzA, Fz, Pz, Cz, Oz).

For the lateral analysis, we used a Region of Interest (ROI: anterior vs. posterior) by Hemisphere (left vs. right) by Lateral site (F7A, F3A, F7, F3, LAT vs. LTP, P3, P3P, T5, OL vs. F8A, F4A, F8, F4, RAT vs. RTP, P4, P4P, T6, OR) design to explore the ERP effects' distribution across the scalp. Interactions with the factor site were followed up by paired t tests at the single-site level.

Plausibility (plausible, mildly implausible, deeply implausible) was a within-subject variable in the initial MANOVAs [next to the (lateral) site, ROI and hemisphere within-subject variables described above]. Significant main effects and interactions of these MANOVAs were followed up by planned simple effect MANOVAs to make comparisons between all pairs of plausibility conditions.

RESULTS

Performance on the Recognition Task

Mean error rate for the sentences for which the participants had to indicate whether they had been presented in the previous block or not was 7.67% (Block 1: 6.67%; Block 2: 6.00%; Block 3: 9.33%; Block 4: 8.67%). Splitting up the error rates per condition resulted in 9.1% for the deeply implausible sentences (i.e., on 9.1% of the deeply implausible sentences, an error was made), for the mildly implausible sentences this was 10%, and for

the plausible sentences, 7.5%. These error percentages indicate that the participants attentively read the sentences during the experiment.

Event-related Potentials

Grand-average waveforms for all the sentence types, time-locked to the onset of the critical noun, are presented in Figure 2. An early ERP response characteristic for visual stimuli was elicited for all sentence types; an N1 followed by a P2 component, which at occipital sites was preceded by a P1 component. Visual inspection of the waveforms suggests different patterns of brain activity for the mildly and deeply implausible sentences as compared to the plausible sentences (see Figures 3 and 4, respectively, for a direct comparison). The waveforms of the mildly implausible sentences suggest that an N400 effect was present; mean amplitude was more negative-going for the mildly implausible than the plausible nouns in the 300 to 500 msec epoch. The waveforms of the deeply implausible sentences, however, suggest a biphasic pattern; an N400 effect as well as a central–posterior P600 effect (mean amplitude was more positive going

for the deeply implausible than the plausible nouns in the 500 to 800 msec window) was present. Visual inspection of Figure 2 suggests that the N400 effect is a bit larger for the deeply implausible than the mildly implausible sentences.

Statistical Analyses

The mean percentage of trials that had to be rejected because of artifacts and excessive EOG amplitude was 3.84% for the deeply implausible, 5.86% for the mildly implausible, and 3.64% for the plausible condition. Report of the ERP results will be restricted to the main effects and interactions that are relevant for the functional interpretation of the condition effects in the present study.

N400 Window (300–500 msec)

The omnibus analysis showed main effects of plausibility for the midline [$F(2, 28) = 9.42$, $p < .001$] and lateral sites [$F(2, 28) = 9.47$, $p < .001$]. Furthermore, a Plausibility × Site interaction was present for the midline [$F(8, 22) = 3.40$, $p < .05$] and lateral sites [$F(2, 28) =$

Figure 2. Grand-average ERP waveforms to the critical noun for all midline and a subset of lateral sites, for all sentence types: plausible (solid line), mildly implausible (dashed line), and deeply implausible (dotted line).
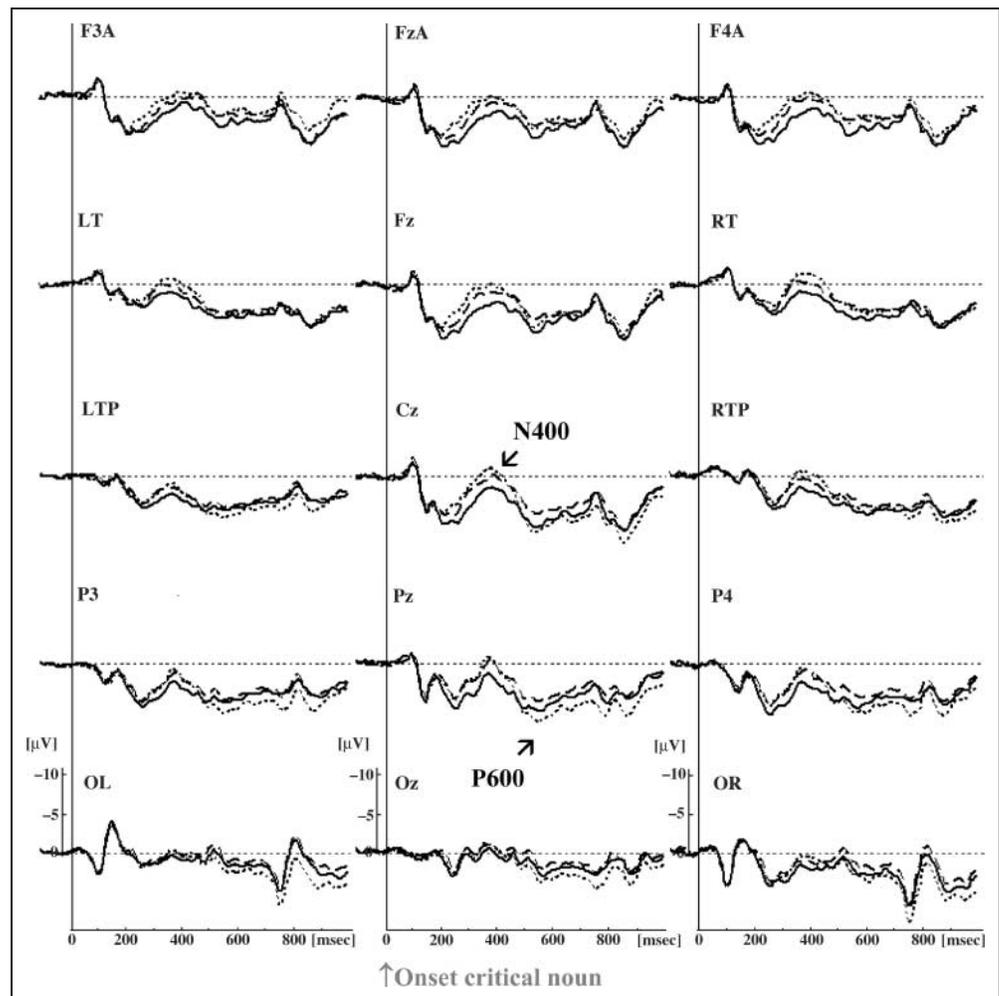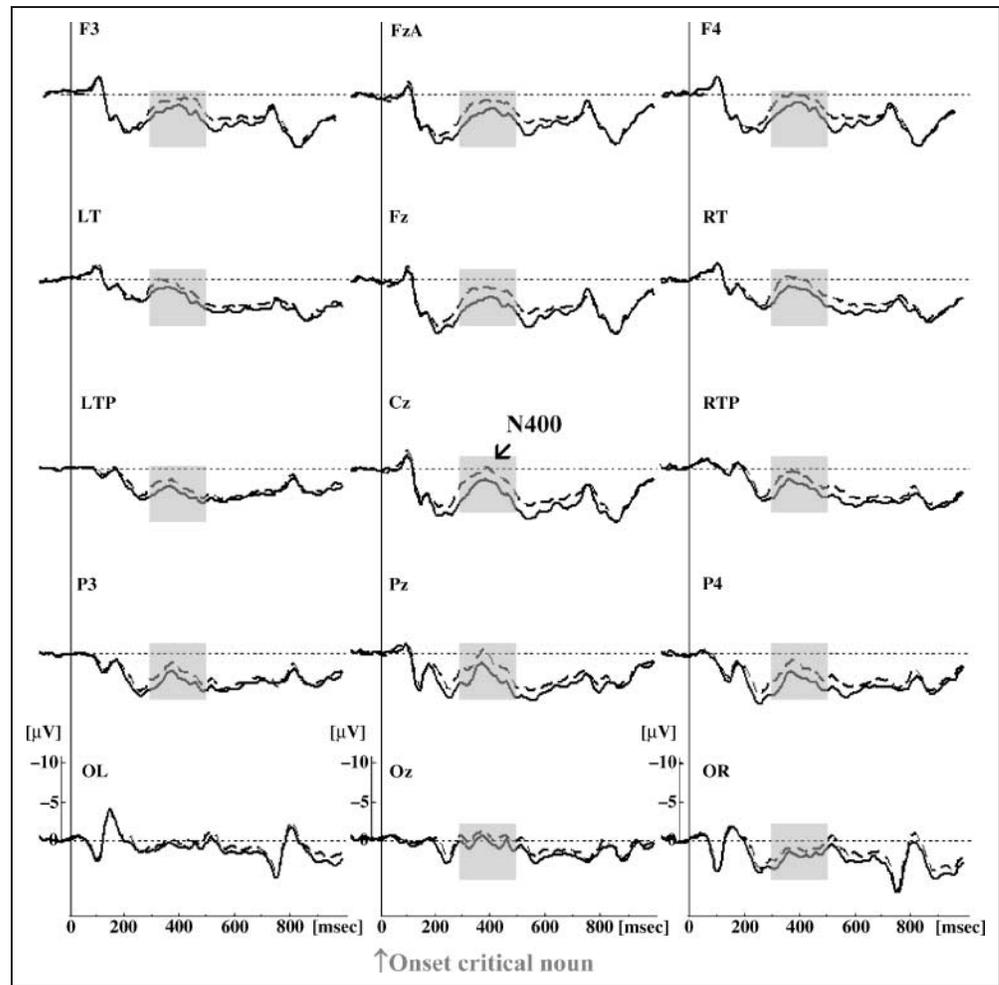
Figure 3. Grand-average ERP waveforms to the critical noun for all midline and a subset of lateral sites, for the mildly implausible (dashed line) versus plausible sentences (solid line).

3.12, $p < .05$]. The analysis for the lateral sites yielded Plausibility $\times$ Hemisphere [$F(2, 28) = 3.77, p < .02$], Plausibility $\times$ ROI [$F(2, 28) = 5.49, p < .05$], Plausibility $\times$ Hemisphere $\times$ Site [$F(8, 22) = 2.46, p < .05$], and Plausibility $\times$ ROI $\times$ Site [$F(8, 22) = 3.15, p < .02$] interactions.

Mildly implausible vs. plausible sentences. Follow-up analyses for the midline sites, comparing the mildly implausible with the plausible sentences, revealed an effect of plausibility [$F(1, 29) = 18.85, p < .001$] in the absence of an interaction of this factor with site, indicating that an N400 effect was obtained across the midline. For the lateral sites, an effect of plausibility [$F(1, 29) = 17.34, p < .001$], a Plausibility $\times$ Site [$F(4, 26) = 3.63, p < .05$], and a Plausibility $\times$ Hemisphere $\times$ Site interaction [$F(4, 26) = 4.35, p < .01$] were present. Single-site analyses indicated that an N400 effect was obtained at anterior frontal (F3A, F4A, F7A, F8A), frontal (F3, F4, F7, F8), anterior temporal (RAT), temporal (LT, RT, T6), temporo-parietal (LTP, RTP), parietal (P3, P4, P3P, P4P), and occipital sites (OR) (see Figure 5 for the topographical map).

Deeply implausible vs. plausible sentences. Follow-up analyses for the midline sites, comparing the deeply implausible with the plausible sentences, revealed an effect of plausibility [$F(1, 29) = 8.56, p < .01$], and a Plausibility $\times$ Site interaction [$F(4, 26) = 4.44, p < .01$]. Single-site analyses indicated that an N400 effect was obtained at the following midline sites: FzA, Fz, Cz, and Pz. For the lateral sites, an effect of plausibility [$F(1, 29) = 13.48, p < .001$], a Plausibility $\times$ Site [$F(4, 26) = 3.70, p < .05$], a Plausibility $\times$ Hemisphere [$F(1, 29) = 7.07, p < .05$], a Plausibility $\times$ ROI [$F(1, 29) = 9.87, p < .01$], and a Plausibility $\times$ ROI $\times$ Site [$F(4, 26) = 6.64, p < .001$] interaction were present. The interactions for the lateral sites reflected that an N400 effect was present at the following bilateral sites: F3A, F4A, F7A, F8A, F3, F4, F7, F8, LAT, RAT, LT, RT, LTP, RTP, as well as at T6. Furthermore, only for the right hemisphere was the N400 effect extended to parietal (P4) and occipital (OR) sites (see Figure 5 for the topographical map).

Deeply implausible vs. mildly implausible sentences. Follow-up analyses comparing the deeply with the mildly implausible sentences revealed no effect of plausibility

Figure 4. Grand-average ERP waveforms to the critical noun for all midline and a subset of lateral sites, for the deeply implausible (dotted line) versus plausible sentences (solid line).
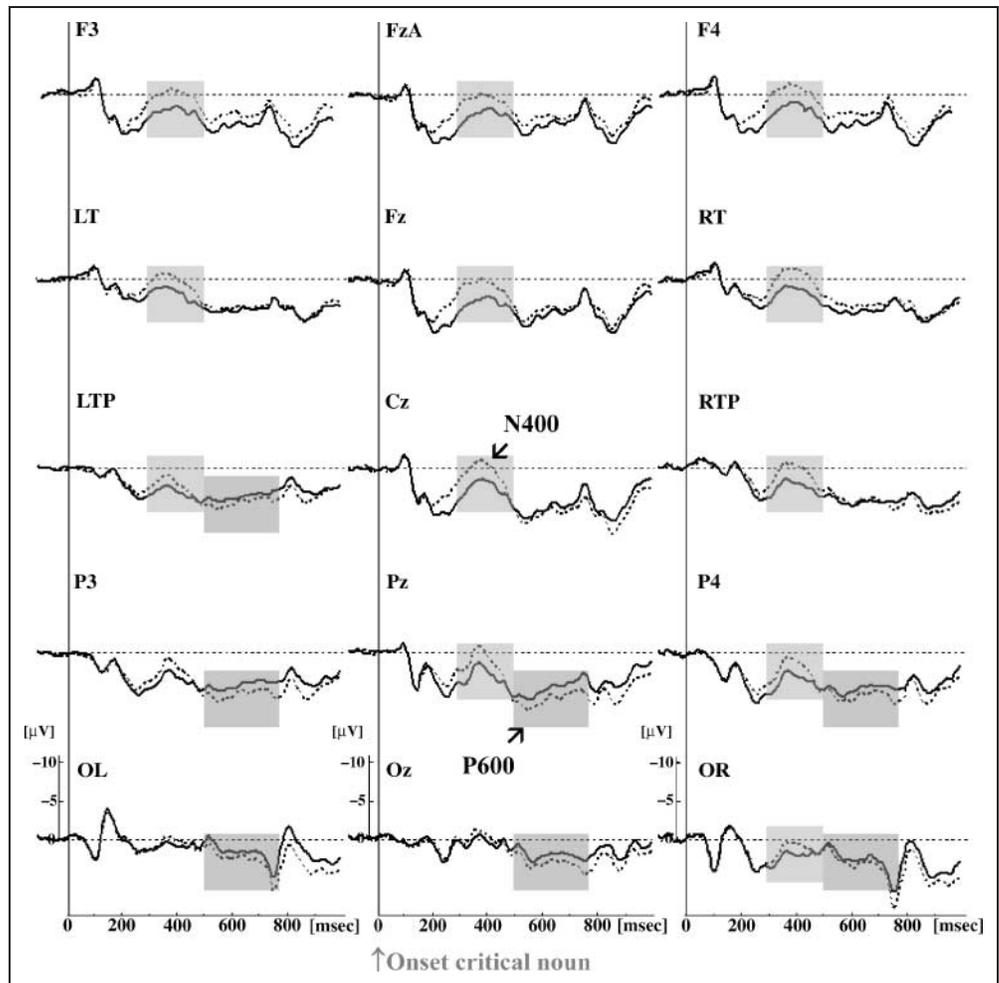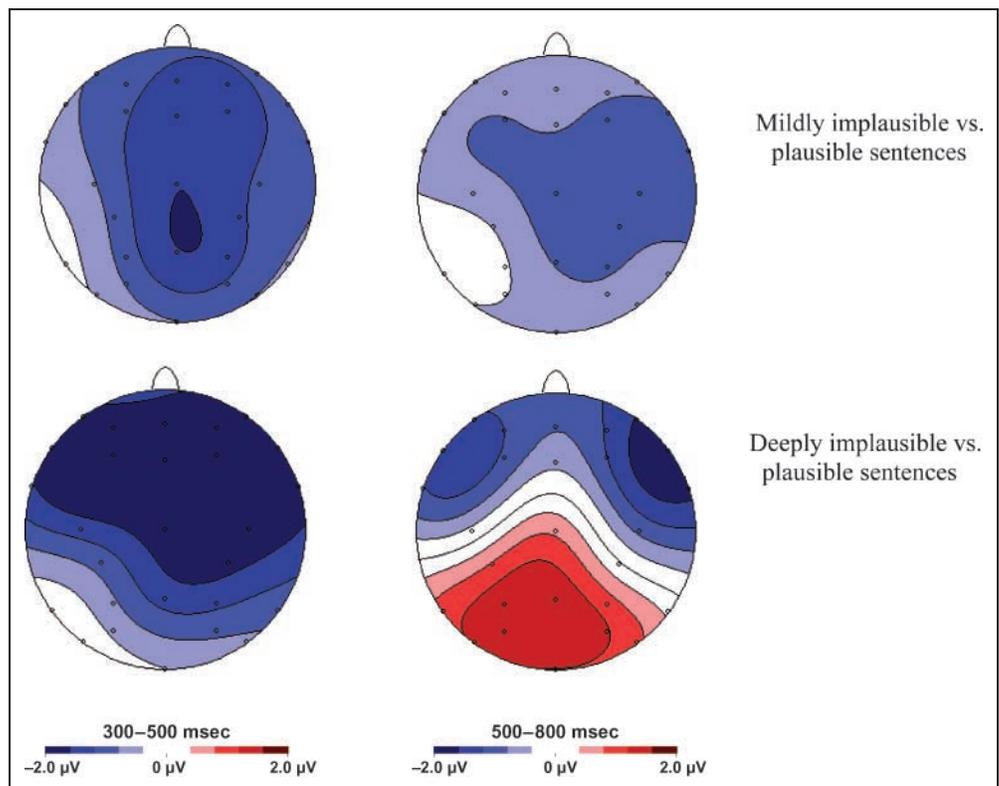

Figure 5. Topographical maps obtained by interpolation from 27 sites for the N400 window (300–500 msec) and the P600 window (500–800 msec). Maps were computed from the difference waves of the mildly implausible versus plausible (first row) and deeply implausible versus plausible (second row) sentences.

[midline: $F(1, 29) = 0.01, p > .05$; lateral: $F(1, 29) = 1.96, p > .05$]. However, for the lateral sites, there was a Plausibility × ROI [$F(1, 29) = 7.01, p < .05$] and a Plausibility × Hemisphere × Site interaction [$F(4, 26) = 3.27, p < .05$]. Separate analyses for the two levels of ROI (anterior, posterior) revealed a larger N400 amplitude to the deeply implausible sentences for the anterior [$F(1, 29) = 6.71, p < .05$], but not for the posterior ROI [$F(1, 29) = 0.05, p > .05$]. Single-site analyses indicated that an N400 effect was obtained for the following subset of sites: F7A, F7, F8, LAT, RAT, and RT.

P600 Window (500–800 msec)

The omnibus analysis showed main effects of plausibility for the midline sites [$F(2, 28) = 6.33, p < .01$] and the lateral sites [$F(2, 28) = 3.89, p < .05$]. Furthermore, a Plausibility × Site interaction was present for the midline [$F(8, 22) = 5.20, p < .001$] and lateral sites [$F(8, 22) = 2.68, p < .05$]. In addition, the analysis for the lateral sites yielded a Plausibility × ROI interaction [$F(2, 28) = 26.35, p < .001$].

Mildly implausible vs. plausible sentences. Follow-up analysis comparing the mildly implausible with the plausible sentences revealed an effect of plausibility at the midline [$F(1, 29) = 6.25, p < .05$] and lateral sites [$F(1, 29) = 7.60, p < .05$]. This indicated that, across the midline and lateral sites, an effect in the opposite direction was obtained; mean amplitude in the P600 window was more negative-going for the mildly implausible than the plausible sentences, reflecting a continuation of the N400 effect in the later window (for discussion, see below; see Figure 5 for the topographical map).

Deeply implausible vs. plausible sentences. Follow-up analysis comparing the deeply implausible with the plausible sentences revealed no effect of plausibility for the midline sites [$F(1, 29) = 0.97, p > .05$]. However, there was a Plausibility × Site interaction [$F(4, 26) = 4.28, p < .01$], which indicated that a P600 effect was obtained at Pz and Oz. For the lateral sites, no effect of plausibility [$F(1, 29) = 0.43, p > .05$] was present as well. However, a Plausibility × Site [$F(4, 26) = 3.84, p < .05$] and a Plausibility × ROI interaction [$F(1, 29) = 47.67, p < .001$] were present. Separate analyses for the two levels of ROI revealed an effect for the anterior [$F(1, 29) = 9.73, p < .01$] and posterior ROI [$F(1, 29) = 7.92, p < .01$]. For the posterior ROI, a P600 effect was obtained at the following sites: T5, LTP, P3, P4, P3P, P4P, OL, and OR. In contrast, for the anterior ROI, an effect in the opposite direction was found at a subset of sites: F3A, F4A, F7A, F8A, F3, F4, F7, F8, LAT, and RAT, reflecting an extension of the N400 effect into the P600 window (see Figure 5 for the topographical map).

Deeply implausible vs. mildly implausible sentences. Follow-up analysis comparing the deeply with the mildly implausible sentences revealed an effect of plausibility [$F(1, 29) = 9.81, p < .01$], and a Plausibility × Site interaction [$F(4, 26) = 9.29, p < .001$] at the midline sites. Single-site analyses indicated a P600 effect was obtained for the following subset of sites: Cz, Pz, and Oz. For the lateral sites, a Plausibility × Site interaction [$F(4, 26) = 5.81, p < .01$], and a Plausibility × ROI interaction [$F(1, 29) = 32.40, p < .001$] were found. Separate analyses for the two levels of ROI revealed an effect of plausibility for the posterior [$F(1, 29) = 16.38, p < .001$], but not for the anterior ROI [$F(1, 29) = 2.98, p > .05$]. Single-site analyses revealed a P600 effect at the following lateral sites: T5, T6, LTP, P3, P4, P3P, P4P, OL, and OR. An effect in the opposite direction was found for two anterior electrodes: LAT and F8.

DISCUSSION

The main results of the present study were as follows. First, as predicted, the mildly implausible sentences elicited an N400 effect at the critical noun when compared to the plausible sentences. Second, an N400 effect was also observed for the deeply implausible sentences, when the critical noun was compared to that of the plausible sentences. This N400 effect for the deeply and mildly implausible sentences was broadly distributed across the scalp. Third, and most importantly, in accordance with the prediction of the monitoring theory, only the deeply implausible sentences elicited a P600 effect when compared to the plausible sentences. The LSA confirmed that the deeply implausible sentences were, indeed, less semantically plausible than the mildly implausible sentences, as reflected by significant differences in SSVs. The P600 effect to deeply implausible sentences resembled the P600 effect found to semantic anomalies and syntactic violations and ambiguity (e.g., Van Herten et al., 2005; Friederici et al., 1993), in terms of the timing and the central–posterior scalp distribution of the effect (see Figure 5).

As was mentioned in the Introduction it is generally believed that the N400 reflects semantic processing, more specifically integration difficulties (e.g., Van Berkum et al., 1999; Chwilla et al., 1998; Holcomb, 1993). In the present study, we found an N400 effect for both the mildly and deeply implausible sentences, indicating that, in both conditions, integration difficulties arose. However, for the mildly implausible sentences, the absence of a P600 effect reflected that these integration difficulties were resolved. In contrast, for the deeply implausible sentences, integration failed, which we propose triggered a process of reanalysis as reflected by the presence of a P600 effect.[3] Recent studies conducted on thematic role animacy violations (e.g., Kuperberg et al., 2003), semantic verb–argument violations (e.g., Kim & Osterhout, 2005), and semantic reversal anomalies (e.g., Van Herten et al., 2005,

2006; Kolk et al., 2003) did not report a biphasic N400–P600 pattern, but only a P600 effect to the semantic anomalies. As argued by Van Herten et al. (2006), in most studies where a ''semantic P600'' has been observed, an N400 reflecting integration difficulties did not occur as well because a plausibility heuristic, taking into account content words alone, led to a plausible interpretation for both the plausible and the implausible sentences. In our implausible sentences, however (e.g., ''The eye consisting of among other things a pupil, iris, and eyebrow/sticker . . .''), both the content words on their own and the regular parse would deliver an implausible interpretation for both the mildly and the deeply implausible sentences. Thus, both should indeed show an N400 effect.

On the other hand, the P600 was thought to reflect syntactic processes, but this view has been challenged because P600 effects have been found to semantic verb-argument violations (e.g., Kuperberg, Caplan, Sitnikova, Eddy, & Holcomb, 2006; Kim & Osterhout, 2005; Kuperberg et al., 2003) and semantic reversal anomalies (e.g., Van Herten et al., 2005; Kolk et al., 2003), as well as pseudohomophones in high-cloze sentences (Vissers et al., 2006), and picture–sentence mismatches (Vissers et al., 2008).

One view is that the P600 reflects syntactic (re)processing consequent upon grammaticality violations, sentence ambiguity, or a high degree of complexity (e.g., Kaan et al., 2000; Münte et al., 1998; Friederici et al., 1993; Hagoort et al., 1993; Osterhout & Holcomb, 1992). However, the sentences in our deeply implausible condition did not contain any syntactic violation, were unambiguous, and had the same structure as the other two experimental conditions, meaning these factors cannot have triggered the P600 effect in the deeply implausible sentences. Kim and Osterhout (2005) and Kuperberg et al. (2003) proposed that semantic verb-argument violations elicited P600 effects because they triggered processes of thematic role reassignment. Thematic role reassignment, however, cannot account for the P600 effect found in the present study because the verb was presented after the critical noun and no roles therefore could have been assigned yet. A second factor proposed by Kuperberg (2007), in her review on P600 effects elicited by semantic anomalies, is animacy. Many verbs used in the studies with semantic verb-argument violations had an inherent thematic structure (agent/experiencer theme), which was violated by an inanimate agent NP. In the present study, about 19% of the sentences of the deeply implausible sentences contained an animacy violation with respect to the mentioned category (e.g., ''Animals at that farm like chickens, pigs, and pits have . . .''). To determine whether animacy violations could have played a critical role in eliciting a P600 effect to deeply implausible sentences, supplementary analyses were conducted. In these analyses, we excluded those items in which an animacy violation occurred. To keep the number of sentences constant across conditions, all three

versions of a trial were removed (i.e., plausible, mildly implausible, and deeply implausible). With these supplementary analyses, essentially the same results as with the original analyses were obtained. In particular, when comparing the mildly and deeply implausible sentences with the plausible sentences, a significant central–posterior distributed P600 effect was present for the deeply implausible sentences but not for the mildly implausible sentences. Based on these results, we can reject the hypothesis that the P600 effect to the deeply implausible sentences was due to animacy violations.

How then can we account for the P600 effect elicited in the present study? The monitoring theory provides a possible answer. According to the monitoring hypothesis, the P600 reflects a more general process of reanalysis to check for possible processing errors (e.g., Vissers et al., 2006, 2007, 2008; Kolk & Chwilla, 2007; Van Herten et al., 2005, 2006; Kolk et al., 2003). The reanalysis is triggered by a strong conflict between an expected and unexpected linguistic event; in our deeply implausible sentences, these events are the exemplar from the expected category and the noun that is actually presented. The expectancy in the present study was created by giving two examples of a certain category, making it very likely a word from this category would follow and not a word from another category. When the conflict is strong enough, as in our deeply implausible sentences, reanalysis is triggered, eliciting a P600 effect. All the abovementioned studies have in common that a certain linguistic event is highly expected but another unexpected linguistic event is encountered. A conflict arises between the two expectancies and reanalysis is triggered to check the input; ''Did I read that correctly?'' This monitoring process is proposed to be reflected by the P600 effect. It is important to point out that according to the monitoring theory, *all* aspects of the input are reanalyzed—that is, the semantic, syntactic, orthographic, as well as phonological aspects of the stimulus all are taken into account (Vissers, 2008). Furthermore, it is assumed that depending on the type of error, the reanalysis process can focus on a certain aspect of the stimulus [e.g., phonological/orthographic aspects for misspellings (Vissers et al., 2006), and semantic aspects for the violations in the present study]. The hypothesized result of the reanalysis process consists of the realization that the perceived error was indeed present and did not stem from a processing error as such.

A question that comes to mind, however, is why low-cloze sentences such as ''He spread the warm bread with socks'' (Kutas & Hillyard, 1980c) do not elicit a P600? In these sentences, a certain linguistic event is highly expected as well and another unexpected linguistic event is encountered, which therefore should create a strong conflict, give rise to reanalysis, and thus, elicit a P600 effect. However, for these semantic anomalies, an N400 effect, and not a P600 effect, has generally been reported in the literature. With the present study, we are

not able to answer this question conclusively because another sentence paradigm was used, in which not cloze probability, but plausibility, was varied. However, we do have some suggestions as to why we did find a P600 effect, whereas other studies with strong semantic violations did not.

One factor that could influence why previous studies with very implausible sentences did not report a P600 effect, whereas the present study did, is the type of sentences that we used, which presumably constrained the range of possible interpretations by creating a high expectancy for an exemplar from a particular category. In a recent study, Federmeier, Wlotko, De Ochoa-Dewald, and Kutas (2007) examined the effects of expectancy (cloze probability) and sentence constraint on the ERP response to words. For strongly constraining sentences (e.g., ''Sam could not believe her story was…''), the best completion (the expected word, e.g., *true*) had a mean cloze value of 83.5%, and the second best completion had a mean cloze value of 4.9%. For the weakly constraining sentences (e.g., ''I was impressed by how much he…''), however, mean cloze value was 26.9% for the expected word (e.g., *knew*), whereas the second best completion had a mean cloze value of 9.3%. The unexpected, low-cloze words (e.g., *published*, in the previous two sentence examples) were matched for cloze probability. By varying constraint, a strong competitor was present when the unexpected word was perceived in the strongly constraining sentences, whereas the same unexpected word perceived in the weakly constraining sentences had a couple of relatively weak competitors. In the experiment, a positivity between 500 and 900 msec was found, following the N400 effect, to unexpected words in strongly constraining sentences. This positivity was not observed when the same unexpected words were used in weakly constraining sentences, or when expected words ended the sentences (Federmeier et al., 2007). Relating this to the present study, in the strongly constraining sentences, a certain word was highly expected based on cloze probability, thereby creating a strong conflict when the unexpected word was perceived and eliciting a positivity. In contrast, the weakly constraining sentences did not create a high expectation for a certain word, and therefore, no strong conflict was triggered upon perceiving the unexpected word. In the present study, the same sentence context was used for all conditions, meaning that a strongly constraining context was present for both mildly and deeply implausible sentences. However, due to our plausibility manipulation and as confirmed by the LSA, the context fit for the critical words in the deeply implausible sentences was smaller than that in the mildly implausible sentences, despite the fact that both critical words had a cloze value of zero. Therefore, in the deeply implausible sentences, a stronger conflict was present between the exemplar from the expected category and the perceived noun, hence, a P600 effect was elicited. A difference with the present study that must

be noted, however, is that the positivity Federmeier et al. (2007) found had a frontal distribution, whereas in the present study, a central–posterior distribution was observed. This difference in topography might be related to the fact that, in the study by Federmeier et al., all the unexpected endings were plausible, whereas in the present study this was not the case.

Another factor that could explain the discrepancy in results between studies (the occurrence of either an N400 or P600 to semantic anomalies) is component overlap. In the present study, it was found that, across the scalp, the N400 effect prolongated into the P600 window when comparing the mildly implausible with the plausible sentences. This was also the case, for the anterior sites, when comparing the deeply implausible with the plausible sentences. When inspecting Figure 3, it can be seen that the negative shift of the mildly implausible sentences in the N400 window continues into the P600 window; after 500 msec, the waveform does not align with that of the plausible sentences. Therefore, the N400 effect could be counteracting the positive shift of the P600 component. When looking at Figure 2, this seems to be a valid hypothesis. In this figure, the mildly and deeply implausible sentences can be compared. As the results have shown, both elicit an almost comparable N400, which could have cancelled out the influence of this component on the P600 window when comparing the sentences to each other and shows the P600 effect for the deeply implausible sentences more clearly than when these are compared to the plausible sentences. An example of a study that reported possible component overlap is the study by Schwarz, Kutas, Butters, Paulsen, and Salmon (1996). They found, subsequent to an N400 effect elicited by semantically unrelated trials in a category priming task, a positivity over the left hemisphere between 600 and 800 msec for elderly participants, but not for young participants. They hypothesized that because the amplitude of the N400 effect was larger for young as compared to elderly participants, the subsequent positivity for young participants was masked. In other words, because the N400 amplitude was smaller in elderly participants, the obscuring effect of this negativity was minimized and the late positivity could be seen.

Like the study by Schwarz et al. (1996) mentioned above, Núñez-Peña and Honrubia-Serrano (2005) and Heinze, Muente, and Kutas (1998) found that, in a category verification task, a positivity followed the N400 effect to nonmembers of a semantic category. Furthermore, there have been various other studies that found that the N400, elicited by semantically incongruous sentence completions, was followed by a larger positivity. For example, Ford et al. (1996) and Woodward, Ford, and Hammett (1993) both presented a subset of congruous and incongruous sentences from among those used in various studies by Kutas and Hillyard (1980a, 1980b, 1980c), and it was found that the N400 to incongruous sentence completions was accompanied by a larger late positivity.

Gunter, Jackson, and Mulder (1992) also found a positivity following the N400 effect to incongruous sentence endings, and Swick, Kutas, and Knight (1998) even hinted at a "post-sentence error monitoring process" for the positivity they found between 600 and 900 msec, following the N400 elicited by incongruous sentence endings.

In light of the studies that show a larger positivity following the N400 effect to semantically incongruous sentences, Van Petten and Luka (2006) speak of a "post-N400 positivity," and they note that little research has been done in trying to determine the relevant factor(s) influencing when a monophasic N400 effect or a bi-phasic pattern will occur. In the present study, we were able to differentiate between these two patterns within participants by manipulating plausibility. With this we do not want to imply that plausibility is the only factor of influence, for example, yet unknown properties of the stimulus materials may play a role, as well as individual processing strategies. However, further research is needed to find out why a larger positivity sometimes follows the N400 to semantically incongruous stimuli and sometimes does not.[4]

When looking at the stimuli of the present experiment, the question might arise whether the positivity that we have found is not a P600 but a P3b component due to an oddball effect. In the classical oddball paradigm on each trial, one of two events can occur with a certain probability (e.g., long auditory tones 80%, and short auditory tones 20%), and the rare events elicit a larger P3b (Donchin, 1981). In the present experiment, the conditions had an equal probability overall, but because we used summations within our sentences (e.g., "… pupil, iris, and sticker …"), some might argue that the positivity we found in our deeply implausible sentences is a P3b component due to an oddball effect, elicited by the rare event of the third noun in the summation not matching the category of the other two. To date, it is debated whether the P600 elicited by syntactic violations and the P3b elicited by rare nonlinguistic events belong to the same family of P300 components. Some (e.g., Coulson, King, & Kutas, 1998; Gunter, Stowe, & Mulder, 1997) argue that the P600 and P3b resemble each other, and (at least in part) reflect a domain-general process elicited by rare events. Others (e.g., Osterhout & Hagoort, 1999; Osterhout, McKinnon, Bersick, & Corey, 1996) have found that these components are at least, to a certain degree, distinct, and therefore, they reason that a part of the neural and cognitive processes should differ. The monitoring theory proposes that the P600 and P3b could be related (e.g., Vissers et al., 2008; Van Herten et al., 2005), in the sense that both components can be triggered by an unexpected event, have shown similar scalp distributions, and fall within the same time range. However, what sets them apart, speaking of the cognitive processes, might be understood by the type and complexity of the information that has to be reanalyzed (linguistic vs. non-

linguistic), which could explain latency differences (see Vissers et al., 2008).

As described in the Introduction, Levelt's (1983) theory on language production assumed that errors are detected by a process of comparison of the intended and the planned or produced utterance. Such a comparison could not underlie the monitoring of perception however, because the intentions underlying the perceived utterance are unknown. We therefore hypothesized that it is a strong conflict between what was expected and what was perceived that triggers the language system to reprocess the input. Because the conflict brings the system into a state of indecision, it functions as a strong bottom–up signal, which does not require a monitoring process to be detected. This is different from Levelt's theory, which entails intention and planning/output to be constantly monitored, in order for discrepancies to be detected. Our view does bear similarity to the conflict monitoring theory in the action domain (e.g., Yeung et al., 2004). This theory assumes that anterior cingulate cortex monitors response conflict when multiple response tendencies are activated, and when detecting such a conflict, prefrontal areas are "warned" to increase cognitive control. This theory does not imply a comparator to detect an error; response conflict that exceeds a certain threshold automatically triggers anterior cingulated cortex, which then informs the brain areas responsible for cognitive control processes.

Another view that shows similarities to our monitoring theory is the view of Kuperberg (2007). In her review, Kuperberg proposes a language comprehension system with at least two interacting processing streams: a semantic memory-based stream and a combinatorial processing stream (sensitive to morphosyntactic and lexical–semantic constraints). A conflict between the outcomes of both streams is thought to trigger continued analysis of the combinatorial stream, reflected by the P600 component. Similar as suggested by the monitoring theory of language perception (e.g., Kolk et al., 2003), Kuperberg proposes that the P600 effect is triggered by a *conflict* between representations. Furthermore, both views assume that the P600 reflects some form of continued (re)analysis of the input. However, we think that a major difference between the two views is the proposed nature and function of the (re)analysis. The processing account Kuperberg proposes is focused on semantic and syntactic aspects of verb-argument structure to determine whether a sentence is acceptable or not. In contrast, the monitoring theory proposes a more general function of the reanalysis, in which all aspects of the input are reanalyzed to find out whether a processing error occurred.

To conclude, in support of the monitoring theory, we propose that the P600 reflects a more general process of reanalysis. The present study shows that only when the conflict between the expected and the unexpected linguistic event is strong enough is reanalysis triggered. It would not be efficient for a monitoring process to be

activated by every conflict; when still possible, we try to integrate the information into the context because we assume that what we read is meaningful and something that is relatively unexpected can still be informative. A mild conflict, such as the mildly implausible sentences in the present experiment, therefore does not trigger re-analysis, and thus, no P600 effect occurs.

APPENDIX A

These are some examples of experimental sentences that were used in the present study. The translation is given in parentheses, and the words given in italics are the critical words (*deeply implausible/mildly implausible/plausible*). (For a list of all experimental sentences, see www.socsci. oru.nl/~nanvdm/Appendix%20A-all.pdf.)

Lichaamsdelen zoals een arm, nek en *telescoop/haar/ teen* hebben elk hun eigen functie. (Parts of the body like an arm, neck and *telescope/hair/toe* each have their own function.)

Zuivelproducten zoals yoghurt, kaas en *bak/ei/melk* vind je in de koeling. (Dairy products like yoghurt, cheese and *bin/egg/milk* can be found in the refrigeration.)

Meubels waaronder een bank, bed en *radar/lamp/ kast* vind je in de woonwinkel. (Furniture among which a couch, bed and *radar/lamp/cupboard* can be found in a furniture shop.)

Dieren op die boerderij zoals kippen, varkens en *kuilen/ cavia's/koeien* hebben geen ruimte om te scharrelen. (Animals at that farm like chickens, pigs and *pits/guinea-pigs/ cows* have no room to scrape.)

Kleding zoals truien, broeken en *sluizen/hoeden/ rokken* ligt in de klerenkast. (Clothes like jumpers, trousers and *sluices/hats/skirts* lie in the closet.)

Wapens zoals een zwaard, mes en *vuilniszak/boog/ pistool* zijn bedoeld om anderen te verwonden. (Weapons like a sword, knife and *rubbish bag/bow/pistol* are meant to hurt others.)

Badkamers met onder andere een douche, toilet en *memo/bidet/bad* vind je in dit hotel. (Bathrooms with among other things a shower, toilet and *note/bidet/bath* can be found in this hotel.)

Vissen zoals de snoek, baars en *vlag/haai/paling* leven in water. (Fishes like the pike, perch and *flag/ shark/eel* live in the water.)

Zeedieren zoals garnalen, inktvis en *vlieger/anemoon/ kreeft* worden in dit restaurant vers bereid. (Marine ani-mals like shrimps, octopus and *kite/anemone/lobster* are prepared freshly in this restaurant.)

Het oog bestaande uit onder andere een pupil, iris en *sticker/wenkbrauw/netvlies* is erg gevoelig. (They eye consisting of among other things a pupil, iris and *sticker/ eyebrow/retina* is very sensitive.)

Notes

1. Recently, Kuperberg (2007) and Kuperberg et al. (2006) have extended their view and put more emphasis on integration costs and factors determining the likelihood that an anomaly will be detected. Specifically, in her recent review, Kuperberg proposes a language comprehension system with at least two interacting processing streams (see below for a discussion of her view vs. the monitoring view).

2. LSA (http://lsa.colorado.edu/) is a method to obtain the estimated correlations of the co-occurrence of word pairs in large text corpora. These estimates have been found to be well correlated with association strength or semantic similarity (Landauer, Foltz, & Laham, 1998). Chwilla and Kolk (2002) showed that LSA is a sensitive method to detect subtle differences in semantic relatedness between words that were ''associatively unrelated'' according to other measures, such as lexical co-occurence and free association. In addition, Chwilla and Kolk (2005) found that conceptual scripts made of word triplets (e.g., PROMISE–SILENCE–GRAVE), that were not semantically and/or associatively related, did have higher LSA values than script-unrelated word triplets (e.g., VACATION–TRIAL–DISMISSAL). Furthermore, Van Herten et al. (2006) used LSA successfully to assess the plausibility of sentence parts of their stimuli.

3. The N400 effect to the mildly implausible sentences was significant at anterior and posterior sites for both the left hemisphere and the right hemisphere. For the deeply implausible sentences, the N400 effect was significant at anterior sites and right posterior sites, whereas the effect for the posterior part of the left hemisphere was limited to one site (LTP). This distribution difference could be taken to indicate that (partly) different cognitive processes contributed to the N400 effect in the mildly and deeply implausible condition. Based on the present data, we cannot rule out this possibility with certainty. However, we think that the same cognitive process (i.e., of semantic integration) elicited the N400 effect in the two conditions. What we propose differed between the two conditions was the resolution of the integration difficulties, which only in the case of the deeply implausible sentences, where the initial attempt at integration failed, triggered reanalysis.

4. In addition to the differences in studies regarding the presence of either a monophasic N400 effect or a biphasic pattern to semantically incongruous endings, inconsistencies have also been shown in the literature on metaphor comprehension. In particular, some studies did not find a positivity following the N400 effect to metaphors (e.g., Arzouan, Goldstein, & Faust, 2007; Pynte, Besson, Robichon, & Poli, 1996), whereas others did find a biphasic pattern (e.g., Coulson & Van Petten, 2002). Coulson and Van Petten (2002) interpreted the late posterior positivity as reflecting the recovery and integration of additional information from semantic memory, which might have been triggered by the earlier semantic mismatch (N400). The monitoring theory could account for the positivities found to metaphors as well; a strong conflict between the representation based on the literal meaning and the metaphorical meaning of the sentence triggers reanalysis.

REFERENCES

Arzouan, Y., Goldstein, A., & Faust, M. (2007). Brainwaves are stethoscopes: ERP correlates of novel metaphor comprehension. *Brain Research, 1160,* 69–81.

Chwilla, D. J., Hagoort, P., & Brown, C. M. (1998). The mechanism underlying backward priming in a lexical decision task: Spreading activation versus semantic matching. *Quarterly Journal of Experimental Psychology, 51A,* 531–560.

Chwilla, D. J., & Kolk, H. J. (2002). Three-step priming in lexical decision. *Memory & Cognition, 30,* 217–225.

Chwilla, D. J., & Kolk, H. J. (2005). Accessing world knowledge: Evidence from N400 and reaction time priming. *Cognitive Brain Research, 25,* 589–606.

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes, 13,* 21–58.

Coulson, S., & Van Petten, C. (2002). Conceptual integration and metaphor: An event-related potential study. *Memory & Cognition, 30,* 958–968.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language, 31,* 218–236.

Donchin, E. (1981). Surprise!…Surprise! *Psychophysiology, 18,* 493–513.

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.

Ford, J. M., Woodward, S. H., Sullivan, E. V., Isaacks, B. G., Tinklenberg, J. R., Yesavage, J. A., et al. (1996). N400 evidence of abnormal responses to speech in Alzheimer's disease. *Electroencephalography and Clinical Neurophysiology, 99,* 235–246.

Friederici, A. D., Pfeifer, E., & Hahne, A. (1993). Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain Research, 1,* 183–192.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2002). *Cognitive neuroscience: The biology of the mind* (2nd ed.). New York: W.W. Norton and Company.

Gunter, T. C., Jackson, J. L., & Mulder, G. (1992). An electrophysiological study of semantic processing in young and middle-aged academics. *Psychophysiology, 29,* 38–54.

Gunter, T. C., Stowe, L. A., & Mulder, G. (1997). When syntax meets semantics. *Psychophysiology, 34,* 660–676.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes, 8,* 439–483.

Heinze, H.-J., Muente, T.-F., & Kutas, M. (1998). Context effects in a category verification task as assessed by event-related brain potential (ERP) measures. *Biological Psychology, 47,* 121–135.

Hoeks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research, 19,* 59–73.

Holcomb, P. J. (1993). Semantic priming and stimulus degradation: Implications for the role of the N400 in language processing. *Psychophysiology, 30,* 47–61.

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.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109,* 679–709.

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes, 15,* 159–201.

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language, 52,* 205–225.

Kolk, H., & Chwilla, D. (2007). Late positivities in unusual situations. *Brain and Language, 100,* 257–261.

Kolk, H. H. J., Chwilla, D. J., Van Herten, M., & Oor, P. J. W. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language, 85,* 1–36.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146,* 23–49.

Kuperberg, G. R., Caplan, D., Sitnikova, T., Eddy, M., & Holcomb, P. J. (2006). Neural correlates of processing syntactic, semantic, and thematic relationships in sentences. *Language and Cognitive Processes, 21,* 489–530.

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research, 17,* 117–129.

Kutas, M., & Hillyard, S. (1980a). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology, 11,* 99–116.

Kutas, M., & Hillyard, S. (1980b). Reading between the lines: Event-related potentials during natural sentence processing. *Brain and Language, 11,* 354–373.

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

Kutas, M., & Van Petten, C. K. (1994). Psycholinguistics electrified: Event-related brain potential investigations. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 83–143). San Diego, CA: Academic Press.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14,* 41–104.

McGuire, P. K., Silbersweig, D. A., & Frith, C. D. (1996). Functional neuroanatomy of verbal self-monitoring. *Brain, 119,* 907–917.

Münte, T. F., Heinze, H.-J., Matzke, M., Wieringa, B. M., & Johannes, S. (1998). Brain potentials and syntactic violations revisited: No evidence for specificity of the syntactic positive shift. *Neuropsychology, 36,* 217–226.

Núñez-Peña, M. I., & Honrubia-Serrano, M. L. (2005). N400 and category exemplar associative strength. *International Journal of Psychophysiology, 56,* 45–54.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh Inventory. *Neuropsychologia, 9,* 97–113.

Osterhout, L., & Hagoort, P. (1999). A superficial resemblance does not necessarily mean you are part of the family: Counterarguments to Coulson, King and Kutas (1998) in the P600/SPS-P300 debate. *Language and Cognitive Processes, 14,* 1–14.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language, 31,* 785–806.

Osterhout, L., McKinnon, R., Bersick, M., & Corey, V. (1996). On the language specificity of the brain response to syntactic anomalies: Is the syntactic positive shift a member of the P300 family? *Journal of Cognitive Neuroscience, 8,* 507–526.

Pynte, J., Besson, M., Robichon, F.-H., & Poli, J. (1996). The time-course of metaphor comprehension: An event related potential study. *Brain and Language, 55,* 293–316.

Schwarz, T. J., Kutas, M., Butters, N., Paulsen, J. S., & Salmon, D. P. (1996). Electrophysiological insights into the nature of the semantic deficit in Alzheimer's disease. *Neuropsychologia, 34,* 827–841.

Stuss, D. T., & Benson, D. F. (1986). *The frontal lobes.* New York: Raven Press.

Swick, D., Kutas, M., & Knight, R. T. (1998). Prefrontal lesions eliminate the LPC but do not affect the N400 during sentence reading. *Journal of Cognitive Neuroscience Supplement, 5,* 29.

Van Berkum, J. J. A., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of Cognitive Neuroscience, 11,* 657–671.

Van Herten, M., Chwilla, D. J., & Kolk, H. J. (2006). When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *Journal of Cognitive Neuroscience, 18,* 1181–1197.

Van Herten, M., Kolk, H. H. J., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research, 22,* 241–255.

Van Petten, C., & Luka, B. J. (2006). Neural localization of semantic context effects in electromagnetic and hemodynamic studies. *Brain and Language, 97,* 279–293.

Vasey, M. W., & Thayer, J. F. (1987). The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution. *Psychophysiology, 24,* 479–486.

Vissers, C. Th. W. M. (2008). *Monitoring in language perception: An electrophysiological investigation* (Doctoral dissertation, Radboud University Nijmegen).

Vissers, C. Th. W. M., Chwilla, D. J., & Kolk, H. H. J. (2006). Monitoring in language perception: The effect of misspellings of words in highly constrained sentences. *Brain Research, 1106,* 150–163.

Vissers, C. Th. W. M., Chwilla, D. J., & Kolk, H. H. J. (2007). The interplay of heuristics and parsing routines in sentence comprehension: Evidence from ERPs and reaction times. *Biological Psychology, 75,* 8–18.

Vissers, C. Th. W. M., Kolk, H. H. J., Van de Meerendonk, N., & Chwilla, D. J. (2008). Monitoring in language perception: Evidence from ERPs in a picture–sentence matching task. *Neuropsychologia, 46,* 967–982.

Woodward, S. H., Ford, J. M., & Hammett, S. C. (1993). N4 to spoken sentences in young and older subjects. *Electroencephalography and Clinical Neurophysiology, 87,* 306–320.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review, 111,* 931–959.