Cognitive Science (2016) 1–32

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12384

# The Importance of Reading Naturally: Evidence From Combined Recordings of Eye Movements and Electric Brain Potentials

Paul Metzner,[a] Titus von der Malsburg,[b] Shravan Vasishth,[a] Frank Rösler[c]

[a]*Department of Linguistics, University of Potsdam*
[b]*Department of Psychology and Department of Linguistics, University of California, San Diego*
[c]*Biological Psychology and Neuropsychology, University of Hamburg*

Received 16 October 2014; received in revised form 7 March 2016; accepted 10 March 2016

**Abstract**

How important is the ability to freely control eye movements for reading comprehension? And how does the parser make use of this freedom? We investigated these questions using coregistration of eye movements and event-related brain potentials (ERPs) while participants read either freely or in a computer-controlled word-by-word format (also known as RSVP). Word-by-word presentation and natural reading both elicited qualitatively similar ERP effects in response to syntactic and semantic violations (N400 and P600 effects). Comprehension was better in free reading but only in trials in which the eyes regressed to previous material upon encountering the anomaly. A more fine-grained ERP analysis revealed that these regressions were strongly associated with the well-known P600 effect. In trials without regressions, we instead found sustained centro-parietal negativities starting at around 320 ms post-onset; however, these negativities were only found when the violation occurred in sentence-final position. Taken together, these results suggest that the sentence processing system engages in strategic choices: In response to words that don't match built-up expectations, it can either explore alternative interpretations (reflected by regressions, P600 effects, and good comprehension) or pursue a "good-enough" processing strategy that tolerates a deficient interpretation (reflected by progressive saccades, sustained negativities, and relatively poor comprehension).

*Keywords:* Reading; Sentence comprehension; ERP; Eye movements; Regressions

## 1. Introduction

When we listen to speech, we process words in the order in which they are uttered and we have little control over their rate. Reading is different. In reading, we look at every

Correspondence should be sent to Titus von der Malsburg, Department of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0109. E-mail: malsburg@ucsd.edu

word for as long as we wish, and we are free to skip and revisit words. Eye-tracking research has demonstrated that we make ample use of this freedom. Words that are difficult to integrate with the evolving interpretation of a sentence are typically fixated longer and can be followed by leftward eye movements (regressions). If a word is easy to process, or if it can even be guessed from the context, we may not look at it at all (Rayner, 1998).

Researchers in psychology and psycholinguistics often use a presentation format where this freedom to navigate the sentence is taken away from the reader. One example is auto-paced word-by-word presentation, otherwise known as rapid serial visual presentation (RSVP). In this form of reading, one word is shown at a time and each word is presented for a fixed duration. This presentation format has primarily been used in research investigating event-related brain potentials (ERPs; Hagoort, Hald, Bastiaansen, & Petersson, 2004; Kutas & Hillyard, 1980; Osterhout & Holcomb, 1992). The motivation for this was that the eye movements necessary for free reading generate electric potentials that contaminate the electroencephalogram (EEG).

The implicit assumption in this research is that comprehension is largely unaffected by this highly constrained form of reading. This assumption is reasonable because, as mentioned above, word-by-word reading and spoken comprehension are constrained in similar ways. On the other hand, a recent study by Schotter, Tran, and Rayner (2014) has shown that interfering with the reader's ability to revisit earlier words can seriously impair comprehension, which suggests that the task demands associated with reading may be more different from those in spoken comprehension than is often appreciated.

An important question is therefore whether comprehension mechanisms in word-by-word reading and free reading differ, and if yes, how they differ. The answers to these questions may considerably improve our understanding of the mechanisms underlying reading comprehension but they also have important methodological implications for ERP research. In this study, we address these questions by comparing sentence comprehension during word-by-word presentation and natural reading. Specifically, we compare classic ERP effects found during word-by-word presentation (N400 and P600) to the analogous ERP effects found in free reading. Further, we investigate how these effects relate to regressive eye movements which seem to play such an important role for reading comprehension (Schotter et al., 2014; see also Godfroid et al., 2015).

### *1.1. Modulating sentence comprehension difficulty*

In order to systematically modulate comprehension difficulty, we adapted the design by Hagoort (2003). Participants read German sentences containing words that violated either syntactic or semantic constraints. We included both syntactic and semantic violations because both have been studied extensively in psycholinguistic research and their ERP correlates are well understood. The task was to judge the well-formedness of the sentences, and we assumed that performance in this task is subserved by sentence comprehension mechanisms. Although Hagoort used Dutch sentences, the language of our

experiment was German. However, due to the similarity of Dutch and German, our material was very close to that used by Hagoort.

An example of a syntactic violation is as follows. At the start of a sentence and in the absence of any other information, the feminine-marked determiner *Die* ('the') raises an expectation for a feminine-marked noun. If instead a masculine noun is encountered, this should be a surprise to the reader. At this stage, the comprehension system can react in one of several ways. One option is to initiate a reanalysis attempt and reject the resulting structure as ungrammatical since there is no alternative interpretation which could resolve the gender mismatch. However, a by-product of failed reanalysis is the increased certainty that the critical word must be wrong, and this should lead to good performance in the judgment task. Syntactic anomalies like these are known to robustly trigger the so-called P600 effects, that is, positive deflections of the EEG around 600 ms after the onset of the critical word (e.g., Osterhout & Holcomb, 1992).

Alternatively, reanalysis may not be initiated, either because the violation is not detected (due to lack of attention, etc.), or because the comprehension system decides to build a partially well-formed representation, treating the interpretation of the sentence as "good enough" (Ferreira & Patson, 2007).

A violation of world-knowledge or semantic constraints can be introduced through nonsensical adjective-noun pairs. For example, the adjective *inquisitive* would raise an expectation for a noun representing an animate referent; surprise should result if the next word instead is an inanimate-referring noun such as *farm*. As in the case of a syntactic violation, this type of violation could either result in the recognition of an anomaly, potentially followed by attempted reanalysis, or in mistakenly treating the adjective-noun combination as an acceptable collocation.

Semantic anomalies like these are known to trigger N400 effects, that is, negative-going deflections of the EEG around 400 ms post-onset (e.g., Kutas & Hillyard, 1980). However, they may also be accompanied by post-N400 P600-effects, in particular, if strong semantic expectations are induced by the preceding context (DeLong, Quante, & Kutas, 2014; Kim & Osterhout, 2005; Van De Meerendonk, Kolk, Chwilla, & Vissers, 2009).[1]

The reanalysis processes that are potentially triggered by syntactic and semantic violations can be seen as part of a broader class of recovery processes in sentence comprehension. A prominent example are those found in "garden-path" sentences. Given a sentence fragment such as *The lawyer examined…*, comprehenders typically interpret the lawyer as the agent doing the examining, and most comprehenders will have a strong expectation that the next constituent will be the object of the examination, as in *The lawyer examined the evidence*. However, the sentence could also continue with *… by the nurse was ill*, that is, with a reduced relative clause which would be ungrammatical under the favored interpretation of *examined* as the main verb of the sentence. The dashed expectation that results from a reduced relative continuation leads to a search for alternative syntactic structures. This search is associated with longer fixation durations and higher rates of regressions in eye-tracking studies (Braze, Shankweiler, Ni, & Palumbo, 2002; Clifton, Staub, & Rayner, 2007; Frazier & Rayner, 1987); and with a centro-parietal negativity

(Hopf, Bader, Meng, & Bayer, 2003), and/or a P600 effect (Gouvea, Phillips, Kazanina, & Poeppel, 2010; Osterhout & Holcomb, 1992) in ERP studies.

A close correspondence has been observed between the recovery process in garden-path sentences and in outright ungrammatical structures as those investigated in the present study: At the earliest moments of processing, the same recovery processes are believed to be initiated in both types of sentences. For example, Hopf et al. (2003) compared ungrammatical and garden-path sentences in German using ERPs, and reported a negativity in the 300–500 ms range, with a similar onset latency, amplitude, and centro-parietal scalp distribution in both types. Similarly, Gouvea et al. (2010) demonstrated that in English, garden-path sentences (conjunct attachment ambiguities) and syntactic violations (agreement mismatch) both trigger P600 effects. Both papers conclude that similar recovery processes are triggered when the anomaly is detected. Of course, in garden paths, the end result should generally be a grammatical structure, whereas in a syntactic violation or a semantic anomaly, the end result should be the full recognition of the violation.

Given these considerations, syntactic violations and semantic anomalies are a good choice for investigating more general recovery processes in reading. They constitute a simple experimental manipulation, are well-studied in the ERP and eye-tracking literature, and trigger a class of recovery processes that is of great importance in sentence comprehension research.

Hagoort (2003), the study that inspired our work, did not just manipulate the type of anomaly (syntactic vs. semantic) but also where this anomaly occurred within the sentence. The motivation for this manipulation was that expectation about upcoming words may be sharpened as more information is accumulated. For example, when we hear *The deteriorating*..., we may not be as sure about the identity of an upcoming noun compared to when we hear *The experienced actor played a difficult*.... One possible outcome of such a change in certainty as a function of word position is that a violation later on in a sentence could result in a greater surprise and easier detection of the violation compared to an earlier position in the sentence. This prediction is consistent with the surprisal account of sentence processing which posits that the difficulty of processing a word is related to the probability of that word conditional on the preceding words (Hale, 2001; Levy, 2008). In other words, this account predicts that a highly unexpected word should cause greater disruption in sentence processing than a less surprising word. Hagoort's results are partly consistent with this prediction: He found that N400 effects in response to semantic anomalies were larger in sentence-final positions than in sentence-medial positions. One possible consequence of higher disruptions in late sentence positions may be improved detection of the anomalies due to a greater mismatch of expected input and actual input.

An alternative possible effect of word position is that, due to higher certainty towards the end of a sentence, the reader pays *less* attention to the visual input. This second possibility can be seen as the modulation of the comprehender's strength of belief about the message: In early parts of the sentence the reader may have weaker prior beliefs about the content and may therefore be more willing to attend to the written text; but in later

parts they may have formed a much stronger belief, so strong that the visual input does not have enough weight to sway that belief. One prediction would be reduced accuracy in detecting the anomaly. However, since the acceptability judgment accuracies in the Hagoort study were close to 100% for all conditions, there was no evidence in favor of either this prediction or the reverse prediction discussed above.

In sum, the position of the anomaly may be an important factor for investigating whether and how the comprehension system detects linguistic anomalies. Following Hagoort, we therefore included syntactic or semantic violations in a sentence-medial versus sentence-final position. As a baseline, we also included analogous sentences that had no anomaly. Thus, the experimental material used in our study followed a $2 \times 3$ factorial design: position (sentence medial vs. final) and violation type (none, syntactic, semantic). Hagoort's study also had a condition with a combined syntactic and semantic violation. We omitted this condition because we were not primarily interested in the joint effect of these violations but rather in the difference in comprehension between word-by-word and free reading.

We tested the six conditions described above using word-by-word presentation (RSVP) and whole-sentence presentation permitting free reading. In both reading conditions, we recorded the EEG. During free reading we also concurrently recorded eye movements. The concurrent recording and joint-analysis of eye movements and brain potentials has been made possible by recent advances in signal processing that enable the removal of eye movement artifacts from EEG recordings (Dimigen, Sommer, Hohlfeld, Jacobs, & Kliegl, 2011; Jung et al., 2000; Makeig, Bell, Jung, & Sejnowski, 1996).

Earlier studies have shown that ERP effects during free reading can be qualitatively similar to those found during word-by-word reading. However, these studies were more limited in scope than ours. For example, Kretzschmar, Bornkessel-Schlesewsky, and Schlesewsky (2009) demonstrated N400 effects in response to world-knowledge violations that were similar to those found in classic studies using word-by-word presentation (e.g., Kutas & Hillyard, 1980). These effects emerged, although no corrections for eye movement artifacts were applied. The N400 effects occurred slightly earlier than in word-by-word reading, but this can be attributed to orthographic pre-activation: Even before the critical word is fixated for the first time, the system receives input from this word through parafoveal preview during the fixation on the previous word (Baccino & Manunta, 2005; Starr & Inhoff, 2004; White, 2008, see also Rayner, 1998, for a review).

Similarly, Dimigen et al. (2011) showed N400 effects in response to words that were relatively unpredictable but otherwise meaningful in the context. That study, too, found that the onset of N400 occurred earlier during free reading than might be expected based on results from RSVP studies. A related study by Dimigen, Sommer, and Kliegl (2007) also showed that long regressive eye movements (two words back or more) can be accompanied by "P600-like" effects. However, the stimuli used in that study were not linguistically controlled and did not contain anomalies or structural ambiguities, which complicates the interpretation of the results.

### 1.2. Predictions

As discussed above, our main goal was to investigate whether and how reading modality influences sentence processing. Given the rich literature on the processing of syntactic and semantic violations, we can derive rather specific predictions for the EEG data, eye movements, and accuracy in the judgment task.

In the aggregate, that is, averaged across all trials, we expected to see qualitatively similar ERP effects for word-by-word presentation and free reading. However, based on the research by Kretzschmar et al. (2009) and Dimigen et al. (2011), we also expected N400 effects and potentially P600 effects to emerge earlier in free reading than in word-by-word reading. Specifically, we expected to see N400 effects in response to semantic violations in medial and final positions. N400 effects in final position were expected to have larger amplitudes reflecting sharpened expectations and hence a stronger mismatch. We also expected N400 effects for syntactic violations but only in sentence-final position (in accordance with Hagoort's results). Further, we expected the canonical P600 effect in response to syntactic violations. Hagoort's data showed numerical trends suggesting P600 effects for semantic violations, but these were not significant. However, a number of related studies suggest that the semantic violations may be accompanied by a biphasic N400-P600 effects (e.g., DeLong et al., 2014; Kim & Osterhout, 2005; Kuperberg, 2007; Van De Meerendonk et al., 2009).

The study by Schotter et al. (2014) suggests that regressions may subserve thorough comprehension (see also Godfroid et al., 2015). The present experimental design combined with the use of coregistration allowed us not only to confirm this but also to shed light on *why* that may be the case. Specifically, we could investigate the neural correlates of regressions and determine whether and how regressions are linked to N400 and P600 effects. There is a range of possible outcomes but based on earlier eye-tracking and ERP studies, one likely possibility is that regressions are linked to P600 effects as both have been observed in sentences that trigger recovery processes in response to incongruous words.

## 2. Materials and methods

### 2.1. Participants

We tested 72 students at the University of Potsdam, Germany (18 male, 54 female, $M_{age}$: 25 years). The ORSEE software was used for participant recruitment (Greiner, 2004). The number of participants was determined before the start of the study. Twenty-four participants were randomly selected to read word-by-word (the same number of participants was tested in the study by Hagoort, 2003). The remaining 48 participants read the text naturally.

Random assignment to reading modalities was used to allow comparisons between the groups. The group-level statistics show that this approach yielded a

homogeneous distribution of participants across groups: The mean age in the word-by-word group was 25.8 and 25.2 years in the group that read naturally. The percentage of women was 72% in the word-by-word group and 76% in the natural-reading group.

The reason for testing twice as many participants in the natural reading condition was that we knew from earlier studies that ERP data recorded during natural reading has a lower signal-to-noise ratio (e.g., Metzner, von der Malsburg, Vasishth, & Rösler, 2015). This is unsurprising because, as discussed above, natural reading gives participants additional degrees of freedom: They can individually adjust the pace, they can skip words, and they can revisit earlier material. All these sources of variance are absent during auto-paced word-by-word presentation. However, having more data also allowed us to conduct a post hoc split of the ERP data in order to investigate the neural correlates of regressive eye movements.

### 2.2. Materials

The tested language was German. The materials followed a 2 × 3, design: the sentences contained a noun that introduced either a syntactic, a semantic, or no violation and that noun occurred in the middle or at the end of the sentence (see Table 1 for example sentences). Sentences with syntactic violations had a noun whose grammatical gender was different from that of its determiner ("the$_{FEM}$ deteriorating farm$_{MASC}$"). Sentences with semantic violations had an adjective that was incongruent with the following noun given commonsense knowledge ("the inquisitive farm").

To control adjectives for frequency, we created clusters using *k*-means clustering on logarithmized frequency (extracted from the lexical database dlexDB, Heister et al., 2011) and paired adjectives belonging to the same cluster. In the sentence-medial conditions, the adjectives used in baseline sentences and syntactic violations had an average

Table 1
Sample set of sentences with English translation. Manipulated word is italicized and the target noun is in boldface. The determiner's gender is indicated by a subscript (*MASC* = masculine, *FEM* = feminine, *NEUT* = neuter). 'Syn' and 'Sem' in parentheses denote whether the violation was syntactic or semantic

| | | | | |
|---|---|---|---|---|
| **Sentence-medial:** | | | | |
| Der$_{MASC}$ | verfallene | **Bauernhof**$_{MASC}$ | braucht eine Renovierung. Er wird von einer Familie bewohnt. | |
| *Die*$_{FEM}$ | verfallene | **Bauernhof**$_{MASC}$ | braucht eine Renovierung. Er wird von einer Familie bewohnt. | (Syn) |
| Der$_{MASC}$ | *neugierige* | **Bauernhof**$_{MASC}$ | braucht eine Renovierung. Er wird von einer Familie bewohnt. | (Sem) |
| The$_{MASC}$/*The*$_{FEM}$ deteriorating/*inquisitive***farm**$_{MASC}$ needs a renovation. It is inhabited by a family. | | | | |
| **Sentence-final:** | | | | |
| Der erfahrene Star spielt die$_{FEM}$ | schwierige | **Rolle**$_{FEM}$. | Er überzeugt seine schärfsten Kritiker. | |
| Der erfahrene Star spielt *das*$_{NEUT}$ | schwierige | **Rolle**$_{FEM}$. | Er überzeugt seine schärfsten Kritiker. | (Syn) |
| Der erfahrene Star spielt die$_{FEM}$ | *elektrische* | **Rolle**$_{FEM}$. | Er überzeugt seine schärfsten Kritiker. | (Sem) |
| The experienced star plays the$_{FEM}$/*the*$_{NEUT}$ difficult/*electric***Role**$_{FEM}$. He convinces his harshest critics. | | | | |

log-frequency of 1.14, and the adjectives giving rise to the semantic violation had an average log-frequency of 1.02. In the sentence-final conditions, the adjectives used in the baseline and syntactic conditions had a average log-frequency of 0.86, and the adjectives in the semantic condition had an average log-frequency of 0.75. Adjectives within an item differed by no more than two characters in length (sentence-medial: $M = 9.02$ in baseline and syntactic violations, $M = 9.04$ in semantic violations; sentence-final: $M = 9.52$ in baseline and syntactic violations, $M = 9.52$ in semantic violations). Neither the differences in frequency nor the differences in length were statistically significant. To avoid an influence of grammatical gender, we balanced the number of nouns with male, female, and neuter gender within each position and condition.

### 2.3. Apparatus

The EEG was recorded using a shielded electrode cap with 32 Ag/AgCl electrodes (Advanced Neuro Technology, Enschede, Netherlands) mounted following a variant of the 10–20 layout. Bipolar electrodes were placed on the left and right outer canthus and the infraorbital ridges of the right eye to record the electrooculogram. Recording was at a sampling rate of 512 Hz and with an anti-aliasing low-pass filter at 138 Hz. Impedances were kept below 5 kΩ and recordings were referenced against the left mastoid. After the experiment, the recordings were computationally rereferenced to linked mastoids. Eye movements were recorded with an EyeLink 1000 eye-tracker (SR Research, Mississauga, ON, Canada) with a sampling rate of 1,000 Hz, a spatial resolution of 0.01°, and an average accuracy of 0.32° in the area where the sentences were presented (0.53° overall).

### 2.4. Procedure

Participants were seated in a dimly lit, electromagnetically shielded, and sound-insulated booth. The eyes were approximately 24″ (60 cm) from the presentation screen, which had a diagonal length of 22″ (56 cm). An experimental session began with 10 practice trials to familiarize the participant with the procedure. In sessions with word-by-word presentation, each trial began with a fixation dot in the center of the screen. After 1,000 ms, the dot was removed and the sentence was displayed word by word in the same position. Every test sentence was presented together with a follow-up sentence to avoid end-of-trial effects in the final region of interest and to increase the difficulty of the judgment task.

Each participant read 360 test/follow-up sentence pairs (60 per condition) randomly interspersed with 180 similar pairs of distractor sentences. Since the position of the violation was manipulated between items, there were three lists of sentences and each participant saw only one version of each sentence (Latin square design).

Each word was presented for 300 ms followed by a 300 ms inter-stimulus interval. The final words of the two sentences were presented together with a period. After the last word of the second sentence, a blank display with a pseudo-randomly varying interval between 1 and 2 seconds was shown. This was followed by the prompt of the judgment

task. Following the design by Hagoort (2003), participants were prompted by a row of asterisks to decide whether or not the sentences they had just read were well-formed. Responses were given with a button press. Following the response, a blank display of 1,150 ms preceded the onset of the next trial.

In sessions testing natural reading, trials began with a fixation dot in the vertical center at the left edge of the display. As soon as the participant had stably fixated the dot, it disappeared and the entire sentence appeared on the screen, offset by 80 pixels to the right in order to induce a saccade to the first word of the sentence. To end the sentence presentation and to proceed to the judgment task, participants had to fixate the lower right corner of the screen.

Sessions lasted between 2.5 and 3.5 h including preparations, breaks, and debriefing. In both presentation forms, participants were encouraged to take a short break every 20 trials. During these breaks, they received feedback about their performance on the judgment task. After 240, 360, and 480 trials, participants had to take mandatory longer breaks to relax their neck muscles and eyes.

### 2.5. Data preprocessing

A velocity-based saccade detection algorithm was used to detect saccades and fixations in the raw eye-tracking data (Engbert & Kliegl, 2003; von der Malsburg, 2015). Fixations shorter than 20 ms and longer than 1,000 ms were removed. This led to the loss of 0.1% of all fixations. The following eye-tracking measures were calculated for the critical word (the noun): first fixation duration, gaze duration, regression probability. First fixation duration is the duration of a first fixation on a word when it is first entered from the left. Gaze duration is the cumulative duration of all fixations during first pass on a word from the first incoming saccade until the first outgoing saccade. Regression probability is estimated by dividing the number of trials with a leftward saccade (after first entering a word from the left and before reading a word to its right) by the total number of trials in a condition.

The EEG data were preprocessed using BrainVision Analyzer 2 (Brain Products, Munich, Germany). We first resampled the signal to 500 Hz and filtered it with a bandpass filter of 0.3 through 70 Hz (both at 48 dB/oct) and a notch filter at 50 Hz (the power frequency in Germany). We then identified muscle artifacts generated by eye movements during reading using independent components analysis (Jung et al., 2000; Makeig et al., 1996). The Infomax algorithm was used for training on all distractor sentences. Components with a frontal or bipolar frontal distribution were removed from the signal such that variance in the eye channels was minimized (see Fig. 1). From the data corrected in this manner, we removed epochs with other muscle artifacts or slow drifts in a semi-automatic procedure. This resulted in the loss of 79 sentence-medial (1.8%) and 52 sentence-final trials (1.2%) in the word-by-word data, and 229 sentence-medial (2.7%) and 150 sentence-final trials (1.7%) in the natural reading data.

In the word-by-word data, the time-lock for the ERP analysis was the onset of the critical word. In the natural reading data, the time-lock was the time when the gaze first

landed on the critical word. Trials in which the target noun was skipped in the first pass were not considered for the ERP analysis, which led to the loss of 444 sentence-medial (5.1%) and 562 sentence-final trials (6.5%) in natural reading sessions. Epochs starting 200 ms preceding the onset of the critical word and ending 1,000 ms later were exported to R (R Development Core Team, 2009) for further processing. Prior to the statistical analysis, the data were baseline-corrected by subtracting the average amplitude in the 100 ms interval preceding the time-lock.

### 2.6. Analysis

Response accuracy in the judgment task was analyzed with two logistic mixed-effects models as implemented in the R package *lme4* (version 1.1.7, using the BOBYQA algorithm for optimization, Bates, Mächler, Bolker, & Walker, 2015). Accuracy was treated as a categorical variable (correct = 1, incorrect = 0). The first model investigated the effect of presentation modality on judgment accuracy. This model included as fixed factors violation type, violation position in the sentence, presentation modality (word-by-word or natural reading), and all two- and three-way interactions of these factors. Violation type was coded using a treatment contrast with the control condition as the baseline. Violation position and modality (the latter a between-participant factor) were coded using sum contrasts (medial position = −1, final position = 1, word-by-word presentation = −1, natural reading = 1). The second model investigated the effect of regressive eye movements on judgment accuracy and used only data from natural reading sessions. The predictors were the same except that modality was replaced by a predictor indicating whether or not a regression had occurred in the respective trial (no regression = −1, regression = 1). Again, all possible interactions were included.

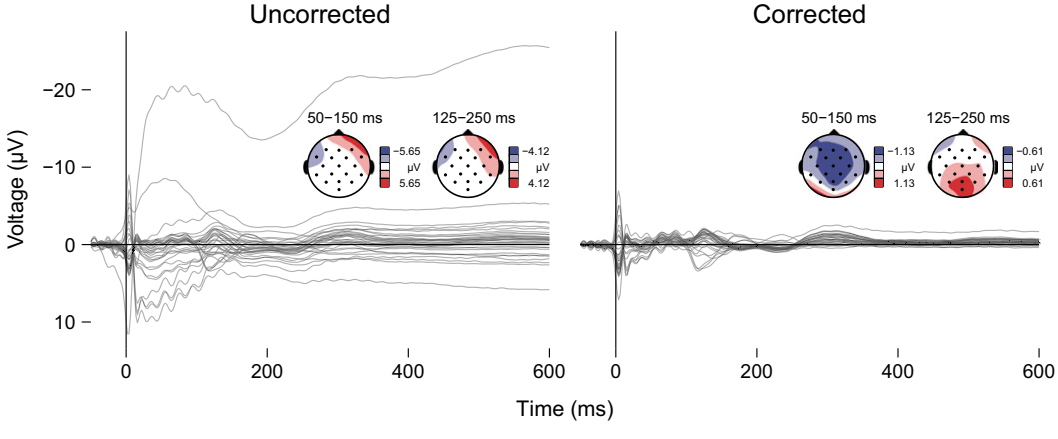


Fig. 1. Grand average ERP for randomly selected fixations before and after artifact correction. Topographic maps show mean amplitude in the N1 and P2 time window.

All models initially had the maximal random-effects structures permitted by the design. In case of convergence failure, we dropped the random effect explaining the least variance and refit the model. We also dropped correlation parameters that had values 1 or −1 (see also Barr, Levy, Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2015; Matuschek, Bates, Kliegl, Vasishth, & Baayen, 2015, for a discussion of this approach). This procedure was repeated until the model converged. All models included random intercepts for subjects and items and all but one (first fixation duration) included random subject and item slopes for violation type. An effect was considered statistically significant if the corresponding absolute $t$ or $z$ statistic was larger than 1.96.

We analyzed eye movement measures at the noun with linear and logistic mixed-effects models. The continuous variables first fixation duration and gaze duration were log-transformed to obtain approximately normally distributed residuals. The occurrence of regressions was treated as a categorical variable (regression: 1, no regression: 0) and modeled using the logit link function. Fixed effects were violation type, violation position, and their interactions. The contrast coding and procedure for determining the random effects structure were the same as described above.[2]

ERPs were analyzed with non-parametric cluster-based randomization tests (Maris & Oostenveld, 2007) which offer an elegant solution for the multiple-testing problem that often arises in the analysis of ERP data. We implemented the procedure as follows. First, paired $t$ tests were performed for the mean amplitude at each time point and electrode with one value per participant and condition. Next, samples with a test statistic that was significant at an $\alpha$ of .05 were clustered using connected-component labeling (Rosenfeld & Pfaltz, 1966). All test statistics within a cluster were then summed up to yield a cluster statistic. To assess each cluster's significance, we generated a distribution representing the null hypothesis by means of a randomization procedure: In each of 1,000 iterations, condition labels were first randomly swapped within participants. With data randomized in this manner, clusters were formed as described above and the largest cluster statistic was entered into the distribution. Clusters found in the original data were considered significant if their test statistic fell in the lower 2.5th or upper 97.5th percentile of this distribution. Note that, depending on the threshold for the individual $t$ tests, these clusters can capture long-lasting effects whose distribution on the scalp changes over time. For example, a positivity that peaks at posterior electrodes around 700 ms after stimulus onset may be connected with an earlier or later positivity at frontal electrodes. For further details, see Maris and Oostenveld (2007).

To investigate the relation between regressive eye movements and ERP effects, we split the data for the natural reading sessions in two subsets: one with trials in which a first-pass regression occurred on the critical word and one in which no such regressions occurred. These subsets were then analyzed individually using the procedure described above. Since the rate of regressions was too low in the baseline condition to conduct statistical tests, we compared ERP data from violation trials with regressions to all baseline trials taken together (irrespective of whether or not a regression had occurred in the baseline condition).

## 3. Results

### 3.1. Judgment accuracy

Fig. 2 shows the participants' performance in the judgment task. We fit two linear mixed models: one combined the data from word-by-word presentation and natural reading to investigate the effect of modality (a between-participants factor); the second model

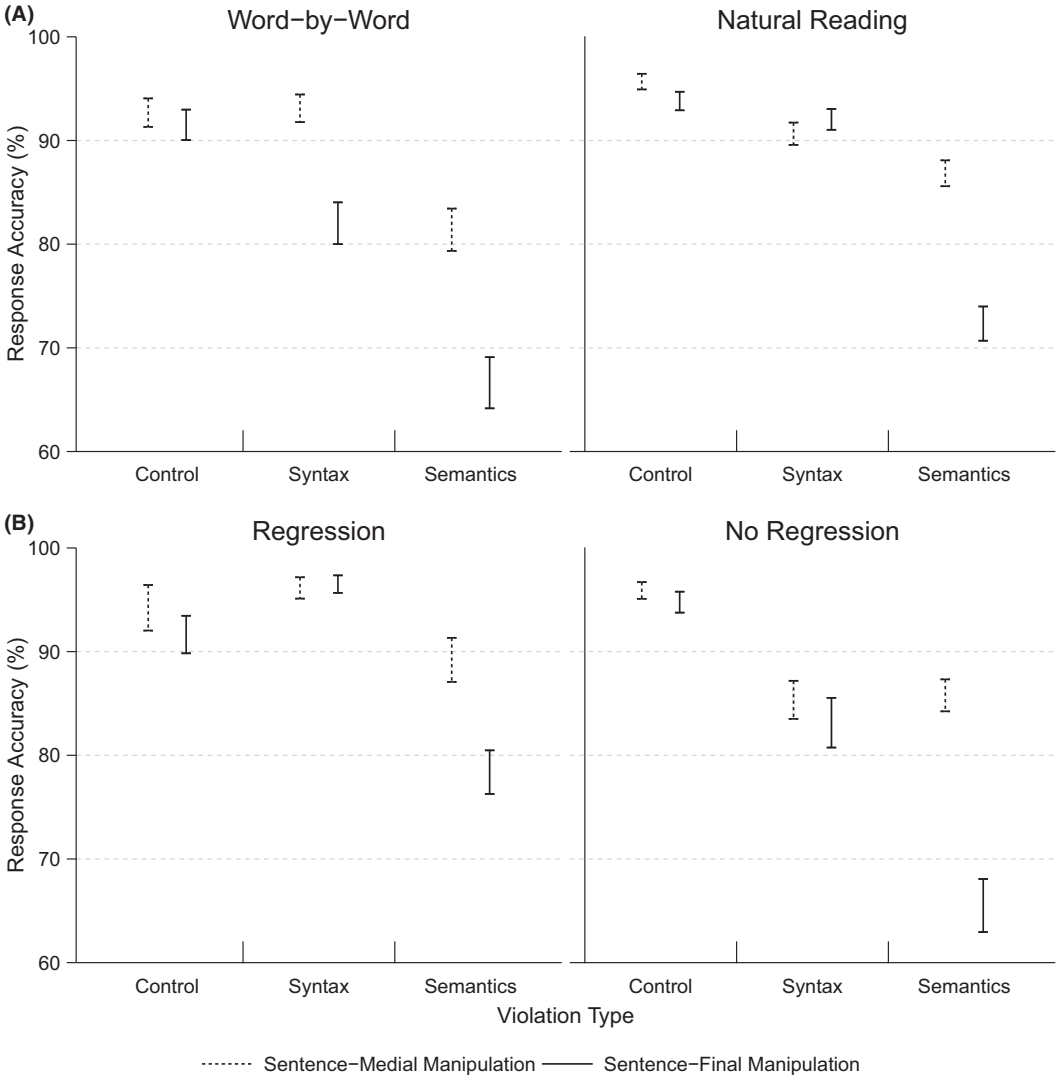


Fig. 2. Accuracy in the judgment task in word-by-word presentation and natural reading (A) and within natural reading sessions, in trials with regression and without regression (B). Dashed lines show sentence-medial manipulations and solid lines sentence-final manipulations. The bars denote 95% confidence intervals.

investigated the effect of regressions on accuracy. See Tables 2 and 3 for the parameter estimates of these two models.

Mean accuracy for target items was 87%, showing that participants were attending to the task. However, both linear mixed models showed that violations had highly significant effects on accuracy: Accuracy was lower when the sentence had a syntactic or a semantic violation (compared to the control condition). Both models also showed that accuracy was on average lower when the violation occurred in sentence-final position versus

Table 2
Summary statistics for the mixed-effects model of response accuracy as a function of violation type, position, and presentation modality. Estimates (Est.) and 95% confidence intervals (CI) are on the logit scale, statistical significance is indicated by test statistics in boldface ($z$). Treatment contrast was used for violation type (baseline = no violation), and sum contrasts were used for all other factors. The label indicates which level was set to 1

| | Est. | 95% CI | $z$ |
|---|---|---|---|
| Syntax | **−0.54** | **[−0.94, −0.14]** | **−2.6** |
| Semantics | **−1.72** | **[−2.12, −1.32]** | **−8.4** |
| Final position | −0.31 | [−0.67, 0.05] | −1.7 |
| Natural reading | **0.51** | **[0.14, 0.88]** | **2.7** |
| Syntax × Final | −0.28 | [−0.71, 0.15] | −1.3 |
| Semantics × Final | **−0.86** | **[−1.31, −0.41]** | **−3.7** |
| Syntax × Natural | −0.07 | [−0.75, 0.61] | −0.2 |
| Semantics × Natural | −0.01 | [−0.68, 0.66] | 0.0 |
| Final × Natural | −0.29 | [−0.74, 0.16] | −1.3 |
| Syntax × Final × Natural | **1.87** | **[1.33, 2.41]** | **6.8** |
| Semantics × Final × Natural | 0.05 | [−0.43, 0.53] | 0.2 |

Table 3
Summary statistics for the mixed-effects model of response accuracy in natural reading sessions as a function of violation type, position, and the occurrence of a regression. Estimates (Est.) and 95% confidence intervals (CI) are on the logit scale, statistical significance is indicated by test statistics in boldface ($z$). Treatment contrast was used for violation type (baseline = no violation) and sum contrasts were used for all other factors. The label indicates which level was set to 1

| | Est. | 95% CI | $z$ |
|---|---|---|---|
| Syntax | **−0.70** | **[−1.27, −0.13]** | **−2.4** |
| Semantics | **−1.86** | **[−2.42, −1.30]** | **−6.5** |
| Final position | **−0.57** | **[−1.09, −0.05]** | **−2.1** |
| Regression | −0.29 | [−0.70, 0.12] | −1.4 |
| Syntax × Final | 0.51 | [−0.10, 1.12] | 1.6 |
| Semantics × Final | **−0.93** | **[−1.54, −0.32]** | **−3.0** |
| Syntax × Regression | **2.00** | **[1.51, 2.49]** | **7.9** |
| Semantics × Regression | **0.93** | **[0.48, 1.38]** | **4.1** |
| Final × Regression | −0.27 | [−0.92, 0.38] | −0.8 |
| Syntax × Final × Regression | 0.59 | [−0.22, 1.40] | 1.4 |
| Semantics × Final × Regression | 0.28 | [−0.47, 1.03] | 0.7 |

medial position (marginally significant in the model testing presentation modality).[3] Further, performance was particularly poor when a semantic violation occurred in final position.

The model testing presentation modality (Table 2) showed that accuracy was significantly improved when the participants read sentences naturally instead of word-by-word. A three-way interaction was found between syntactic violation, modality, and position. This interaction was due to accuracy being lower in medial position when the sentence was presented word-by-word.

The model testing the effect of regressions (Table 3) showed that accuracy was higher in sentences with violations when a regression occurred. This effect was bigger for syntactic than for semantic violations.

The overall accuracy in our experiment (87%) was lower than in Hagoort's study (94%). This may be due to differences in design (we had two sentences in each trial but Hagoort just one), material, language (German vs. Dutch), instructions, and subject population. However, our pattern of results resembles that reported by Hagoort (Hagoort did not report accuracy broken down by sentence position.)

To check whether performance was deteriorating due to fatigue, we also conducted a split-half analysis (not reported in detail) which showed that the overall pattern of effects was remarkably similar in the two halves of the experiment; if anything, there was a small improvement in the second half, probably due to increased familiarity with the task.

### 3.2. Eye-tracking data

Fig. 3 shows means and 95% confidence intervals for the eye-tracking measures (first fixation duration, gaze duration, regression probability). The parameter estimates from the mixed-effects model analysis are listed in Table 4.

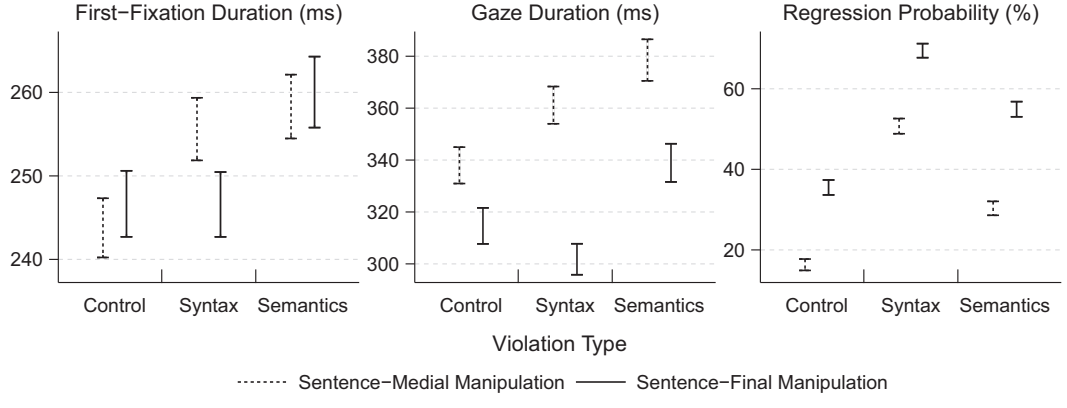


Fig. 3. 95% confidence intervals for the means of first fixation duration, gaze duration, and regression probability at the noun. Dashed lines show sentence-medial conditions and solid lines sentence-final conditions.

Table 4
Summary statistics for the mixed-effects models of first fixation duration, gaze duration, and regression probability in natural reading sessions as a function of violation type, and position. Estimates (Est.) and 95% confidence intervals (CI) are on the log scale for first fixation and gaze duration and on the logit scale for regression probability. Statistical significance is indicated by test statistics in boldface ($t$ and $z$). Treatment contrast was used for violation type (baseline = no violation) and a sum contrast was used for position (final position = 1)

| | First Fixation Duration | | | Gaze Duration | | | Regression Probability | | |
|---|---|---|---|---|---|---|---|---|---|
| | Est. | 95% CI | $t$ | Est. | 95% CI | $t$ | Est. | 95% CI | $z$ |
| Syntax | **0.02** | **[0.01, 0.03]** | **2.39** | **0.03** | **[0.02, 0.04]** | **2.43** | **1.75** | **[1.66, 1.84]** | **18.68** |
| Semantics | **0.06** | **[0.05, 0.07]** | **7.16** | **0.10** | **[0.09, 0.11]** | **10.08** | **0.90** | **[0.85, 0.95]** | **17.24** |
| Final position | 0.00 | [−0.02, 0.02] | −0.14 | **−0.08** | **[−0.11, −0.05]** | **−2.64** | **1.13** | **[1.01, 1.25]** | **9.50** |
| Syntax × Final | **−0.04** | **[−0.06, −0.02]** | **−2.60** | **−0.10** | **[−0.12, −0.08]** | **−4.52** | **−0.27** | **[−0.40, −0.14]** | **−2.02** |
| Semantics × Final | −0.01 | [−0.02, 0.00] | −0.45 | −0.03 | [−0.05, −0.01] | −1.77 | 0.07 | [−0.03, 0.17] | 0.73 |

Consistent with earlier research (e.g., Braze et al., 2002), the eye movement data showed that, compared to the control condition, readers slowed down at the critical word when a syntactic or semantic violation was encountered. This effect was found in first fixation duration and gaze duration. Similarly, regression rates were also increased in response to syntactic and semantic violations. Rates of regressions were also substantially increased in sentence-final positions.

In final position, gaze durations were shorter than in medial position. An interaction between syntactic violation and position was seen in all measures:[4] In the duration measures, a syntactic violation led to shorter fixations times sentence-finally compared to sentence-medially suggesting less processing difficulty in final position. However, regressions showed the opposite pattern: a syntactic violation led to higher regression rates sentence-finally compared to sentence-medially, which suggests *higher* processing difficulty. This seemingly paradoxical effect can be explained by rapid regressions in response to the syntactic violation that cut the first pass short. We tested this using a linear mixed model (not reported in detail) that examined the effect of regression (as a binary predictor) on fixation durations; in both first fixations and gaze duration, regressive eye movements were associated with significantly shorter durations. For this reason, we can conclude that readers experienced greater processing difficulties sentence-finally despite the shortened reading times.

### 3.3. Event-related potentials

All ERP results are summarized in Table 5. For simplicity, we will refer to negativities and positivities when we really mean relative negativities and positivities compared to control sentences.

For each ERP effect we report its onset, peak, and offset. It is important to understand that the estimated on- and offset times mark the time where the test statistic first and last surpasses the threshold in at least one electrode. By definition these are the regions where

Table 5
Summary of ERP results in the word-by-word reading and natural reading experiments

| | Word-by-Word Presentation | | Natural Reading | |
|---|---|---|---|---|
| | Syntax | Semantics | Syntax | Semantics |
| Medial | P600 | N400/P600 | P600 | N400/P600 |
| Final | N400/P600 | N400 | N400/P600 | N400/P600 |

| Regression-Contingent Analysis | | | |
|---|---|---|---|
| | | Syntax | Semantics |
| Regression | Medial | P600 | P600 |
| | Final | N400/P600 | N400/P600 |
| No regression | Medial | – | – |
| | Final | Sustained Negativity | Sustained Negativity |

the effect is weakest and noise in the data can therefore push around the detected boundaries of an effect. Specifically, it is possible that one electrode reaches the threshold considerably earlier than the others due to noise, and this means that the on- and offsets are not necessarily representative of all electrodes involved in the effect. A corollary from this is that negativities and positivities can overlap in time. When this happens, this means that some electrodes still showed the negative-going effect while others already showed the positive-going effect. For these reasons, inferences about the timing of effects should primarily be based on the peaks of effects which are highly reliable. In addition we refer the reader to the plots in the Appendix which shows the waveforms for all electrodes.

#### *3.3.1. Word-by-word presentation*

In the condition with word-by-word presentation, we found results similar to those reported by Hagoort (2003, see Fig. 5). Syntactic violations in sentence-medial position led to a P600-like centro-parietal positivity (544–1,000 ms, peak: 821 ms, $p < .001$); in sentence-final position, an N400-like effect (312–526 ms, peak: 428 ms, $p < .05$) was followed by a late centro-parietal positivity (570–1,000 ms, peak: 688 ms, $p < .001$).

Semantic violations in sentence-medial position elicited a short-lived centro-parietal negativity (476–520 ms, peak: 504 ms, $p < .01$) and a centro-parietal positivity (684–1,000 ms, peak: 836 ms, $p < .01$); in sentence-final position, a sequence of four centro-parietal negativities together constituted an N400 effect (128–656 ms, peaks: 170, 230, 358, 536 ms; all $p < .001$).[5]

#### *3.3.2. Natural reading*

During natural reading (Fig. 4), syntactic violations in sentence-medial position elicited a P600-like centro-parietal positivity (478–1,000 ms, peak: 828 ms, $p < .001$); in sentence-final position, a similar centro-parietal positivity (590–1,000 ms, peak: 776 ms, $p < .001$) was preceded by an N400-like occipito-parietal negativity (130–434 ms, peak: 216 ms, $p < .01$).

Semantic violations led to an N400/P600 response in both sentence-medial and sentence-final position. Sentence-medially, it comprised an occipito-parietal negativity (230–372 ms, peak: 268 ms, $p < .05$) and a centro-parietal positivity (760–1,000 ms, peak: 980 ms, $p < .01$). Sentence-finally, both effects occurred slightly earlier (104–522 ms, peak: 360 ms, $p < .001$, and 696–988 ms, peak: 842 ms, $p < .01$, respectively).

#### *3.3.3. Regression-contingent analysis*

In sentences with syntactic violations, the eyes regressed in 60% of the cases and in sentences with semantic violations in about 43% of the cases. As shown in Fig. 3, regression probabilities were on average higher in violation than control sentences, and higher in final position compared to medial position. To compare ERP effects in trials with and without regressions, we split the data into two subsets, one in which regressions occurred during first pass, and the other in which no first-pass regressions occurred from the noun. We will refer to these as the regression and no-regression trials, respectively. We applied the same statistical analysis as for the full data-set (see Fig. 4).

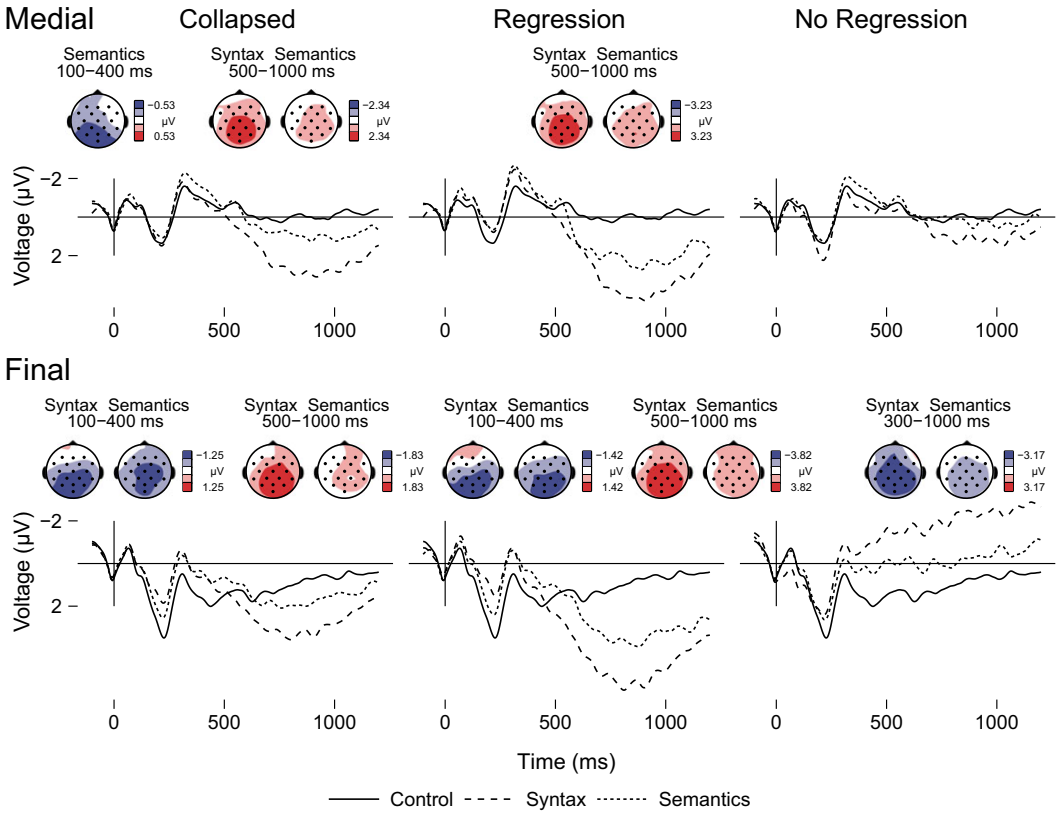


Fig. 4. Average ERPs from natural reading sessions at electrode Pz for control sentences (solid line), syntactic violations (dashed line), and semantic violations (dotted line) at the noun in sentence-medial and sentence-final position. Isovoltage maps show the topographic distributions of amplitude differences (violation minus control) from 100 to 400 ms (N400), from 500 to 1,000 ms (P600), and from 300 to 1,000 ms (sustained negativities).

*3.3.3.1. Regression trials*: Considering trials in which a first-pass regression occurred at the target noun, in sentence-medial conditions both violation types elicited P600 effects (syntax: 290–1,000 ms, peak: 828 ms, $p < .001$; semantics 540–1,000 ms, peak: 848 ms, $p < .001$). In sentence-final position, both violations elicited an N400/P600 response; in syntactic violations, a centro-parietal negativity (24–378 ms, peak: 164 ms, $p < .05$) was followed by a centro-parietal positivity (244–1,000 ms, peak: 868 ms, $p < .001$); and in semantic violations, an occipito-parietal negativity (98–392 ms, peak: 360 ms, $p < .05$) was followed by a centro-parietal positivity (412–1,000 ms, peak: 842 ms, $p < .001$).

*3.3.3.2. No-Regression trials*: Considering trials in which no first-pass regression occurred at the target noun, neither violation type showed any ERP effects in sentence-medial conditions. In sentence-final conditions, both violation types elicited the same response: a sustained, centro-parietal negativity. In syntactic violations, this was reflected

by a single effect (310–1,000 ms, peak: 586 ms, $p < .001$). In semantic violations, the effect was constituted by two disjoint effects (336–646 ms, peak: 592 ms, $p < .01$; 652–774 ms, peak: 692 ms, $p < .05$) (Fig. 5).

## 4. Discussion

The goal of this study was to establish whether any differences exist between reading studies using the word-by-word presentation method and natural reading. To this end, we compared judgment accuracy and ERP responses in both reading modalities. In the natural reading study, we also performed regression-contingent analyses of judgment accuracy and ERP responses in order to investigate the role of regressive eye movements for sentence comprehension.

The judgment accuracy data provide strong evidence that comprehension is, on average, better when sentences are read naturally than when they are presented word-by-word. Here, we are assuming that judgment accuracy indexes comprehension ability. This assumption is plausible because in order to detect the semantic anomalies the participants had to process for comprehension. Since the comprehenders did not know in advance whether a sentence would contain a semantic, a syntactic, or no violation, they had to process all sentences for comprehension.

The ability to revisit earlier material available in natural reading seems to be the key to explaining the difference in judgment accuracy between word-by-word and natural

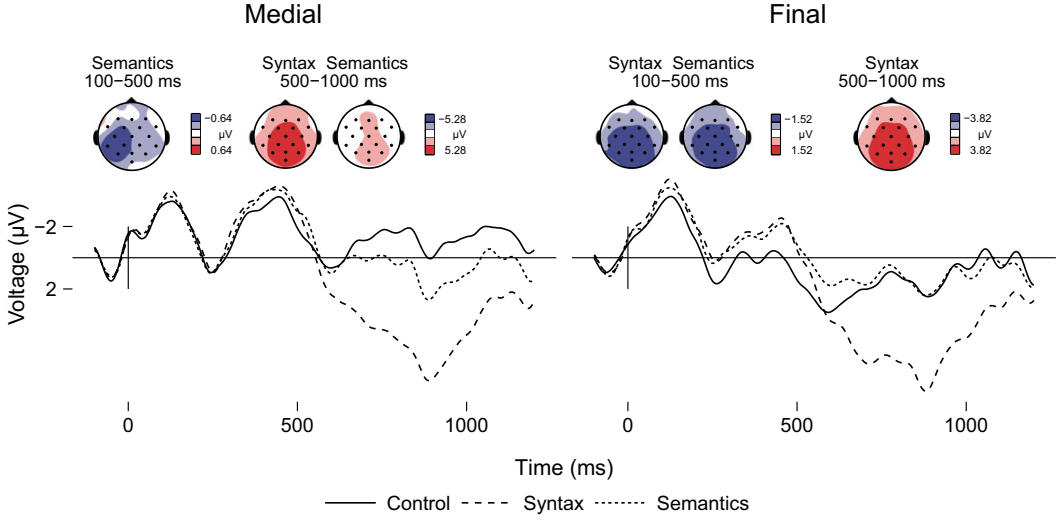


Fig. 5. Average ERPs from word-by-word sessions at electrode Pz (low-pass-filtered at 10 Hz, negativity plotted upwards) for control sentences (solid line), syntactic violations (dashed line), and semantic violations (dotted line) at the target noun in sentence-medial and sentence-final position. Bicubic spline-interpolated isovoltage maps above the waveforms show the topographic distributions of mean amplitude differences (violation minus control) from 100 to 500 ms (N400) and from 500 to 1,000 ms (P600).

reading. If the eyes made a regressive saccade, comprehension improved substantially. If the eyes did not make a regressive saccade, accuracy in the violation conditions was as low or lower than that observed during word-by-word presentation. The latter result suggests that the freedom to control viewing times alone does not improve comprehension in these sentences. In the current setting, the added benefit of natural reading really is the ability to revisit earlier material.

Regarding the ERP responses in word-by-word presentation and in natural reading, as Table 5 shows, these were remarkably similar in the two modalities. This shows that results from word-by-word reading do largely generalize to free reading. However, there were also some interesting differences demonstrating that coregistration gives us a more comprehensive picture: First, like earlier studies (Dimigen et al., 2011; Kretzschmar et al., 2009), we found that ERP effects appeared earlier in free reading than in word-by-word reading. This has been explained in terms of orthographic preactivation through parafoveal preview. Lexical processing basically gets a head start in free reading because visual information about a word enters the system even before the word is first fixated. One important consequence is that researchers have to be careful when they compare the time-course of effects in RSVP studies to effects found in other experimental paradigms.

Second, for sentence-final semantic violation, word-by-word presentation showed an N400 effect, whereas the natural reading study showed both N400- and P600-effects. Originally, the P600 effects have been associated with syntactic processing (hence the alternate name syntactic positive shift). However, a number of recent psycholinguistic studies suggest that P600 effects may not be specific to syntactic processing (DeLong et al., 2014; Kim & Osterhout, 2005; Van De Meerendonk et al., 2009) but may also be triggered by semantic mismatches. Our results are consistent with this (we return to this issue below). Further, our analysis also showed N400 effects for both semantic violations and syntactic violations, suggesting that these effects aren't domain-specific either.[6]

Third, in natural reading, a sustained negativity was found in both syntactic and semantic violations but only in sentence-final position and only in trials without regression. These effects clearly demonstrate the added benefit of coregistration. In the grand mean these negativities were cancelled out by the stronger P600 effects (see Fig. 4), but the eye-tracking signal allowed us to separate them out.

The fact that trials with regressions show P600 effects but trials without show sustained negativities suggests that the comprehension system can respond in at least two qualitatively different ways to the violations. The question is then what are the processing mechanisms that give rise to the two types of effects?

Regarding the sustained negativities, these effects appeared in sentence-final but not in sentence-medial position. This is consistent with our prediction that sharper expectations in sentence-final position should elicit stronger effects. However, the precise nature of these effects is unclear. Sustained negativities have been found to index specific processing demands imposed on the system to solve a task. In many studies, it has been observed that the amplitude of such slow negativities correlates with the amount of load and their

topography with specific resources that are invoked by a task (e.g., during retrieval of spatial, facial, or linguistic long-term memories or during handling distinct—spatial vs. linguistic—working memory tasks; see Khader et al., 2007; Rolke, Heil, Hennighausen, Häussler, & Rösler, 2000; Rösler, Heil, & Röder, 1997). In the present study, the sustained negativity was present only after violations in sentence-final position and when no regression was initiated and its amplitude was much more pronounced in the syntactic than the semantic condition. Taking the negativity as a sign of working memory resource mobilization, the pattern of results suggests the following: First, it was harder for the system to wrap-up non-matching syntactic pieces of information than to interpret a semantic anomaly. Second, when a mid-sentence violation was read without a regression, no sustained negativity was observed. This may be taken as evidence that either no anomaly was detected, or, if it was detected, the wrap-up process was suspended to a later stage, possibly to the end of the sentence. However, when we tested for such negativities in the final region of sentences, we found no effects.[7] At this point, we may be tempted to conclude that readers did not detect sentence-medial violations in trials without regressions, but this conclusion cannot be reconciled with the relatively high judgment accuracy in these trials (87% correct). The explanation for these differences between no-regression trials with sentence-medial and sentence-final violations therefore remains an open problem for the time being.

We found strong evidence suggesting that regressions and P600 effects are linked in free reading. This observation is consistent with results from a corpus study showing that backward-directed eye movements in reading are accompanied by a "P600-like" effect if these regressions landed two or more words back (Dimigen et al., 2007). However, as mentioned earlier, the stimuli in that study were not linguistically controlled, which complicated the interpretation. Our data show that P600 effects and regressions occur jointly not only in response to syntactic violations but also in response to semantic violations. This indicates that a P600 effect is not specifically related to recovery processes handling syntactic mismatches. Instead, the hidden process seems to be of a more general nature. This is in line with recent findings and theoretical accounts which suggest that the P600 effect is not a language-specific and, least of all, syntax-specific phenomenon. Rather, it is argued that it belongs to the family of P3 or P300 effects observed also in a wide range of paradigms with non-linguistic material (Sassenhagen & Bornkessel-Schlesewsky, 2015; Van De Meerendonk et al., 2009).

The defining characteristics of the P3 are positive polarity with a peak at Pz around 300 ms or later, a duration of several hundred milliseconds and a widespread topography extending from left to right superior temporal and from occipital to central sites. The antecedent condition that seems to be common to all situations in which such a positivity is observed is a strong conflict between currently held expectations and what is actually perceived. The functional significance of such a mismatch between expectations and reality can be seen quite differently depending on the specific task demands and stimulus domains involved. Consequently, over the years rather different interpretations have been proposed, such as context updating (Donchin, 1981), stimulus evaluation (Duncan-Johnson, 1981), reactivating well-established stimulus response

connections (Verleger, Baur, Metzner, & Śmigasiewicz, 2014), decision making on what to do next (Hillyard & Kutas, 1983), or initiating executive control (Rösler, 1983).

A more general idea is that of Nieuwenhuis, Aston-Jones, and Cohen (2005), which, by and large, subsumes all these accounts and also those motivated by language studies (see Friederici, 2011; Hagoort, Brown, & Groothusen, 1993, for linguistic interpretations of the P600 effects). In brief, the theory of Nieuwenhuis et al. (2005) says that the P3/P600 component reflects an norepinephrine-induced phasic enhancement of neural responsivity (gain) in the neocortex which is triggered by the outcome of a task-relevant decision process to guide the response. Van de Meerendonk and colleagues translated this very general idea to the P600 findings reported for the linguistic domain and argued that the P3/P600 accompanies the detection of a conflict that may result from any aspect of the input, as semantic, syntactic, orthographic and phonological expectation mismatches which then trigger reanalysis. They hypothesize that "as a result of the reanalysis, it becomes clear [to the system] that no processing error occurred and the perceived error was indeed present. In addition, ... depending on the type of error, the reanalysis can be focused on certain aspects of the stimulus." With the last statement, the authors concede that the topography of the P3 may slightly vary with the specific task demands, and that there can be monophasic P600-effects as well as biphasic N400-P600 effects. A monophasic effect seems to arise when an expectation conflict occurred but the sentence meaning is immediately reconstructable from a reordering or patching of the elements while a biphasic N400-P600 seems to indicate that a prediction error occurred but that a new interpretation of the meaning of the phrase has to be computed by some additional processing effort, which, among others, may induce an activation and binding of additional long-term memory representations. As outlined by Kutas and Federmeier (2011), activation and binding of long-term memory representations is the most likely correlate of genuine, that is, monophasic N400 effects.

This monitoring account also nicely covers the results of the present study. It suggests that whenever a prediction conflict was detected in the natural reading situation in the middle of a sentence (indicated by the P600) a regression was triggered in order to sample additional evidence. When such a conflict was detected at a sentence final position both semantic and syntactic violations triggered a biphasic N400/P600 effect accompanied by a regression. This suggests that, on the one hand, the P600 at the end of the sentence has the same functional property as in the middle of a sentence, that is, it initiates a regression, but that on the other hand, at the end of the sentence it also triggers an immediate integration/reanalysis process (N400).

Considering all three effect patterns, the sustained negativity in no-regression trials with sentence-final violations, the phasic P600 in regression trials following mid-sentence violations, and the phasic N400-P600 in regression trials following sentence-final violations, suggests that the positivity is most likely not indicating the reanalysis effect per se but that it more likely indicates the detection of a prediction error, which then triggers different action strategies; among others this can be a regression or an immediate attempt to solve the conflict. The N400 and the sustained negativity, on the other hand, seems to be manifestations of additional mental processes in order to solve the

conflict, that is, memory search and binding which are invoked to resolve a perceived ambiguity.

One might argue that, according to this model, the positivity as a sign of prediction error detection should always precede the negativity which is seen as a manifestation of reanalysis. However, things might be more complicated. First of all, it must not necessarily be the case that ERP components are always clearly separated in time, that is, that each effect can be seen as a clearly distinct component. ERP effects originating from different neural generators can show different degrees of overlap (Rösler, 1983). Depending on the exact latencies and temporal extensions of the underlying effects, an overlap in the surface EEG can be partial such that distinct peaks are still discernible, or complete, such that one effect is more or less fully "swallowed" by the other. Moreover, as we do not know and cannot see the exact beginning of an effect (the onset of the underlying generator activity) by observing surface potentials, it is possible that a process whose peak is seen later in the ongoing EEG started even earlier than another process which appears with an earlier peak. These timing and overlap relations depend on the location, temporal and spatial extension, and the direction of the dipole moment of the generators. Thus, it could be that the negativity originating from cortical generators peaks earlier over the scalp than a positivity which is originating from subcortical generators and which may also bear effects on the autonomous nervous system (Rösler, Hasselmann, & Sojka, 1987). Moreover, it is conceivable that in some cases the one or the other effect is "swallowed" completely by its counterpart; for example, the pronounced and prolonged negativity at sentence final position in no-regression trials may have "swallowed" a smaller positivity.

Based on the above considerations, it appears that the sentence processing system may resort to at least two different strategies when a word is encountered that does not match the expectations. The system can back off and actively recruit additional information that may help to resolve the problem (reflected by regressions and P600 effects) or alternatively it can plod on, tolerating the inconsistency for the time being (reflected by sustained negativities and the absence of regressions). In the first case, the recruited information will raise confidence about the status of the critical word, and this will lead to increased performance in the judgment task. In the second case ("plodding on"), there remains higher uncertainty about whether or not there is a violation and that will lead to comparatively lower performance.

This view is consistent with two studies by von der Malsburg and Vasishth (2011, 2013) who argued that readers orchestrate a number of sentence processing strategies when reading garden-path sentences. These strategies differ with respect to the depth of processing and von der Malsburg and Vasishth showed that working memory constraints may be one factor determining which strategy is adopted. This proposal is in spirit similar to the "good-enough" account of sentence processing which assumes that the sentence processing system does not necessarily aim for a complete and fully consistent interpretation if there is a chance that a slightly deficient interpretation will do (Ferreira, 2003; Ferreira & Patson, 2007; Ferreira, Christianson, & Hollingworth, 2001; Swets, Desmet, Clifton, & Ferreira, 2008).

## 5. Conclusions

Consistent with prior research (e.g., Schotter et al., 2014), our results show that the ability to revisit earlier material is crucial for thorough reading comprehension. However, our study goes beyond that by systematically investigating the neural correlates of regressions in order to better understand their precise function. Our results show that regressions are strongly associated with the well-known P600 effect. This finding constrains the interpretation of both regressions and the P600 effect and thereby links two large bodies of research. One interpretation of this finding is that regressions reflect the parser's attempt to explore alternative interpretations in response to words that do not match built-up expectations.

When readers did not regress, we found a qualitatively different ERP response, suggesting that an alternative processing strategy was pursued. The sustained centro-parietal negativities observed in these cases may reflect a strategy under which the parser tolerates a deficient interpretation, which may allow merely "good enough" comprehension. More research is needed to better understand this alternative processing strategy and how the parser orchestrates the in-depth and good-enough strategies.

Finally, our results demonstrate that linking behavioral and electrophysiological measures of reading behavior results in a much richer and less ambiguous picture of the processes underlying our ability to read. Coregistration therefore has the potential to considerably aid the development of theories of reading and sentence processing.

## Acknowledgments

Paul Metzner and Titus von der Malsburg were funded through a grant from the German Research Foundation (DFG) awarded to Shravan Vasishth and Frank Rösler within the DFG Research Group 868, Mind and Brain Dynamics. Titus von der Malsburg was also supported through a Feodor Lynen Research Fellowship awarded to him by the Alexander von Humboldt Foundation and through NIH grant HD065829 awarded to Roger Levy and Keith Rayner. The original experimental design by Peter Hagoort was jointly adapted to German with modifications by all authors. Paul Metzner was responsible for the preparation of the experiment, data collection, analyses, and preparation of the plots. Titus von der Malsburg devised the regression-contingent analyses and wrote the software for fixation detection and the randomization test used for the ERP analysis. The manuscript was jointly prepared by all authors.

## Notes

1. We will use the labels N400, P600, and N400-P600 to refer to these effects, that is, deflections with the indicated peak latencies and polarities relative to ERP onset and the control condition.

2. Von der Malsburg and Angele (2016) demonstrated that analyses of multiple dependent eye-tracking measures need to be corrected for multiple testing. In the present case, this is not necessary because the purpose of the eye-tracking analysis merely was to establish that effects found in earlier research (e.g., Braze et al., 2002) were also present in our data.
3. As mentioned earlier, effects with absolute $t$ or $z$ scores above 1.96 were considered significant. Effects with absolute $t$ or $z$ scores between 1.64 and 1.96 were considered marginally significant. These conventional criteria approximate significance at the levels $\alpha = 0.05$ and 0.1.
4. Fig. 3 suggests that the increase in regressions as a function of sentence position was the same for control sentences and sentence with a syntactic anomaly. However, that is only true on the percentage scale. The scale that matters and on which the analysis was carried out is the logit, and on that scale the difference between medial and final regression rates is larger for control sentences.
5. It may be surprising that this N400 effect was reflected in four separate effects, but this is a consequence of the clustering method that was used for the data analysis. The test statistic is computed for each sample; that is, once for every 2 milliseconds of the total epoch. If one of these samples has a below-threshold difference between conditions, that is enough to split a larger effect in two pieces. This, however, is just a consequence of the method and should not stop us from interpreting the results as one prolonged effect as long as there are no a priori reasons to assume that these separate regions have distinct functional meanings.
6. Note that Hagoort (2003) also found a syntactic N400 effect in sentence final position.
7. This analysis followed a similar analysis reported in Hagoort (2003). In the interest of brevity, the details of this analysis are not included in the present article.

## References

Baccino, T., & Manunta, Y. (2005). Eye-fixation-related potentials: Insight into parafoveal processing. *Journal of Psychophysiology*, *19*(3), 204–215.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. Available at http://arxiv.org/abs/1506.04967. Accessed June 16, 2015.

Braze, D., Shankweiler, D., Ni, W., & Palumbo, L. C. (2002). Readers' eye movements distinguish anomalies of form and content. *Journal of Psycholinguistic Research*, *31*(1), 25–44.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. V. Gompel (Ed.), *Eye movements: A window on mind and brain* (Chap. 15, pp. 341–374). Amsterdam, Netherlands: Elsevier Science Ltd.

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150–162.

Dimigen, O., Sommer, W., & Kliegl, R. (2007). Long reading regressions are accompanied by a P600-like brain potential. In R. Kliegl & R. Engbert (Eds.), *Proceedings of the 14th European Conference on Eye Movements (ECEM) (p. 63)*. Potsdam, Germany: University of Potsdam, Journal of Eye Movement Research.

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., & Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: Analyses and review. *Journal of Experimental Psychology: General*, *140*(4), 552–572.

Donchin, E. (1981). Surprise!... surprise? *Psychophysiology*, *18*(5), 493–513.

Duncan-Johnson, C. C. (1981). P300 latency: A new metric of information processing. *Psychophysiology*, *18* (3), 207–215.

Engbert, R., & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, *43*(9), 1035–1045.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*(2), 164–203.

Ferreira, F., & Patson, N. D. (2007). The "good enough" approach to language comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83.

Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research*, *30*(1), 3–20.

Frazier, L., & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language*, *26*(5), 505–526.

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, *91*(4), 1357–1392.

Godfroid, A., Loewen, S., Jung, S., Park, J.-H., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge. *Studies in Second Language Acquisition*, *37*, 269–297.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes*, *25*(2), 149–188.

Greiner, B. (2004). An online recruitment system for economic experiments. In K. Kremer & V. Macho (Eds.), *Forschung und wissenschaftliches Rechnen 2003* (Vol. 63, pp. 79–93). Göttingen, Germany: Gesellschaft für wissenschaftliche Datenverarbeitung.

Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *Journal of Cognitive Neuroscience*, *15*(6), 883–899.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439–483.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438–441.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In A. Kehler, L. Levin, & D. Marcu (Eds.), *Proceedings of NAACL 2001* (pp. 1–8). Pittsburgh, PA: Association for Computational Linguistics.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB —eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, *62*(1), 10–20.

Hillyard, S. A., & Kutas, M. (1983). Electrophysiology of cognitive processing. *Annual Review of Psychology*, *34*, 33–61.

Hopf, J.-M., Bader, M., Meng, M., & Bayer, J. (2003). Is human sentence parsing serial or parallel? Evidence from event-related brain potentials. *Cognitive Brain Research*, *15*(2), 165–177.

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, *37*(2), 163–178.

Khader, P., Knoth, K., Burke, M., Ranganath, C., Bien, S., & Rösler, F. (2007). Topography and dynamics of associative long-term memory retrieval in humans. *Journal of Cognitive Neuroscience*, *19*(3), 493–512.

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, *52*(2), 205–225.

Kretzschmar, F., Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2009). Parafoveal versus foveal N400s dissociate spreading activation from contextual fit. *NeuroReport*, *20*(18), 1613–1618.

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

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*(1), 621–647.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Makeig, S., Bell, A. J., Jung, T.-P., & Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems*, *8*, 145–151.

von der Malsburg, T. (2015). Saccades: An R package for detecting fixations in raw eye tracking data. Available at https://cran.r-project.org/web/packages/saccades. Accessed March 18, 2015.

von der Malsburg, T., & Angele, B. (2016). False positives and other statistical errors in standard analyses of eye movements in reading. Manuscript under revision. Available at http://arxiv.org/abs/1504.06896. Accessed February 2, 2016.

von der Malsburg, T., & Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language, 65*(2), 109–127.

von der Malsburg, T., & Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language and Cognitive Processes*, *28*(10), 1545–1578.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

Matuschek, H., Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Balancing type I error and power in linear mixed models. Available at http://arxiv.org/abs/1511.01864. Accessed November 5, 2015.

Metzner, P., von der Malsburg, T., Vasishth, S., & Rösler, F. (2015). Brain responses to world-knowledge violations: A comparison of stimulus- and fixation-triggered event-related potentials and neural oscillations. *Journal of Cognitive Neuroscience*, *27*(5), 1017–1028.

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus-norepinephrine system. *Psychological Bulletin*, *131*(4), 510–532.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806.

R Development Core Team. (2009). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Rolke, B., Heil, M., Hennighausen, E., Häussler, C., & Rösler, F. (2000). Topography of brain electrical activity dissociates the sequential order transformation of verbal versus spatial information in humans. *Neuroscience Letters*, *282*(1–2), 81–84.

Rosenfeld, A., & Pfaltz, J. L. (1966). Sequential operations in digital picture processing. *Journal of the ACM*, *13*(4), 471–494.

Rösler, F. (1983). Endogenous ERPs and cognition: Probes, prospects, and pitfalls in matching pieces of the mind-body puzzle. In A. W. Gaillard & W. Ritter (Eds.), *Tutorials in event related potential research: Endogenous components* (Vol. 10, pp. 9–35). Advances in Psychology. North-Holland.

Rösler, F., Hasselmann, D., & Sojka, B. (1987). Central and peripheral correlates of orienting and habituation. In R. J. Johnson, J. Rohrbaugh, & R. Parasuraman (Eds.), *EEG supplement: Vol. 40. Current trends in event-related potential research* (pp. 366–372). Amsterdam: Elsevier Science Publishers.

Rösler, F., Heil, M., & Röder, B. (1997). Slow negative brain potentials as reflections of specific modular resources of cognition. *Biological Psychology*, *45*(1–3), 109–141.

Sassenhagen, J., & Bornkessel-Schlesewsky, I. (2015). The P600 as a correlate of ventral attention network reorientation. *Cortex*, *66*, A3–A20.

Schotter, E. R., Tran, R., & Rayner, K. (2014). Don't believe what you read (only once): Comprehension is supported by regressions during reading. *Psychological Science*, *25*(6), 1218–1226.

Starr, M. & Inhoff, A. (2004). Attention allocation to the right and left of a fixated word: Use of orthographic information from multiple words during reading. *European Journal of Cognitive Psychology*, *16*(1–2), 203–225.

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, *36*(1), 201–216.

Van De Meerendonk, N., Kolk, H. H., Chwilla, D. J., & Vissers, C. T. (2009). Monitoring in language perception. *Language and Linguistics Compass*, *3*(5), 1211–1224.

Verleger, R., Baur, N., Metzner, M. F., & Śmigasiewicz, K. (2014). The hard oddball: Effects of difficult response selection on stimulus-related P3 and on response-related negative potentials. *Psychophysiology*, *51*(11), 1089–1100.

White, S. J. (2008). Eye movement control during reading: Effects of word frequency and orthographic familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 205–223.

## Appendix : ERP waveform plots with multiple electrodes

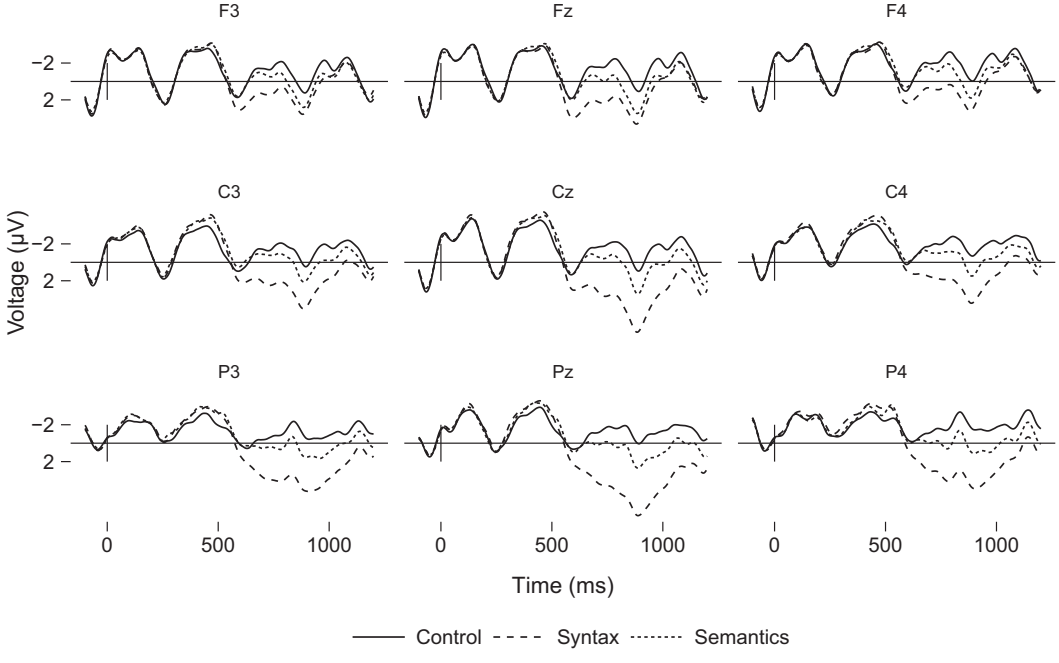


Fig. A1. Word-by-word presentation, sentence-medial violation.

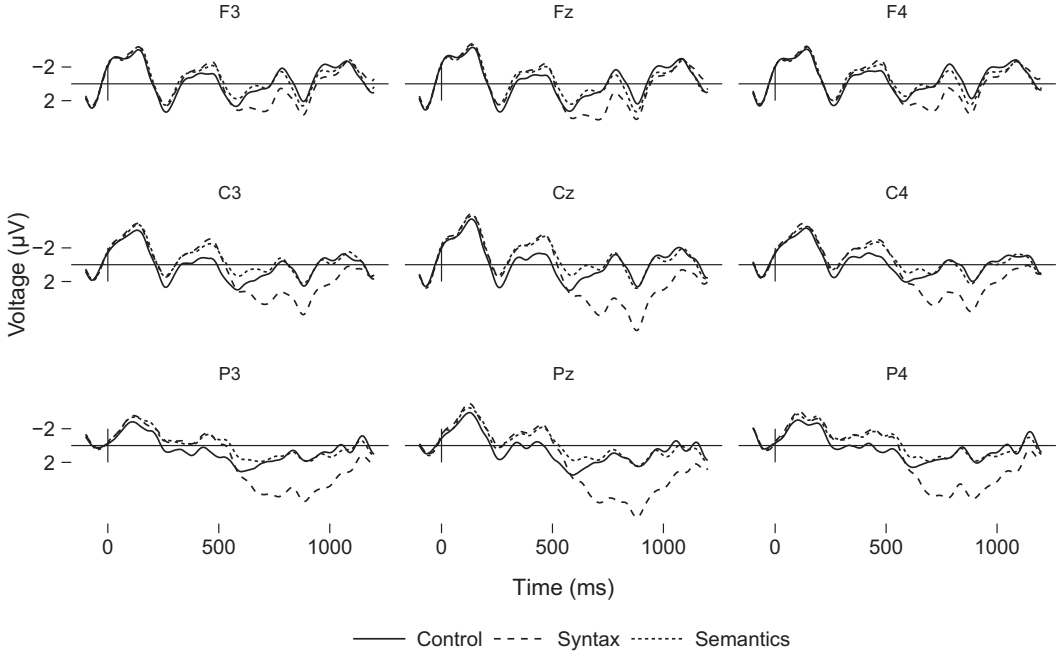


Fig. A2. Word-by-word presentation, sentence-final violation.

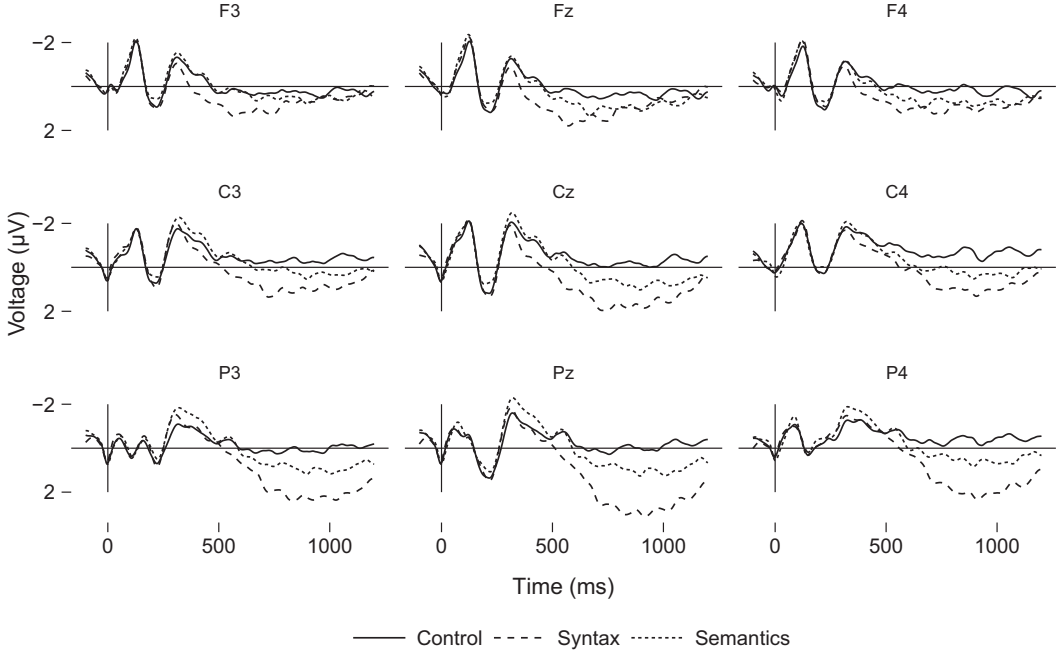


Fig. A3. Natural reading, sentence-medial violation.

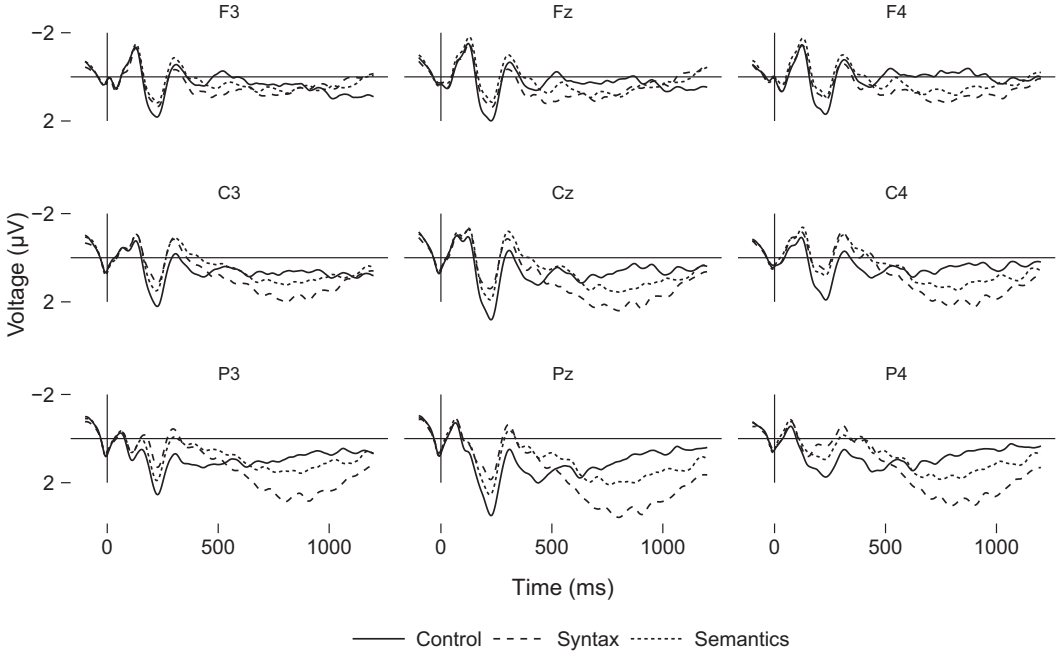


Fig. A4. Natural reading, sentence-final violation.

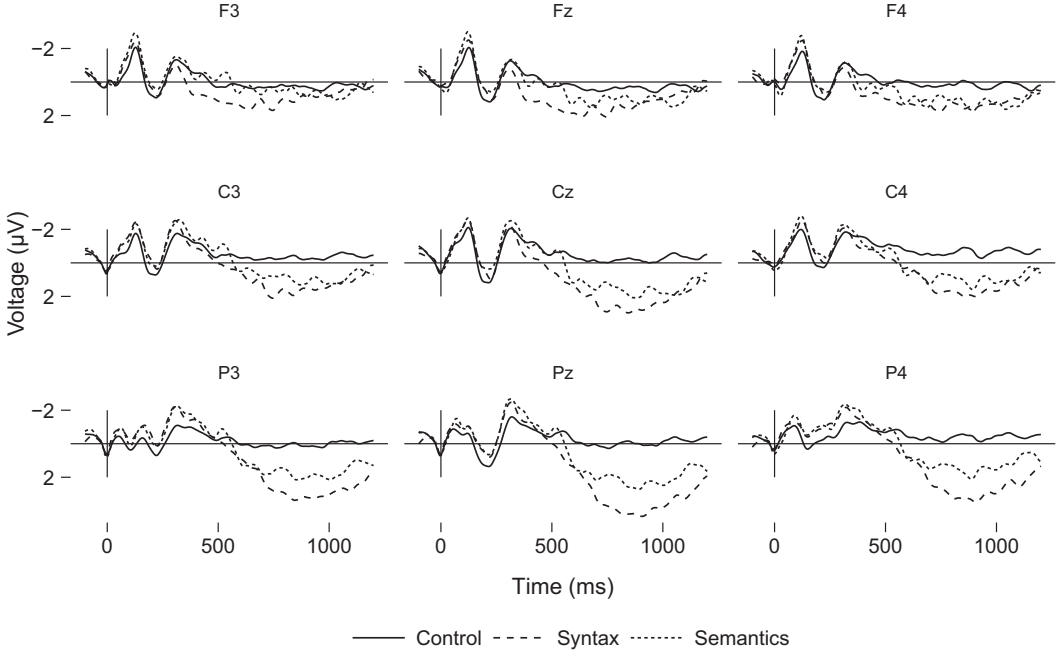


Fig. A5. Natural reading, regression, sentence-medial violation.

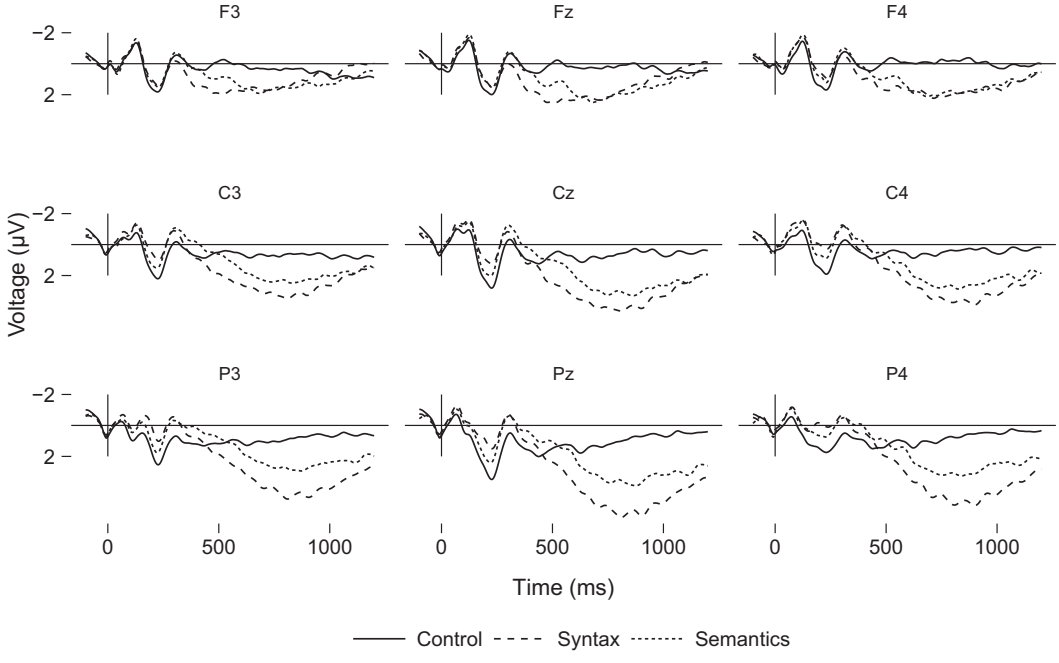


Fig. A6. Natural reading, regression, sentence-final violation.

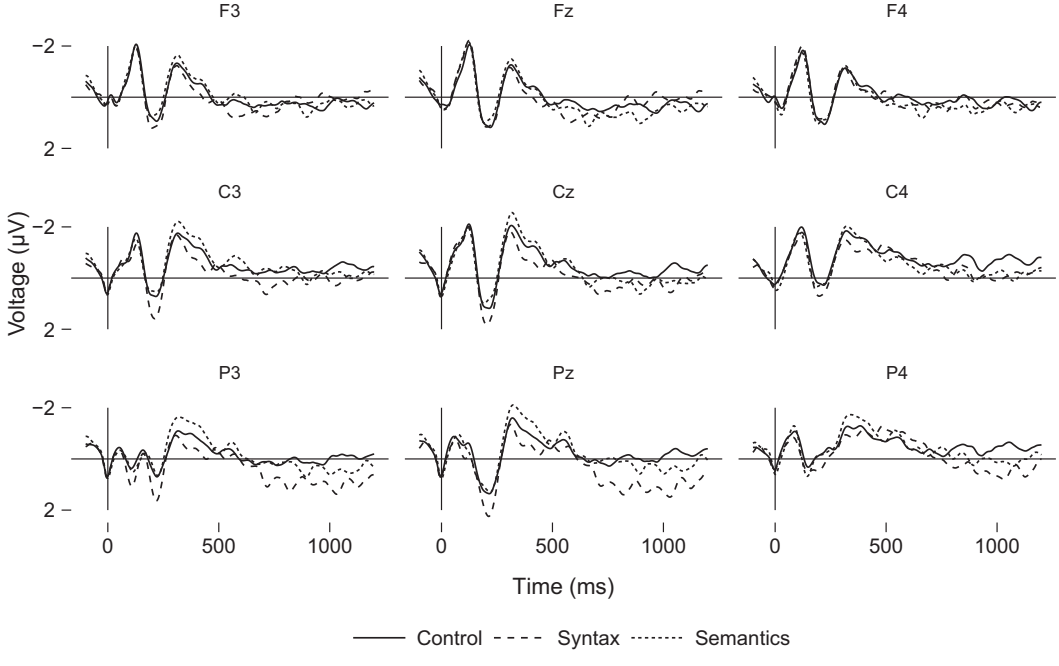


Fig. A7. Natural reading, no regression, sentence-medial violation.

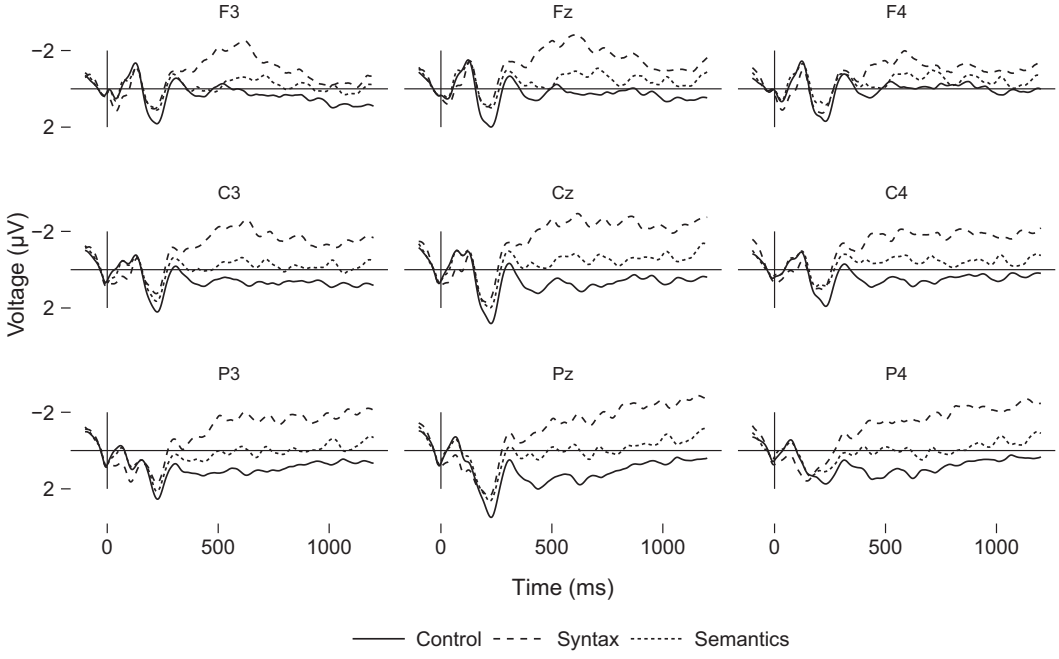


Fig. A8. Natural reading, no regression, sentence-final violation.