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

Paul Metzner
Department of Linguistics, University of Potsdam, Karl-Liebknecht-Straße 24–25, 14476 Potsdam, Germany

Titus von der Malsburg
Department of Psychology and Department of Linguistics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0109, USA

Shravan Vasishth
Department of Linguistics, University of Potsdam, Karl-Liebknecht-Straße 24–25, 14476 Potsdam, Germany

Frank Rösler
Department of Biological Psychology and Neuropsychology, University of Hamburg, Von-Melle-Park 11, 20146 Hamburg, Germany

**Corresponding Author:** Titus von der Malsburg, University of California, San Diego, Department of Psychology, 9500 Gilman Drive, La Jolla, CA 92093-0109

E-mail: malsburg@ucsd.edu
Abstract

How important is the ability to freely control eye movements for reading comprehension? We investigated this question using event-related brain potentials recorded while participants read either word-by-word (also known as RSVP) or naturally. Additionally, eye movements were recorded concurrently with brain potentials during natural reading. Word-by-word presentation and natural reading both elicited similar N400 and P600 effects in response to syntactic and semantic violations. However, comprehension accuracy was higher in natural reading than in word-by-word presentation and particularly high when participants regressed to earlier portions of the sentence after encountering the violation. A more fine-grained ERP analysis showed that P600 effects, which are believed to reflect recovery processes, only occurred in trials with regressive eye movements. In trials without regressions, we instead found either a sustained, centro-parietal negativity starting at around 320 ms post-onset or no effect depending on the position of the violation within the sentence. Thus, the combined analysis of eye movements and ERPs reveals that the sentence comprehension system engages in strategic choices when confronted with difficult material and that the ability to reread earlier parts of the sentence is the key to thorough comprehension.

Keywords: ERP; eye movements; reading; sentence comprehension; regressions
The Importance of Reading Naturally: Evidence from Combined Recordings of Eye Movements and Electric Brain Potentials

**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 can look at every word for as long as we wish, and we are not forced to read the words sequentially. Eyetracking 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 more frequently trigger leftward eye movements (regressions). If a word is easy to process, or when it can 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. An example is auto-paced word-by-word presentation (otherwise known as rapid serial visual presentation). In this form of reading, one word is shown at a time and each word is presented for a fixed duration. For example, word-by-word presentation has been used in research investigating event-related brain potentials (ERPs) during reading (Kutas & Hillyard, 1980; Hagoort, Hald, Bastiaansen, & Petersson, 2004; Osterhout & Holcomb, 1992). The motivation for using this presentation modality is that electric potentials generated by eye movements in natural reading would contaminate the recordings of the electroencephalogram (EEG).

One assumption made by experiments using word-by-word presentation is that, despite the highly constrained form of reading, comprehension accuracy is largely unaffected. This is a reasonable assumption because the rate of speech and the order or words during listening comprehension is also beyond the comprehender’s control.

However, word-by-word reading takes away the ability to make regressive eye movements, which are crucial markers of processing difficulty in sentence comprehension (Clifton, Staub, & Rayner, 2007). Indeed, a recent study has shown that interfering with the reader’s ability to pick up visual information during regressions causes comprehension accuracy to fall (Schotter, Tran,
& Rayner, 2014). Schotter and colleagues masked words immediately after they were read for the first time. This masking led to lower comprehension accuracy compared to a natural reading condition.

So what is the difference between word-by-word versus natural reading with respect to comprehension accuracy and the ERP response? In natural reading, is there a difference in the ERP response when a regression occurs versus when it does not? We address these questions by directly comparing sentence comprehension difficulty during word-by-word presentation and in natural reading while recording EEG signals. We show that comprehension improves in natural reading compared to word-by-word presentation, and that regressive eye movements reveal the strategic choices made by the comprehension system. Thus, natural reading furnishes important information about sentence comprehension processes that cannot be uncovered with word-by-word presentation.

**Modulating sentence comprehension difficulty**

In order to systematically modulate comprehension difficulty, we adapted a design by Hagoort (2003): We had participants read German sentences containing words that violated either grammar (syntax) or common world-knowledge (semantics). We included both syntactic and semantic violations because both have been studied extensively in psycholinguistic research and their ERP correlates are relatively well understood.

An example of a syntactic violation in German 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 human sentence comprehension system can react in one of several ways. One option is to initiate a recovery attempt by registering the gender mismatch and rejecting the resulting structure as ungrammatical; such a syntactic recovery attempt generally elicits a P600 effect (e.g., Osterhout & Holcomb, 1992). Alternatively, no recovery may be initiated, either because the violation is not detected (due to lack of attention,
etc.), or because the comprehension system builds a partially well-formed representation, treating the sentence with a violation as “good enough” (Ferreira & Patson, 2007).

A world-knowledge (semantic) violation can be triggered by using an adjective such as *neugierig*, (‘inquisitive’). This word raises an expectation for a noun representing an animate referent; surprise should result if the next word is instead an inanimate-referring noun such as *Bauernhof* (‘farm’). As in the syntactic violation, this type of violation should either result in the recognition of an anomaly, or in mistakenly treating the adjective-noun combination as an acceptable collocation. Such semantic anomalies are known to trigger an N400 effect (e.g., Kutas & Hillyard, 1980). Hereafter, when we write N400 or P600, we mean an N400 effect, and a P600 effect.

Thus, both syntactic and semantic violations generally trigger recovery processes. Such processes are part of a broader class of recovery process in sentence comprehension, a prominent example of which is triggered by “garden-path” sentences. Given a sentence fragment such as *The lawyer examined…*, native speakers of English expect a sentence in which the lawyer is doing the examining and most readers will have a strong expectation that the next constituent will be the object of *examined*, as in *The lawyer examined the evidence*. However, the sentence could also continue with *…by the nurse was ill*, i.e., 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 eyetracking studies (Frazier & Rayner, 1982; Braze, Shankweiler, Ni, & Palumbo, 2002; Clifton et al., 2007); and with a centro-parietal negativity (Hopf, Bader, Meng, & Bayer, 2003), and/or the P600 effect (Osterhout & Holcomb, 1992; Gouvea, Phillips, Kazanina, & Poeppel, 2010) in ERP studies.

A close correspondence has been observed between the recovery process in garden-path sentences and outright ungrammatical structures: 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 a P600. 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 a recognition of ungrammaticality/anomaly. Nevertheless, as Gouvea et al. (2010) and Hopf et al. (2003), have demonstrated, in the first moments of detecting the violation/anomaly, the recovery process has an ERP response similar to that of the garden path.

Given these considerations, syntactic violations and semantic anomalies are a good choice for investigating recovery processes in reading. They constitute a simple experiment manipulation, are well-studied in the ERP and eyetracking literature, and trigger a class of recovery process that is of great importance in sentence comprehension research. Accordingly, in our experiment, we had participants read sentences with words that were inconsistent with their currently maintained expectation. Violating this expectation was intended to either trigger a recovery process that would lead to a recognition of ungrammaticality (syntactic violation) or anomaly (semantic violation), or to lead to a misjudgment of the sentence as grammatical/non-anomalous.

As in the Hagoort (2003) study that inspired our design, we were also interested in the effect of the position of the anomaly within a sentence. When we start to hear or read a sentence, we may be less certain about the identity of the upcoming word or part of speech compared to when we reach the end of the sentence. 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 of expectation 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 is what Hagoort found for the semantic violation in his study; he found a larger N400 effect in sentence-final position, compared to effects seen in the sentence-medial condition. Hagoort suggested that this larger amplitude may have to do with “the strength of the semantic constraints increas[ing] towards the end of the sentence” (p. 894). In other words, with increasing information about the sentence, the prediction regarding the upcoming word becomes sharper. Such a sharpened expectation account would predict higher grammaticality judgment accuracy in sentence-final position. Although Hagoort did not report accuracy as a function of position in his study, it is likely that no such effect was found since the accuracies across conditions were close to 100%. This high accuracy was probably because the experimental task was relatively easy: a single sentence followed by the judgment task.

An alternative possible effect of word position on linguistic violations is that, due to higher certainty towards the end of a sentence, the reader pays less attention to a violation. The 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 be therefore more willing to attend to the sentence; but in later parts they may have formed a much stronger belief, so strong that a violation does not have enough weight to sway that belief. A clear prediction of such an attenuation in attention is reduced accuracy in detecting the violation/anomaly. Since the acceptability judgment accuracies in the Hagoort study were close to 100% for all conditions, there was not much evidence in favor of this position.

The effect of sentence position is therefore a potentially important factor for investigating how and whether the comprehension system detects an anomaly. We therefore included syntactic or semantic violations in a sentence-medial versus sentence-final position. In order to have a baseline, the sentence-medial and sentence-final conditions each had a control sentence that had no violation. In other words, we had a $2 \times 3$ factorial design: position (sentence medial vs final) and violation type (control, syntactic, semantic). The Hagoort study had an additional condition with a combined syntactic and semantic violation. We did not have such a condition in order to avoid complexity in the experiment design, and because, unlike Hagoort, we were not primarily
interested in the joint effect of syntactic and semantic violations, but rather in the differences between recovery processes in word-by-word presentation versus natural reading.

We conducted two experiments using the same items. One was a classical reading study with word-by-word presentation. In the second study, we allowed participants to read the sentences on a computer screen at their own pace, while concurrently recording their electroencephalogram and their eye movements. This approach has been made possible by recent advances in signal processing and computation that enable the removal of eye movement artifacts from EEG recordings (Makeig, Bell, Jung, & Sejnowski, 1996; Dimigen, Sommer, Hohlfeld, Jacobs, & Kliegl, 2011). The analysis focused on known markers of processing difficulty in the EEG (N400 and P600 effects) and on readers’ performance in the task probing comprehension of the text. Having an electrophysiological and behavioral signal allowed us to analyze brain potentials contingent on the reading strategies observed using eyetracking.

Predictions

As discussed above, our main goal was to investigate whether experiment modality matters in reading studies. Either we would see identical effects in word-by-word presentation and natural reading, or there would be theoretically important differences. Given that our study modified a design by Hagoort (2003), it is possible to derive quite specific predictions for the effect of position on syntactic and semantic violations. In sentence-medial position, Hagoort found (inter alia) that semantic violations elicited an N400 effect, and syntactic violations a P600; the amplitude of the P600 in the syntactic violation was larger than in the semantic violation. In sentence-final position, an N400 was seen in both syntactic and semantic violations; in semantic violations, the amplitude of the N400 was larger in sentence-final position than in the sentence-medial position.

Based on Hagoort’s results, in both modalities we expected (relative to the control condition) a P600 for syntactic violations, and an N400 for semantic violations. P600 effects in response to semantic violations have been reported by some authors (e.g., Kim & Osterhout,
2005) but they are believed to occur only under specific linguistic constraints involving semantic role reversal that are not given in our design (see Bornkessel-Schlesewsky & Schlesewsky, 2008, for a discussion). In the natural reading modality, we expected higher regression probability in violations compared to the control, in both sentence-medial and sentence-final position; higher regression probability is a classic marker of recovery processes in eyetracking research (Frazier & Rayner, 1982; Braze et al., 2002). We had no clear predictions about the association between ERP components and regressive eye movements.

We also expected that sentence-medial conditions would show a different pattern of effects than sentence-final conditions, due to the greater predictability of the final word in the latter. Given the Hagoort results, we expected an N400 sentence-finally for both violation types. In addition, in the natural reading we expected an overall higher regression probability sentence-finally than sentence-medially; this is because a higher regression probability is a well-known consequence of sentence-final wrap-up processing (Rayner, Kambe, & Duffy, 2000).

**Materials and methods**

**Participants**

We tested 72 students at the University of Potsdam, Germany (18 male, 54 female, mean 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. We tested twice as many participants in the natural reading condition because we expected a lower signal-to-noise ratio in that condition based on previous studies that were conducted in our lab (Metzner, von der Malsburg, Vasishth, & Rösler, 2014).
Design

German was the target language. There were six conditions: the sentences contained a noun that introduced either a syntactic, a semantic, or no violation that 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$_{\text{fem}}$ deteriorating farm$_{\text{masc}}$”). Sentences with semantic violations had a noun that was incongruent with the preceding adjective given commonsense knowledge (“the$_{\text{masc}}$ inquisitive farm$_{\text{masc}}$”).

To control adjectives for frequency, we created clusters using $k$-means clustering on logarithmic frequency (extracted from the lexical database dlexDB, Heister et al., 2011) and paired only adjectives from the same cluster. Sentence-medial adjectives had an average log-frequency of 1.10 (baseline sentences and syntactic violations: $M = 1.14$, $SE = 0.02$; semantic violations: $M = 1.02$, $SE = 0.02$). In sentence-final violations, the adjective had an average log-frequency of 0.83 (baseline sentences and syntactic violations: $M = 0.86$, $SE = 0.02$; semantic violations: $M = 0.75$, $SE = 0.02$). Adjectives within an item differed by no more than two characters in length (sentence-medial: $M = 9.02$, $SE = 0.16$ in baseline and syntactic violations, $M = 9.04$, $SE = 0.16$ in semantic violations; sentence-final: $M = 9.52$, $SE = 0.15$ in baseline and syntactic violations, $M = 9.52$, $SE = 0.15$ in semantic violations). To avoid an influence of grammatical gender, we balanced the number of male, female, and neutral gender within each position and condition.

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
rereferenced to linked mastoids. Eye movements were recorded with an EyeLink 1000 (SR Research, Mississauga, Ontario, Canada) with a sampling rate of 1000 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).

**Procedure**

Participants sat in a dimly lit, electromagnetically shielded, and sound-insulated booth. The eyes were approximately 60 cm from the presentation screen, which had a diagonal length of 22 in. An experimental session began with ten practice trials to familiarize participants with the procedure. In sessions with word-by-word presentation, each trial began with a fixation dot in the center of the screen. After 1000 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. Each participant read 360 test/follow-up sentence pairs randomly interspersed with 180 similar pairs of distractor sentences. 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 was presented, based on the procedure described by Hagoort (2003), a blank display with a pseudo-randomly varying interval between one and two seconds was shown. This was followed by a judgment task. Also 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 and to respond accordingly with a button press. Following the response, a blank display of 1150 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 px 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.
In both presentation forms, participants were encouraged to take a short break every twenty trials. During these breaks, they received feedback about their performance on the judgment task. After 240, 360, and 480 trials, participants had to take longer breaks to relax their neck muscles and eyes.

**Data preprocessing**

A velocity-based saccade detection algorithm was used to detect saccades and fixations in the raw eyetracking data (Engbert & Kliegl, 2003). Fixations shorter than 20 ms and longer than 1,200 ms were removed. This led to the loss of 0.3% of all fixations. The following eyetracking 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 entered from the left (i.e., during first pass). 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; only regressions during first pass are considered as they are widely assumed to index early processing events (Vasishth, von der Malsburg, & Engelmann, 2012).

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. We then identified muscle artifacts generated by the eye movements during reading using independent components analysis. 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 electrodes was minimized (see Fig. 1). From the data corrected in this manner, we removed epochs with 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 1000 ms later were exported to R (R Core Team, 2013) for further processing. Prior to the statistical analysis, the data were baseline-corrected by subtracting the average amplitude in a 100 ms interval preceding the time-lock.

Analysis

Response accuracy in the judgment task was analyzed with two logistic mixed-effects models as implemented in the R package \textit{lme4} (version 1.1.7, using the BOBYQA algorithm for optimization, Bates, Maechler, Bolker, and Walker, 2014). 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. 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 (sum contrast). Again all possible interactions were included.

The models had a full variance-covariance specification for the random effects (intercepts and slopes for the within-participant fixed effects, with correlation estimated). In cases where a model did not converge, we dropped the random effect with the least variance and refit the model. This procedure was repeated until the model converged. An effect was considered statistically
significant at $\alpha = 0.05$ if the corresponding absolute $t$ or $z$ statistic was greater 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.

ERPs were analyzed with nonparametric 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 1000 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 of this cluster-based randomization approach, 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 a violation had occurred).

Results

Judgment accuracy

Fig. 2 shows 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 investigated the effect of regressions on accuracy. See Table 2 and Table 3 for the parameter estimates from the 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 vs medial position (marginally significant in the model testing presentation modality).

The model testing presentation modality (Table 2) showed that accuracy was significantly improved when the participants read sentences naturally instead of word-by-word. For syntactic violations, a three-way interaction showed a lower accuracy on average when the violation occurred in final position, and when sentences were presented word-by-word.

The model testing the effect of regressions showed that accuracy was higher in sentences with violations when a regression occurred.
Eyetracking data

Fig. 3 shows means and 95% CIs for the eyetracking measures (first fixation duration, gaze duration, regression probability). The parameter estimates from the linear mixed model analysis are in Table 4.

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 occurred. This effect was found in first fixation duration and gaze duration. The gaze duration data also showed that the critical noun was read faster when it was in sentence-final position. An interaction between syntactic violation and position was seen in all measures. In duration data, a syntactic violation led to shorter fixations times sentence-finally compared to sentence-medially. The first-pass regression data showed the opposite pattern to that seen in the fixation durations: a syntactic violation led to higher regression probability sentence-finally compared to sentence-medially. The shorter fixation durations sentence-finally were due to regressive eye movements cutting short the first pass. This association in the sentence-final conditions was established by fitting a linear mixed model that examined the effect of regression (as a binary predictor) on fixation durations; in both first fixations and gaze duration, a regressive eye movement led to significantly shorter durations.

Event-related potentials

All ERP-related results for the word-by-word study, the natural reading study, and the regression-contingent analyses in the natural reading study are summarized in Table 5. We discuss these below. For simplicity, instead of saying “relative negativities and positivities compared to control sentences”, we use the abbreviations negativities and positivities.

Word-by-word presentation. In the word-by-word presentation experiment, we found results similar to those of Hagoort (2003). Syntactic violations in sentence-medial position led to a P600-like centro-parietal positivity from 544 to 1000 ms (peak at 812 ms, \(p < .001\)); in sentence-final position, an N400-like effect from 312 to 526 ms (peak at 428 ms, \(p < .05\)) was
followed by a late centro-parietal positivity from 570 to 1000 ms (peak at 688 ms, \( p < .001 \)).

Semantic violations in sentence-medial position elicited a short-lived centro-parietal negativity from 476 to 520 ms (peak at 504 ms, \( p < .01 \)) and a centro-parietal positivity from 684 to 1000 ms (peak at 836 ms, \( p < .01 \)); in sentence-final position, a sequence of four centro-parietal negativities from 128 to 656 ms together constituted an N400 effect (peaks at 170, 230, 358, 536 ms; all \( p < .001 \)).

Natural reading. In the ERP recorded during natural reading (Fig. 4), syntactic violations in sentence-medial position elicited a P600-like centro-parietal positivity from 478 to 1000 ms (peak at 828 ms, \( p < .001 \)); in sentence-final position, a similar centro-parietal positivity from 590 to 1000 ms (peak at 776 ms, \( p < .001 \)) was preceded by an N400-like occipito-parietal negativity from 130 to 434 ms (peak at 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 from 230 to 372 ms (peak at 268 ms, \( p < .05 \)) and a centro-parietal positivity from 760 to 1000 ms (peak at 980 ms, \( p < .01 \)). Sentence-finally, both effects occurred slightly earlier from 104 to 522 ms (peak at 360 ms, \( p < .001 \)) and from 696 to 988 ms (peak at 842 ms, \( p < .01 \)).

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 higher in violations than control sentences, and higher in final position compared to medial position. To examine the ERP 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 similar statistical analyses on these two subsets that were conducted on the full data-set (see Fig. 4).

Regression trials. Considering trials in which a first-pass regression occurred at the target noun, in sentence-medial conditions both violation types elicited a P600 effect; in syntactic violations, it ranged from 290 to 1000 ms (peak at 828 ms, \( p < .001 \)) and in semantic violations
from 540 to 1000 ms (peak at 848 ms, \( p < .001 \)). In sentence-final position, both violations elicited an N400/P600 response; in syntactic violations, a centro-parietal negativity from 24 to 378 ms (peak at 164 ms, \( p < .05 \)) was followed by a centro-parietal positivity from 244 to 1000 ms (peak at 868 ms, \( p < .001 \)); and in semantic violations, an occipito-parietal negativity from 98 to 392 ms (peak at 360 ms, \( p < .05 \)) was followed by a centro-parietal positivity from 412 to 1000 ms (peak at 842 ms, \( p < .001 \)).

**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, a single effect from 310 to 1000 ms (peak at 586 ms, \( p < .001 \)) represented the sustained negativity. In semantic violations, it comprised two disjoint effects from 336 to 646 ms (peak at 592 ms, \( p < .01 \)) and from 652 to 774 ms (peak at 692 ms, \( p < .05 \)).

**Discussion**

Our main goal was to establish whether any differences exist between reading studies using the word-by-word presentation method and natural reading. Towards this end, we compared judgment accuracy and ERP responses in the two modalities. In the natural reading study, we also performed regression-contingent analyses of judgment accuracy and ERP responses. We discuss the implications of each of these.

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. The ability to revisit earlier material, only available in natural reading, seems to be the key to explaining this difference. If the eyes make a regressive saccade, comprehension improves substantially. If the eyes don’t make a regressive saccade, accuracy in the violation conditions is 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
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, but there were two differences: (i) in the semantic violation which occurred sentence-finally, word-by-word presentation showed an N400 effect, whereas the reading study showed both an N400 and a P600; and (ii) in the natural reading study, in no-regression trials of the sentence-final conditions, a sustained negativity was seen in both syntactic and semantic violations. The P600 in the natural reading study is not too surprising given that P600 effects have been observed in other settings testing semantic violations (Kim & Osterhout, 2005). However, this effects shows that semantic P600 effects are not limited to semantic violations involving thematic role reversal.

Regarding the sustained negativity, at first sight, this might look like a reflex of wrap-up effects that occur at the end of a sentence. But if that were so, we would see a similar sustained negativity in the sentence-final region in the sentence-medial conditions. Instead, in sentence-medial conditions, we see no effects in the sentence-final region when we compare regression and no-regression trials (to conserve space, we do not present these analyses in detail). Thus, it is unlikely that the sustained negativity reflects only wrap-up effects. This sustained negativity, found only in the sentence-final conditions when no regression occurred, suggests that the sentence-final wrap-up process interacts with the recovery process. One possibility is that the sustained negativity reflects the “good-enough” toleration strategy discussed earlier. The weakness of such an explanation is that we should have seen such a sustained negativity in the no-regression trials for the sentence-medial conditions as well. Instead, we see no effect. It is worth investigating in future work, perhaps with higher statistical power, whether a sustained negativity is generated in medial positions as well.

What new insights about the comprehension process could be uncovered using natural reading? The regression-contingent analyses suggest an answer. Syntactic violations did not just elicit stronger P600 effects, they also prompted regressive eye movements more often than semantic violations did. This suggests that regressions may be linked to the P600 effect and thus
to recovery processes. The idea that regressions are associated with P600 effects also receives support from a corpus study showing that backward-directed eye movements in reading are accompanied by a “P600-like” effect that is larger if the regression is longer (Dimigen, Sommer, & Kliegl, 2007). This study used simple sentences suggesting that even when the material is easy to process, misinterpretations may occur and may demand recovery procedures similar to those triggered by our material. In our experiment, P600 effects occurred when the reader carried out a regressive saccade from the mismatching word to earlier words; no P600 effects were found in no-regression trials. Along with the fact that judgment accuracy was lower in no-regression trials, the absence of a P600 suggests that a recovery was either initiated less often in these trials, or not at all. The regression-linked P600 suggests that regressive eye movements support the recovery process in sentence comprehension. When faced with processing difficulty, the sentence comprehension system may carry out a recovery process or alternatively back off to tolerating the violation/anomaly. When a recovery process is started, regressive eye movements improve its probability of success, leading to a correct detection of ungrammaticality; but when the comprehension system backs off to a toleration strategy, no regression tends to occur. Similar to this finding, earlier research on the processing of garden-path sentences also showed that readers engage in strategic choices when faced with words that challenge the currently maintained interpretation. von der Malsburg and Vasishth, 2013 used a novel analysis method to investigate regression scanpaths in response to garden-pathing and found that readers differ strongly in how they orchestrate various reading strategies and that this variability is partly explained by individual differences in working memory resources.

The word-by-word presentation format did not reveal this variability in processing strategies because this format does not provide us with a behavioral signal that would allow us to analyze subsets of the ERP data separately. Thus, the ERP signal was potentially aggregated over two sets of processing strategy: recovery and the toleration strategy. In word-by-word presentation, we also found P600 effects when a violation occurred; that is, in a situation where regressions were impossible, readers may have deployed a “covert reanalysis” strategy (Lewis,
1998). Covert reanalysis refers to a recovery process that is initiated without moving the eyes leftwards. However, this covert recovery was not deployed often in word-by-word reading or it was less efficient, as can be seen from the lower accuracy in the judgment task during word-by-word presentation compared to the natural reading study. Clearly, the toleration strategy was used more often.

Thus, although word-by-word reading and natural reading can yield comparable ERP results (Kretzschmar, Bornkessel-Schlesewsky, and Schlesewsky, 2009; Dimigen et al., 2011; but see Metzner et al., 2014), co-registration using natural reading reveals two qualitatively different reading strategies (attempted recovery vs. toleration of the inconsistency) that were not detected in data obtained using word-by-word presentation alone. Co-registering the EEG signal in addition to fixation data also reveals that one function of regressive eye movement in reading is to support the recovery process when comprehension difficulties arise.
Author notes

This work was funded by a grant from the Deutsche Forschungsgemeinschaft to S. Vasishth and F. Rösler within the DFG Research Group 868, Mind and Brain Dynamics. The original experimental design by Peter Hagoort was jointly adapted to German with modifications by all authors. P. Metzner was responsible for the preparation of the experiment, data collection, analyses, and preparation of the plots. T. 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.
References


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.

<table>
<thead>
<tr>
<th>Sentence-medial:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Der\textsubscript{MASC} verfallene Bauernhof\textsubscript{MASC} braucht eine Renovierung. Er wird von einer Familie bewohnt.</td>
<td>Die\textsubscript{FEM} verfallene Bauernhof\textsubscript{MASC} braucht eine Renovierung. Er wird von einer Familie bewohnt.</td>
<td>(Syn)</td>
</tr>
<tr>
<td>Der\textsubscript{MASC} neugierige Bauernhof\textsubscript{MASC} braucht eine Renovierung. Er wird von einer Familie bewohnt.</td>
<td>(Sem)</td>
<td></td>
</tr>
<tr>
<td>The\textsubscript{MASC}/The\textsubscript{FEM} deteriorating/inquisitive farm\textsubscript{MASC} needs a renovation. It is inhabited by a family.</td>
<td>The\textsubscript{FEM} schwierige Rolle\textsubscript{FEM} - Er überzeugt seine schärfsten Kritiker.</td>
<td>(Syn)</td>
</tr>
<tr>
<td>Der erfahrene Star spielt die\textsubscript{FEM} schwierige Rolle\textsubscript{FEM} - Er überzeugt seine schärfsten Kritiker.</td>
<td>(Sem)</td>
<td></td>
</tr>
<tr>
<td>Der erfahrene Star spielt das\textsubscript{NEUT} schwierige Rolle\textsubscript{FEM} - Er überzeugt seine schärfsten Kritiker.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Die\textsubscript{FEM} elektrische Rolle\textsubscript{FEM} - Er überzeugt seine schärfsten Kritiker.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The experienced star plays the\textsubscript{FEM}/the\textsubscript{NEUT} difficult/electric role\textsubscript{FEM}. He convinces his harshest critics.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 bold face (z).

<table>
<thead>
<tr>
<th></th>
<th>Est.</th>
<th>95% CI</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax</td>
<td>-0.54</td>
<td>[-0.94, -0.14]</td>
<td>-2.6</td>
</tr>
<tr>
<td>Semantics</td>
<td>-1.72</td>
<td>[-2.12, -1.32]</td>
<td>-8.4</td>
</tr>
<tr>
<td>Final Position</td>
<td>-0.31</td>
<td>[-0.67, 0.05 ]</td>
<td>-1.7</td>
</tr>
<tr>
<td>Natural Reading</td>
<td>0.51</td>
<td>[0.14, 0.88  ]</td>
<td>2.7</td>
</tr>
<tr>
<td>Syntax × Final</td>
<td>-0.28</td>
<td>[-0.71, 0.15 ]</td>
<td>-1.3</td>
</tr>
<tr>
<td>Semantics × Final</td>
<td>-0.86</td>
<td>[-1.31, -0.41]</td>
<td>-3.7</td>
</tr>
<tr>
<td>Syntax × Natural</td>
<td>-0.07</td>
<td>[-0.75, 0.61 ]</td>
<td>-0.2</td>
</tr>
<tr>
<td>Semantics × Natural</td>
<td>-0.01</td>
<td>[-0.68, 0.66 ]</td>
<td>0.0</td>
</tr>
<tr>
<td>Final × Natural</td>
<td>-0.29</td>
<td>[-0.74, 0.16 ]</td>
<td>-1.3</td>
</tr>
<tr>
<td>Syntax × Final × Natural</td>
<td>1.87</td>
<td>[1.33, 2.41]</td>
<td>6.8</td>
</tr>
<tr>
<td>Semantics × Final × Natural</td>
<td>0.05</td>
<td>[-0.43, 0.53]</td>
<td>0.2</td>
</tr>
</tbody>
</table>
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 regressions. Estimates (Est.) and 95% confidence intervals (CI) are on the logit scale, statistical significance is indicated by test statistics in bold face (z).

<table>
<thead>
<tr>
<th>Interaction</th>
<th>Est.</th>
<th>95% CI</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax</td>
<td>-0.70</td>
<td>[-1.27, -0.13]</td>
<td>-2.4</td>
</tr>
<tr>
<td>Semantics</td>
<td>-1.86</td>
<td>[-2.42, -1.30]</td>
<td>-6.5</td>
</tr>
<tr>
<td>Final position</td>
<td>0.57</td>
<td>[-1.09, 0.05]</td>
<td>-2.1</td>
</tr>
<tr>
<td>Regression</td>
<td>-0.29</td>
<td>[-0.70, 0.12]</td>
<td>-1.4</td>
</tr>
<tr>
<td>Syntax × Final</td>
<td>0.51</td>
<td>[-0.10, 1.12]</td>
<td>1.6</td>
</tr>
<tr>
<td>Semantics × Final</td>
<td>-0.93</td>
<td>[-1.54, -0.32]</td>
<td>-3.0</td>
</tr>
<tr>
<td>Syntax × Regression</td>
<td>2.00</td>
<td>[1.51, 2.49]</td>
<td>7.9</td>
</tr>
<tr>
<td>Semantics × Regression</td>
<td>0.93</td>
<td>[0.48, 1.38]</td>
<td>4.1</td>
</tr>
<tr>
<td>Final × Regression</td>
<td>-0.27</td>
<td>[-0.92, 0.38]</td>
<td>-0.8</td>
</tr>
<tr>
<td>Syntax × Final × Regression</td>
<td>0.59</td>
<td>[-0.22, 1.40]</td>
<td>1.4</td>
</tr>
<tr>
<td>Semantics × Final × Regression</td>
<td>0.28</td>
<td>[-0.47, 1.03]</td>
<td>0.7</td>
</tr>
</tbody>
</table>
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, position, and regression. 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 bold face ($t$ and $z$).

<table>
<thead>
<tr>
<th></th>
<th>First Fixation Duration</th>
<th>Gaze Duration</th>
<th>Regression Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>95% CI</td>
<td>$t$</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.02</td>
<td>[0.01, 0.03]</td>
<td>2.39</td>
</tr>
<tr>
<td>Semantics</td>
<td>0.06</td>
<td>[0.05, 0.07]</td>
<td>7.16</td>
</tr>
<tr>
<td>Final Position</td>
<td>0.00</td>
<td>[−0.02, 0.02]</td>
<td>−0.14</td>
</tr>
<tr>
<td>Synax × Final</td>
<td>−0.04</td>
<td>[−0.06, −0.02]</td>
<td>−2.60</td>
</tr>
<tr>
<td>Semantics × Final</td>
<td>−0.01</td>
<td>[−0.02, 0.00]</td>
<td>−0.45</td>
</tr>
</tbody>
</table>
Table 5
*Summary of ERP results in the word-by-word reading and natural reading experiments.*

<table>
<thead>
<tr>
<th>Word-by-Word Presentation</th>
<th>Natural Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position</td>
</tr>
<tr>
<td>Medial</td>
<td>P600</td>
</tr>
<tr>
<td>Final</td>
<td>N400/P600</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Regression-Contingent Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
</tr>
<tr>
<td>Yes</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Figure 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.
Figure 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. Accuracy was better when the text was read naturally than when it was read word-by-word. The benefit of natural reading was mainly driven by trials in which the eyes regressed. When they did not regress, the accuracy was similar to that observed during word-by-word reading.
Figure 3. First fixation duration, gaze duration, and regression probability at the noun in sentence-medial and sentence-final items. Color indicates condition with black for control sentences, red for syntactic violations, and blue semantic violations. The bars denote 95\% confidence intervals. Dashed lines show sentence-medial manipulations and solid lines sentence-final manipulations.
different brain responses to syntactic and semantic violations. Separating the trials using the behavioral signal thus revealed two qualitatively sustained negativities that were cancelled out by the stronger P600 effects in the analysis of the full data set. However, the separate analyses for trials with and without regressions show that P600 effects are strongly associated with regressions. In trials without regressions, we instead found sustained negativities that were cancelled out by the stronger P600 effects in the analysis of the full data set. Separating the trials using the behavioral signal thus revealed two qualitatively different brain responses to syntactic and semantic violations.

Figure 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 1000 ms (P600), and from 300 to 1000 ms (sustained negativities). The collapsed data (comprising trials with and without regressions) largely replicate earlier results by Hagoort (2003). However, the separate analyses for trials with and without regressions show that P600 effects are strongly associated with regressions. In trials without regressions, we instead found sustained negativities that were cancelled out by the stronger P600 effects in the analysis of the full data set. Separating the trials using the behavioral signal thus revealed two qualitatively different brain responses to syntactic and semantic violations.
Figure 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 1000 ms (P600).