Neural dynamics of sentiment processing during naturalistic sentence reading

NeuroImage 218 (2020) 116934

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Neural dynamics of sentiment processing during naturalistic sentence reading

Christian Pfeiffer[a,b,*], Nora Hollenstein[c], Ce Zhang[c], Nicolas Langer[a,b,d]

[a] Methods of Plasticity Research Laboratory, Department of Psychology, University of Zurich, Switzerland
[b] University Research Priority Program (URPP) Dynamics of Healthy Aging, Zurich, Switzerland
[c] Department of Computer Science, ETH, Zurich, Switzerland
[d] Neuroscience Center Zurich (ZNZ), Zurich, Switzerland

ABSTRACT

When we read, our eyes move through the text in a series of fixations and high-velocity saccades to extract visual information. This process allows the brain to obtain meaning, e.g., about sentiment, or the emotional valence, expressed in the written text. How exactly the brain extracts the sentiment of single words during naturalistic reading is largely unknown. This is due to the challenges of naturalistic imaging, which has previously led researchers to employ highly controlled, timed word-by-word presentations of custom reading materials that lack ecological validity. Here, we aimed to assess the electrical neural correlates of word sentiment processing during naturalistic reading of English sentences. We used a publicly available dataset of simultaneous electroencephalography (EEG), eye-tracking recordings, and word-level semantic annotations from 7129 words in 400 sentences (Zurich Cognitive Language Processing Corpus; Hollenstein et al., 2018). We computed fixation-related potentials (FRPs), which are evoked electrical responses time-locked to the onset of fixations. A general linear mixed model analysis of FRPs cleaned from visual- and motor-evoked activity showed a topographical difference between the positive and negative sentiment condition in the 224–304 ms interval after fixation onset in left-central and right-posterior electrode clusters. An additional analysis that included word-, phrase-, and sentence-level sentiment predictors showed the same FRP differences for the word-level sentiment, but no additional FRP differences for phrase- and sentence-level sentiment. Furthermore, decoding analysis that classified word sentiment (positive or negative) from sentiment-matched 40-trial average FRPs showed a 0.60 average accuracy (95% confidence interval: [0.58, 0.61]). Control analyses ruled out that these results were based on differences in eye movements or linguistic features other than word sentiment. Our results extend previous research by showing that the emotional valence of lexico-semantic stimuli evoke a fast electrical neural response upon word fixation during naturalistic reading. These results provide an important step to identify the neural processes of lexico-semantic processing in ecologically valid conditions and can serve to improve computer algorithms for natural language processing.

1. Introduction

The written word has fundamentally shaped human cultural and cognitive evolution and still today remains a primary medium for information storage (e.g., Wikipedia) and human communication (e.g., email and social media). Nonetheless, very little is currently known about how human readers extract and process meaning from written text.

The cognitive and neural processes of language processing in humans have been of major interest in cognitive science, neurolinguistics, and neuropsychology (Currie, 1990; Manning et al., 1999; Mason and Just, 2006). Previous research has used custom reading material to study specific aspects of linguistic material (e.g., phonetics, morphology, and semantics) in highly controlled experimental settings in order to collect repeated self-reports, forced-choice ratings, eye-movement data, or electrical or functional neuroimaging data ("The Oxford Handbook of Neurolinguistics," 2019). The neural correlates of reading have traditionally been studied with serial word-by-word presentation with a fixed presentation time, a condition that eliminates important aspects of the normal reading process and precludes direct comparisons between neural activity and oculomotor behavior (Dimigen et al., 2011; Kliegl et al., 2012). However, the electrical neural correlates of naturalistic reading of real sentences has been investigated less frequently due to a number of challenges related to identifying the exact timing and type of visual stimuli presented during reading, as well as the contamination of electrical neuroimaging data with eye-movement-related motor- and visual-evoked potentials. Indeed, because of an excellent temporal resolution and comparably low cost, electroencephalography (EEG) in combination with eye-tracking have become important tools for studying the temporal dynamics of naturalistic reading (e.g., Dimigen et al., 2011; Frey et al., 2018; Hollenstein et al., 2018; Loberg et al., 2018; Sato and

* Corresponding author. Methods of Plasticity Research Laboratory, Department of Psychology, University of Zurich, Andreasstrasse 15, CH-8050, Zürich, Switzerland.

E-mail address: christian.pfeiffer@uzh.ch (C. Pfeiffer).

https://doi.org/10.1016/j.neuroimage.2020.116934
Received 20 June 2019; Received in revised form 24 April 2020; Accepted 7 May 2020
Available online 13 May 2020

Mizuhara, 2018). In this context, fixation-related potentials (FRPs), which are the evoked electrical responses time-locked to the onset of fixations, have been studied and have received broad interest by naturalistic imaging researchers for free viewing visual perception (e.g., Rämä and Baccino, 2010), brain-computer interfaces (e.g., Finke et al., 2016), and naturalistic reading (e.g., Dimigen et al., 2011). In naturalistic reading paradigms, FRPs allow the study of the neural dynamics of how novel information from currently fixated text affects the ongoing language comprehension process. Evidence for this proposition has been provided by Dimigen et al. (2011), who showed that naturalistic reading of unexpected vs. expected words induced an N400 response in the FRP signals, previously observed for experimental paradigms using single word presentations (for a review, see Kutas and Federmeier, 2011). In addition, Frey et al. (2018) found modulations of slow-wave components of FRPs that depended on whether participants performed a memorization or decision-making task while reading, and Sato and Mizuhara (2018) found differences in early (100–200 ms) and late (400–500 ms) FRP components between words subsequently forgotten or remembered by the participants. Collectively, these studies indicate that FRPs provide useful information at high temporal resolution about the cognitive-neural processes that underlie naturalistic reading in humans.

Research on how humans process naturalistic language is paralleled by another line of research in artificial intelligence, called natural language processing, which aims at developing computer algorithms for decoding the meaning from natural language material. In this context, an important topic is sentiment analysis (Cambria et al., 2013; Liu and Zhang, 2012), which aims at detecting emotions and opinions expressed in text for applications such as hate-speech or sarcasm detection (Mamidi et al., 2019). Sentiment analysis has mainly focused on text-based processing using linguistic models (Agarwal et al., 2015) or machine-learning-based prediction of sentiment annotations of humans (Yang et al., 2012). More recently, however, other researchers (including ourselves) have proposed that text-based sentiment analysis can be considerably improved by contemplating neuro-cognitive data produced by human readers during naturalistic reading (Chanel et al., 2006; Hollenstein et al., 2018; Mishra et al., 2016; Raudonis et al., 2013). These signals include eye-movement parameters such as fixation duration and number of fixations, as well as the FRPs elicited by fixating words of different sentiment. Related work has used EEG power spectra data from participants watching movie clips (Nie et al., 2011; Wang et al., 2014) or viewing pictures of human faces (Li and Lu, 2009) to decode the polarity of the evoked sentiment (i.e., positive vs. negative). In a recent study, we showed improvement of decoding performance for relationship classification, entity recognition, and sentiment analysis using gaze position and EEG activity, in addition to text-based features (Hollenstein et al., 2019). However, the spatiotemporal neural dynamics of sentiment processing in humans remain largely unknown. With a few exceptions, FRPs have not been used to assess emotional processing in humans. Guérin-Dugué et al. (2018) recorded FRPs from participants freely viewing images of faces with different emotional expressions and found FRP differences depending on the emotion expressed at 200–300 ms after fixation onset. Simola et al. (2013) showed images of pleasant and unpleasant scenes to participants and found FRP differences at 400–500 ms. Based on these results for free-viewing image exploration, it remains unknown whether similar FRP differences are evoked by naturalistic reading of text with positive vs. negative sentiment.

In this study, we investigated the neural dynamics of sentiment processing in participants silently reading sentences from English movie reviews while simultaneous EEG and eye-tracking signals were recorded from 7129 words in 400 sentences (data taken from Hollenstein et al., 2018). The sentences were presented to the subjects in a naturalistic reading scenario, where the complete sentence was presented on the screen and the subjects read each sentence at their own speed. This allowed readers themselves to determine how long they fixated on each word and on which word to fixate next. By simultaneously acquiring EEG and eye-movement data, we determined the exact timing and gaze position with respect to word boundaries while subjects were reading sentences, a phenomenon that allowed us to extract EEG signals for word-level processing. In order to extend current insights into lexico-semantic processing during naturalistic reading, we aimed to identify whether and how words with different sentiment connotation (i.e., positive vs. negative emotional valence) would affect the FRP responses to word fixations during naturalistic reading. Moreover, in line with recent work on natural language processing (Hollenstein et al., 2018), we aimed to assess whether the word sentiment could be decoded from FRP data in a data-driven fashion.

2. Methods

The data used in this study were taken from the ZuCo dataset (Hollenstein et al., 2018), an openly available dataset of EEG and eye-tracking data from subjects reading English sentences (https://doi.org/10.17605/OSF.IO/Q3ZWS). A detailed description of the entire ZuCo dataset, including individual reading speed, lexical performance, average word length, average number of words per sentence, skipping proportion on word level, and effect of word length on skipping proportion, can be found in Hollenstein et al. (2018). In the following section, we will describe the methods relevant to the subset of data used in the present study.

2.1. Participants

The ZuCo dataset comprises recordings from 12 healthy adults (5 females, 22–54 years, all right-handed) who are native English speakers (average score of 95% in the Lexical Test for Advanced Learners of English; Lemhöfer and Broersma, 2012). All participants gave their written informed consent prior to participation in the study. Data from all participants of the ZuCo dataset were included in the present study.

2.2. Materials and procedure

ZuCo contains data from three tasks: Normal Reading, Task-Specific Reading, and Sentiment Reading (see Hollenstein et al., 2018). In the present study, we focused our analysis on the data from the Sentiment Reading task (see below) because, compared to the other tasks, the reading material from the Sentiment Reading task contained frequent use of positive and negative expressions. Moreover, human annotations of sentiments at word, phrase, and sentence level were available (provided by Socher et al., 2013).

The linguistic material consisted of English sentences extracted from movie reviews from the Stanford Sentiment Treebank (Socher et al., 2013). The Stanford Sentiment Treebank comprises 11,855 single sentences parsed into individual phrases and annotated by three human judges. Sentiment labels are available for the word, phrase, and sentence level and consist of the average rating across subjects on a 5-point scale ranging from -2 (very negative) to 0 (neutral) to $+2$ (very positive). For the ZuCo dataset, a total of 400 sentences were randomly selected: 140 positive, 137 negative, and 123 neutral (based on human annotation). Each sentence was individually shown to the participant on a computer screen. The text had black color, Arial font, 20-point font size, and was presented on a gray background. Letters were 0.8 mm high, corresponding to a 0.674° visual angle. Words were double-spaced, lines were triple-spaced, and each line consisted of a maximum of 80 letters or 13 words. Long sentences spanned multiple lines with a maximum of 7 lines and only one sentence was presented at a time (see Fig. 1A for an exemplar sentence).

Participants were equipped with a control pad and used their right index finger to trigger the onset of the next sentence. Participants were orally instructed before the experiment to carefully monitor the elicitation of emotions and opinions while reading each sentence and to occasionally answer a control question. Throughout the experiment, control questions were presented for 47 out of 400 sentences,

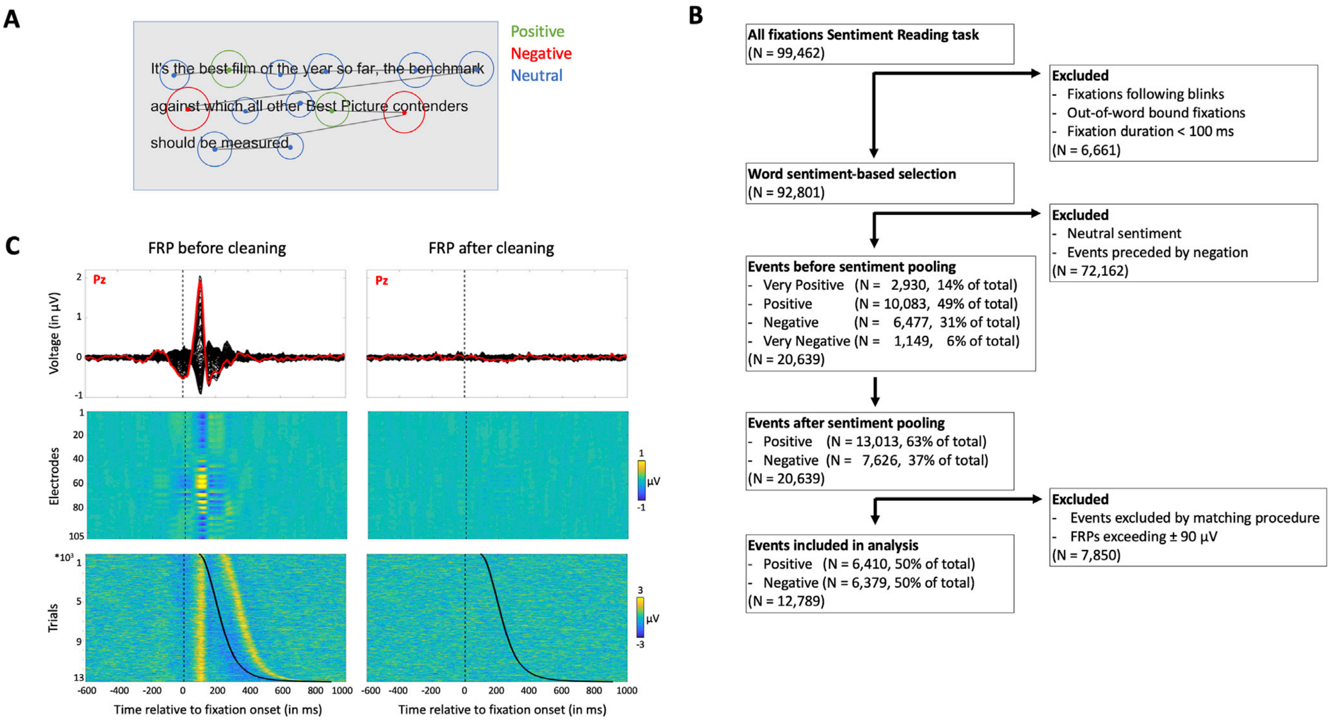

Fig. 1. A. Exemplar sentence and eye-movement sequence during naturalistic reading. Dots represent fixations, circle sizes represent fixation durations, lines connecting dots represent saccades, and colors represent sentiment type (positive, negative, or neutral) of the fixated word. B. Fixation event selection procedure. C. Comparison of fixation-related potentials (FRPs) before and after cleaning by electroencephalography (EEG) deconvolution modeling. FRPs before cleaning consist of both high-amplitude visual-motor signals and lower amplitude signals related to lexico-semantic processing. FRPs after cleaning consist only of lower amplitude signals from lexico-semantic processing and of random noise retained after removal of visual-motor signals. From top to bottom: Butterfly plot of average FRPs across all trials included in the analysis (black lines are FRPs for different channels, red line is the FRP of channel Pz); average FRPs across all channels; single-trial FRPs for channel Pz sorted by fixation duration and averaged over 50 adjacent epochs (black line represents fixation duration).

immediately after the participant had finished reading and pushed the button. Control sentences were the same for all participants. The following question was presented on the screen: *"Based on the previous sentence, how would you rate this movie? (very bad) |1–2 - 3–4 - 5 | (very good). Please press the corresponding number on the keyboard."* Ratings for the control questions were given with the control pad numbers 1 (very bad) to 5 (very good) and there was no time limit for providing the response. On average, participants correctly rated 79.53% (standard deviation [SD] = 11.22%) of the control questions. An overview of the sentiment ratings for control sentences is reported in Hollenstein et al. (2018).

Before the experiment, 3–5 practice sentences (including a control question) were presented to each participant to familiarize them with the task. The 400 sentences of the Sentiment Reading task were presented in 6 blocks of 60 sentences (10–15 min per block) and a final block of 40 sentences (8–12 min) in order to allow for regular re-calibration of the eye-tracker between blocks. The order of blocks and sentences within blocks was identical for all subjects. The blocks were presented in two sessions (4 blocks in the first session and 3 blocks in the second session, at the same time of day) in order to reduce fatigue in participants, who completed two additional tasks for the ZuCo study (Hollenstein et al., 2019). Between recording sessions, the proportion of sentences with negative (33%), positive (33%), or neutral (33%) sentiment out of the total number of sentences as well as the proportion of words with negative (6%), positive (10%), or neutral (84%) sentiment out of the total number of words were matched.

2.3. *Data acquisition*

Data acquisition took place in a sound-attenuated and dark Faraday recording cage. Participants were comfortably seated at a table in front of a 24-inch monitor (ASUS ROG, Swift PG248Q, display dimensions 531.4 × 298.9 mm, resolution 800 × 600 pixels [resulting in a 400 × 298.9 mm display], and vertical refresh rate of 100 Hz) placed 68 cm from the participant. A stable head position was ensured via a chin rest. Participants were instructed to stay as still as possible during the tasks. They were offered snacks and water during the breaks and were encouraged to rest. The experiment was programmed in MATLAB 2016b (The MathWorks Inc., Natick, MA, US), using the PsychToolbox extension. The order of the reading paradigms (i.e., Normal Reading, Task-Specific Reading, and Sentiment Reading) and sentence presentation was the same for all participants. Participants completed the tasks sitting alone in the room while two experimenters monitored their progress in the adjoining room.

2.3.1. *Eye-tracking acquisition*

An infrared video-based eye-tracker (EyeLink 1000 Plus, SR Research, http://www.sr-research.com/) with a sampling rate of 500 Hz and an instrumental spatial resolution of 0.01° was used to record gaze position and pupil size during the experiment. The eye-tracker was calibrated with a 9-point grid before each recording block. Specifically, participants were asked to direct their gaze in turn to a dot presented at each of nine locations in a random order. In a validation step, the calibration was repeated until the error between two measurements at any point was less than 0.5°, or the average error for all points was less than 1°.

2.3.2. *EEG acquisition*

High-density EEG data were recorded at a sampling rate of 500 Hz with a bandpass filter of 0.1–100 Hz, using a 128-channel EEG Geodesic Hydrocel system (Electrical Geodesics, Eugene, Oregon). The recording reference was at Cz. For each participant, head circumference was measured, and an appropriately sized EEG net was selected. The impedance of each electrode was checked prior to recording, to ensure good contact, and was kept below 40 kΩ. Good electrode impedance levels were checked and restored after every third block of 60 sentences (approximately every 30 min).

2.4. Data preprocessing

2.4.1. Eye-tracking preprocessing

The EyeLink 1000 tracker processes eye-position data: It identifies saccades, fixations, and blinks. Saccades are detected by the velocity and acceleration of the eye movements. Here, SR-research default system parameters have been used to define saccades: an acceleration threshold of 8000° per sec^2, a velocity threshold of 30° per sec, and a deflection threshold of 0.1°. Fixations were defined as time periods without saccades. The dataset therefore consists of (x,y) gaze location entries for individual fixations. Coordinates were given in pixels with respect to the monitor coordinates (the upper left corner of the screen was (0,0) and down/right was positive). Further, a blink can be regarded as a special case of a fixation, where the pupil diameter is either zero or outside a dynamically computed valid pupil, or the horizontal and vertical gaze positions are zero. For later EEG analysis, we only extracted fixations within the boundaries of each displayed word (Fig. 1B). In naturalistic reading, gaze fixations typically fall on or in-between adjacent words (Dimigen et al., 2011). In order to determine the word currently fixated by the participant, we defined word boundaries (Beymer and Russell, 2005; Hara et al., 2012; Tateosian et al., 2015), i.e., rectangular regions of interest around each word, extended laterally to cover half of the space between the word inside the boundary and the subsequent word in the line. This design resulted in boundaries that were non-overlapping and covered the entire space between subsequent words. Raw data from the eye-tracker showed slightly more variability of gaze positions along the y-axis as compared to the x-axis, similar to Ehinger et al. (2019). Thus, in our data, gaze position along the y-axis was occasionally close to but outside of the vertical word bounds. Given that naturalistic reading occurs within lines of text and not between the lines, we corrected the y-axis data using the following procedure. Fixations located 50 pixels above the first line or below the last line were excluded from analysis (i.e., out-of-bound fixations). Next, we applied a Gaussian mixture model (GMM) on y-axis gaze data, with the number of Gaussians set equal to the number of lines in the current trial (between 1 and 5; for details, see Hollenstein et al., 2018). As a result, each gaze position was clearly assigned to a specific text line. We used the corrected y-axis gaze positions for subsequent analyses. Fixations that were shorter than 100 ms were excluded from the analyses because they are unlikely to reflect fixations relevant for reading.

2.4.2. EEG preprocessing

EEG data were preprocessed with the Automagic toolbox for MATLAB (version: 1.9, https://github.com/methlabUZH/automagic; Pedroni et al., 2019). One-hundred-and-five EEG channels were used for scalp recordings and nine electrooculography (EOG) channels were used for artifact removal. The remaining channels (lying mainly on the neck and face) were discarded before data analysis (see Langer et al., 2012). Bad electrodes were identified and replaced. Identification of bad electrodes was based on the EEGLab plugin clean_rawdata (http://sccn.ucsd.edu/wiki/Plugin_list_process). This plugin removes flatline, low-frequency, and noisy channels. A channel was defined as a bad electrode when recorded data from that electrode were correlated at less than 0.85 to an estimate based on other channels (channel criterion). Furthermore, a channel was defined as a bad channel if it had more line noise relative to its signal compared to all other channels (4 standard deviations). Finally, if a channel had a flatline longer than 5 s, it was considered to be bad. In a next step, we ran the EEG processing pipeline “PREP” for robust average referencing (Bigdely-Shamlo et al., 2015), including using the CleanLine plugin for EEGLAB (Mullen, 2012) for removing power line noise at 50, 100, 150, 200, and 250 Hz. Next, the EEG data were band-pass filtered between 1 and 50 Hz with a Hamming windowed-sync finite impulse response zero-phase filter (EEGLAB function pop_eegfiltnew.m) for detrending the data and to remove high-frequency components of no interest. The filter order was defined to be 25% of the lower passband edge. In this study, we used Multiple Artifact Rejection Algorithm (MARA), a supervised machine-learning algorithm that evaluates Independent Component Analysis (ICA) components, for automatic artifact rejection. MARA has been trained on manual component classifications; thus, it captures the wide range of artifacts that manual rejection detects (Winkler et al., 2014, 2011). MARA has proven especially effective at detecting and removing eye and muscle artifact components. Specifically, MARA evaluates each component on the six algorithm features from the spatial, spectral, and temporal domain (Winkler et al., 2011, 2014). Subsequently, bad electrodes were interpolated by using a spherical spline interpolation eeg_interp.m. We quantified the quality of EEG data using four quality measures implemented in the Automagic toolbox (see Pedroni et al., 2019). One indicator for good quality of the data is the ratio of identified bad and, consequently, interpolated channels (RBC). The more channels that are interpolated, the more of the signal of interest is lost and, hence, the worse the data quality. All subjects had less than 15% RBC. A second quality measure is the ratio of data with overall high amplitude (OHA), which is defined by calculating the ratio of data points (i.e., electrodes x timepoints) that have a higher absolute voltage magnitude of 30 μV. The EEG data of all subjects exhibited an OHA of less than 10%. Similarly, the third quality measure is the ratio of timepoints of high variance (THV). THV is identified where the standard deviation of the voltage measures across all channels exceeds 15 μV. The THV for all subjects was below 10%. Finally, the ratio of channels of high variance (CHV), for which the standard deviation of the voltage measures across all time points exceeds 15 μV, was assessed. The CHV was below 15% for all subjects. During data acquisition, event triggers at the start and the end of each sentence presentation were simultaneously sent from the stimulus presentation computer to both the EEG recording and eye-tracking systems. After data preprocessing, these event triggers served to temporally synchronize the EEG and eye-tracking data using the “EYE-EEG extension” (Dimigen et al., 2011). The synchronization is performed at the event triggers for sentence onset and offset by fitting linear functions to the latencies recorded in the EEG and eye-tracking data and subsequently merging the EEG and eye-tracking data. Synchronization quality was ensured by comparing the trigger latencies recorded in the EEG and eye-tracker data. All synchronization errors did not exceed one sample (2 ms), which is to be expected because the same sampling rate (500 Hz) was used for both EEG and eye-tracking data acquisition. Finally, the synchronized EEG and eye-tracking data were downsampled to 125 Hz using the EEGLAB function pop_resample.

2.4.3. EEG deconvolution modeling

Free viewing is an important characteristic of naturalistic behavior and imposes challenges for the analysis of electrical and functional neuroimaging data. We note that free viewing in the context of our study refers to the participant’s ability to perform self-paced reading given the experimental requirement to keep the head still during data recording. In the case of EEG recordings during naturalistic reading, the self-paced timing of eye fixations—with average durations of 200–250 ms and variable onset asynchronies (e.g., Dimigen et al., 2011)—leads to a temporal overlap between successive fixation-related events, including short-latency high-amplitude visuo-motor potentials and mid-to long--latency lower-amplitude lexico-semantic processing related potentials. There is also a contamination of the signal of interest (i.e., lexico-semantic processing) with stereotypical high-amplitude evoked electrical responses to saccadic eye movements and visual processing upon fixation onset. In order to isolate the signals of interest and correct for temporal overlap in the continuous EEG, several authors have proposed methods using linear-regression-based deconvolution modeling for estimating the overlap-corrected underlying neural responses to events of different types (e.g., Ehinger and Dimigen, 2019; Smith and Kutas, 2015a, 2015b) for detailed discussions). Events of interest were electrical responses to saccadic eye movement, visual-evoked responses, blinks, and button-press related motor responses. Here, we used the *unfold* toolbox for MATLAB (https://github.com/unfoldtoolbox/unfold/; Ehinger and Dimigen, 2019). Deconvolution modeling is based on the assumption that in each channel the recorded signal consists of a combination of time-varying and partially overlapping event-related responses and random noise. Thus, the model estimates the latent event-related responses to each type of event based on repeated occurrences of the event over time. First, a design matrix is created that includes the onset latencies and temporal offset from event onset for different events within a chosen time window. Based on Ehinger and Dimigen (2019), we modeled the EEG during naturalistic reading using eye-tracking based information about fixation onsets, saccade onsets and their amplitudes, blinks, and button press responses. We included the following event types and formulas in our model:

- Fixation: y ~1 + duration
- Saccade: y ~1 + spl(amplitude,10)
- Blink: y ~ 1
- Keypress: y ~ 1

Note that the dependent variable *"y"* corresponds to the EEG data from a given channel, *"~"* refers to being equivalent to or being modeled by, *"1"* refers to the intercept term of the model, "duration" is the fixation duration, and "spl(amplitude,10)" refers to the use of a spline predictor with 10 splines for modeling non-linear relationships between EEG responses and saccade amplitude (see Ehinger and Dimigen, 2019 for a similar approach). Next, the design matrix was time expanded to a −600 ms–1000 ms time window around the event onset, and finally the model coefficients (hereinafter betas) were estimated for each channel and subject separately by fitting the combined design matrix to the EEG recorded in each channel. We note that the above listed equations for the different types of events serve to construct a single time-expanded design matrix, which is fit to the continuous EEG data specifically for each EEG channel and independent of the other EEG channels. Thus, the specific order by which the equations are entered in the toolbox does not affect the deconvolution outcome (i.e., the beta estimates). The estimated betas reflect the average responses over all events for each type of event (e.g., evoked electrical responses to saccade-related eye movement, visual-evoked responses, blinks, and button-press-related motor responses), from which overlapping activity was removed. Note that the outcome of deconvolution modeling is one set of beta estimates for each model predictor/type of event (e.g., fixation), but the betas are not estimated for individual events (for similar approaches, see Brodbeck et al., 2018; Dimigen et al., 2011; Smith and Kutas, 2015b). We did not aim to statistically analyze the extracted betas because for group-level analysis our rather small sample size (N = 12) would have lacked statistical power and our primary aim was not to test for differences in visual-motor responses evoked by naturalistic reading. Instead, we aimed to remove high-amplitude and temporally overlapping visual and motor responses from the continuous EEG in order to identify the neural dynamics of word sentiment processing. Accordingly, we used the beta estimates from deconvolution modeling for data cleaning purposes (based on Ehinger and Dimigen, 2019). Specifically, based on the assumption that the continuous EEG consists of a linear combination of temporally overlapping visual, motor, and lexico-semantic processing signals and random noise, we computed a continuous time series of model-predicted activation reflecting only the temporally overlapping visual and motor response (model-predicted EEG). This was achieved by convolving the design matrix of events with the beta estimates from deconvolution modeling, resulting in a model-predicted EEG that consists of the same number of electrodes and time points as the continuous EEG. It is important to note that the model-predicted EEG reflects only visual-motor responses. The design matrix and beta estimates used for computing the model-predicted EEG were based on different types of visual-motor events but did not include information about the type of linguistic material (e.g., sentiment information). Subsequently, the model-predicted EEG was subtracted from the continuous EEG for data cleaning:

- Cleaned EEG = EEG – Model-Predicted EEG.

The resulting cleaned EEG thus corresponds to EEG data from which the high-amplitude temporally overlapping visual and motor-evoked responses were removed, and lower amplitude lexico-semantic processing related signals and residual noise were retained (see Fig. 1C for an illustration of the effect of data cleaning).

2.4.4. *Fixation-event selection*

In this study, fixations corresponded to reading- and non-reading-related events, where the majority were word fixations (93,116 of 99,462 fixations, 94%), defined as a fixation within a word boundary (see Fig. 1A for an overview). Given that the main aim of this study was to identify the differences in FRPs from words with positive vs. negative sentiment, we performed the following event selection procedure. First, fixations out of word bounds and fixations shorter than 100 ms were removed, as they were unlikely related to lexico-semantic processing (Sereno and Rayner, 2003). Fixations on neutral words were removed (which in large part consist of filler or stop words, such as "and" and "the"). Next, we excluded words with positive or negative sentiment that were preceded by negation particles (i.e., "no", "not", "without", "nor", and "neither") within three words prior to the current word (<1% of words). For the remaining words, only a small fraction had a strong sentiment, whereas the majority had a moderate sentiment (based on the annotations from the Stanford Sentiment Treebank; Socher et al., 2013). Accordingly, we pooled strongly negative (label "-2") and moderately negative (label "-1") words into the Negative condition and strongly positive (label "+2") and moderately positive (label "+1") words into the Positive condition (Fig. 1B).

Finally, we performed a trial sampling procedure to remove word-fixation-related differences between experimental conditions that were unrelated to the emotional valence of the words, including fixation duration, word length, fixation onset probability of preceding and subsequent words, and number of trials per condition (Hauk and Pulvermüller, 2004; Thibadeau et al., 1980). Specifically, we used a stratified random sampling procedure where the trials were grouped by sentiment (positive or negative), word length (1–20 characters), and fixation onset probability of previous or subsequent fixation. Next, we randomly selected trials from each group such that between the positive and negative sentiment condition the number of trials was matched. The number of trials for selection was given by the number of trials available in the group with a minimum number of trials. This approach resulted—within and across subjects—in a matched number of trials, matched probability distribution for word length, and matched fixation onset distributions (Fig. 4). We note that this procedure slightly reduced the number of accepted events relative to the number of available events for both the positive and negative sentiment condition. This phenomenon was related to the word length matching procedure. Specifically, for each subject and each word length, we first determined the number of events for each condition and then matched the number of events between conditions by randomly selecting from the condition with more events the same number of events as present in the condition with fewer events. Thus, across subjects and word lengths trials were removed from both the positive and negative condition in order to achieve a match of word-length distributions between the positive and negative condition for each subject (see Fig. 4D). Finally, we excluded events where the FRP exceeded a ±90 μV amplitude threshold to remove transient noise from the EEG (see below). We illustrate the workflow of event selection in Fig. 1B. Out of the final selection of 12,789 trials, for 6983 trials (55%) the sentiment at word- and sentence-level was congruent, for 2167 trials (17%) word- and sentence-level sentiment were incongruent, and the remaining 3639 trials (28%) contained words with positive or negative sentiment in neutral sentences.

2.4.5. Data segmentation

FRPs were extracted by segmenting the continuous EEG into epochs −600 ms–1000 ms relative to fixation onset events (similar to Ehinger and Dimigen, 2019). Epochs were extracted from both EEG without deconvolution (FRPs) and from EEG cleaned by deconvolution modeling (cleaned FRPs). FRP epochs exceeding a ±90 μV amplitude threshold were removed, and for consistency between analyses, the same epochs were removed from cleaned FRPs, in order to exclude transient noise (238 of 13,027 epochs, 2%). The total number of accepted epochs across subjects was 12,789 (6379 for the positive condition and 6410 for the negative condition, Fig. 1B) for both the FRP and cleaned FRP dataset.

2.5. Data analysis

We analyzed reading-related FRP data in two ways. First, the *FRP analysis* served to identify the spatiotemporal differences in FRPs between the positive and negative sentiment conditions. Second, we used *decoding analysis* to predict the sentiment label (positive or negative) based on a data-driven selection of FRP features. Finally, we conducted *control analysis* of eye-tracker data and linguistic features to exclude that systematic differences in eye-movement behavior and linguistic material selected for the positive and negative sentiment condition confounded our EEG analyses.

2.6. FRP analysis

We performed the FRP analysis on two independent aspects of the global electrical field: response strength and response topography (Brunet et al., 2011; Murray et al., 2008; Tzovara et al., 2012). Response strength was assessed by global field power (GFP; Lehmann and Skrandies, 1980), which is the standard deviation of the voltages across all channels at a given time point and reflects the global response strength independent of topographical configuration. We analyzed GFP using a timewise general linear mixed model (GLMM) to control for random subject effects. We complemented this analysis with an electrode-by-time GLMM and a nonparametric cluster-based permutation test to identify physiologically plausible electrode clusters where response strength differences are observed (Maris and Oostenveld, 2007). Finally, we tested for differences in response topography between the positive and negative sentiment conditions using the topographic consistency test (TCT; König and Melie-García, 2010), which assesses differences in the spatial configuration of the underlying neural generators independent of global electrical field strength (Lehmann and Skrandies, 1980). We conducted all analyses on reading-related FRPs, which contain oculomotor, visual, and lexico-semantic processing related signals, and on cleaned FRPs, which are FRPs from which high-amplitude oculomotor and visual processing related signals were removed by deconvolution modeling. All analyses used the Word Sentiment as predictor (see below). In addition, we conducted an analysis using Word, Phrase, and Sentence Sentiment as predictors. In the main manuscript, we report only the analyses for cleaned FRPs using the Word Sentiment predictor, because across the different analyses the results were highly similar, and model fits were slightly better for analyses including the Word Sentiment predictor rather than Word, Phrase, and Sentence predictors. The results of the remaining analyses can be found in the Supplementary Material.

2.6.1. GFP GLMM analysis

We computed GFP (Lehmann and Skrandies, 1980) for each time point and trial and subjected these data to a timewise statistical analysis using a GLMM. The GLMM takes advantage of the large number of fixations available for each subject and provides a more accurate and generalizable estimate of the effects, improved statistical power, and non-inflated type I errors (Singmann and Kellen, 2017). Group-level analyses, in which for each subject only the GFP for one average FRP for the positive and one for the negative sentiment condition are available, were not carried out because these analyses would have lacked statistical power due to a small sample size ($N = 12$). The GFP-GLMM analysis of cleaned FRPs was carried out using the fixed effect predictor Word Sentiment (levels: negative or positive) and the random effect predictor Subject (12 levels) for a subject-wise random intercept and subject-wise random slope. The GLMM was computed using the MATLAB function *fitglme* using a normal distribution, identity link function, and La Place fit method. After computing the GLMM across time points and electrodes, the fixed effect statistics for the Word Sentiment predictor were extracted. An alpha threshold of $p < 0.05$ was used and correction for temporal autocorrelation was based on a >40-ms duration criterion (i.e., >5 sampling points; Lehmann and Skrandies, 1980).

2.6.2. Electrode-by-time GLMM analysis

To compare the spatiotemporal differences between positive and negative sentiment during word processing, we carried out analyses across all trials from all subjects using the same GLMM as for GFP analysis. After computing the GLMM across time points and electrodes, the fixed effect statistics for the Word Sentiment predictor were extracted. An a priori alpha threshold of 0.05 was applied and correction for multiple comparisons across electrodes and time points was based on a >40-ms duration (i.e., >5 sampling points; Lehmann and Skrandies, 1980) and >5 electrodes (i.e., >5% of all electrodes, similar to Matusz et al., 2015) criterion. This measure is in line with the spatiotemporal clustering commonly observed for EEG event-related potentials (Mensen and Khatami, 2013; Murray et al., 2008).

2.6.3. Cluster-based permutation t-test

We complemented the electrode-by-time analysis with a cluster-based permutation *t*-test (Maris and Oostenveld, 2007), implemented in the FieldTrip toolbox for MATLAB (Oostenveld et al., 2011). The cluster-based permutation test was applied to the FRPs and tested for differences between the trials for the positive and negative sentiment condition. GLMM analysis is currently not available in the toolbox; hence, we used a paired-samples test between trials from the positive and negative sentiment condition. It showed highly similar results to the electrode-by-time GLMM (see Figs. 2 and 3). This analysis identified data samples showing significant t-values ($p < 0.05$, two-tailed) that were clustered based on temporal and spatial proximity (i.e. >4 neighboring electrodes). Each cluster was assigned to cluster-level statistics corresponding to the sum of the t-values of the samples belonging to that cluster. The type I error rate was controlled by evaluating the maximum cluster-level statistics by randomly shuffling condition labels 1000 times to estimate the distribution of maximal cluster-level statistics obtained by chance and applying a two-tailed Monte-Carlo p value. This procedure was applied at the sensor level in the time window from −600 to 1000 ms relative to fixation onset.

2.6.4. TCT

The TCT (König and Melie-García, 2010) is a permutation-based test that assesses the consistency of the topographical distribution of voltages across electrodes for repeated observations. In the present study, we were interested in the consistency of the difference between the positive and negative sentiment condition. Thus, we computed the difference between cleaned FRPs from the positive minus negative sentiment condition. The difference signals were computed by ranking the trials for each condition by subject, word length, and fixation duration, and then computing the differences between matched trials. In the case of an unequal number of trials between conditions, leftover trials were discarded (<1% of trials). We note that a GLMM analysis was not available for this test, but random effects between subjects were partially accounted for by computing the positive-negative differences within subjects. The topographic consistency test uses an electrode-level randomization method to estimate a data distribution under the null hypothesis, whereby at each time point the GFP of the grand average across trials is computed and compared to GFP values computed on 5000 random shuffles of electrode positions. The p value is the tail probability of the grand average GFP being larger than the permutation-based GFP values. The alpha threshold was set to $p < 0.05$, and correction for temporal autocorrelation was based on a >40-ms duration criterion (i.e., >5 sampling points, Lehmann and Skrandies, 1980).

2.7. Decoding analysis

We complemented the descriptive FRP analysis, which does not allow interpreting results beyond the data used for analysis, with a predictive analysis aimed at decoding word-level sentiment from unseen (hold-out) FRP trials (see Breiman, 2001; Yarkoni and Westfall, 2017 for a discussion). The decoding analysis was independent of the FRP analysis and had the goal to test whether there are any detectable differences in FRPs for reading words with positive vs. negative sentiment. This analysis did not aim at developing algorithms for brain-computer interfaces. Instead, we aimed at maximizing the sensitivity of our analysis for detecting sentiment-related differences in FRPs by using event-matched trials (i.e., 12,789 trials from 12 subjects, Fig. 1B) from cleaned FRPs, because visuo-motor artifacts of no interest were removed from these data. The analysis focused on the 0-500-ms post-fixation interval (63 sampling points), because pre-fixation EEG were unlikely to contribute to reading-related sentiment processing. Thus, a total of 6615 features were used (i.e., 105 EEG channels x 63 sampling points). Given that single-trial EEG has a low signal-to-noise ratio (SNR), due to high-amplitude environmental and physiological electrical signals not time-locked to fixation onset, we aimed to increase SNR by trial averaging. Similar to Tuckute et al. (2019), we compared decoding performance for single-trial to trial-averaged data. Trial averages were computed within subjects for random selections of 10, 20, or 40 trials of the same word-level sentiment (i.e., positive or negative) and similar word length. For example, a 40-trial average was computed on FRPs of 40 words of the same sentiment and similar word length that were presented in different sentences. A maximum of 40 trials per average were used (similar to Tuckute et al., 2019) as a trade-off between SNR improvement and reducing the number of trials available for decoding analysis. If the number of trials per subject was not evenly divisible by the desired number of trials per average (i.e., 10, 20, or 40 trials), the remaining trials were averaged and included in the analysis. Thus, a maximum of 12 trials (one per subject) contained fewer trials than the desired number of trials per average (i.e., 30–80% of the desired number of trials). The resulting number of samples was thus 12,789 samples for single-trial, 1300 samples for 10-trial, 650 samples for 20-trial, and 328 samples for 40-trial averages.

Subsequently, the data were randomly split 100 times into a training (95%) and test set (5%) using stratification. That is, the proportion of samples per subject and sentiment condition relative to the total number of samples was matched between training and test set (i.e., a majority-class baseline classifier exhibited a mean test set accuracy of 0.5 [95% confidence interval [CI]: 0.5, 0.5], see Supplementary Material). For each data split, we normalized the features using scikit-learn (https://scikit-learn.org/stable/, StandardScaler class), a machine learning framework for Python, by computing feature-wise mean and standard deviation parameters from training set data. We subsequently applied these parameters for feature normalization of the training and test set. Parameter estimation was based only on the training set to prevent data leakage from training to test set (Kononenko and Kukar, 2007). This procedure was followed by dimensionality reduction, which served to avoid overfitting of classifiers because of the large number of available features (i.e., 6615 features) relative to a small number of samples (328–12,789 samples) in our dataset. Dimensionality reduction for the analyses reported in the main manuscript was performed using neighborhood component analysis (NCA; Goldberger et al., 2005), a supervised classification method based on k-nearest-neighbors classification that maximizes differences between classes for a desired number of k features (i.e., number of retained components after dimensionality reduction). We chose NCA instead of principal component analysis (PCA) because previous research has shown that for EEG data PCA mainly retains high-amplitude noise for the first principal components (e.g., Artoni et al., 2018), which is not the case for NCA (see Goldberger et al., 2005 for a comparison). We report in the Supplementary Material a comparison of decoding results for NCA and PCA, both of which showed similar results. Similar to feature normalization, we used scikit-learn (NeighborhoodComponentAnalysis class) and only trained the NCA classifier on the training set, in order to prevent data leakage between training and test samples (Kononenko and Kukar, 2007). The number of k features to retain after dimensionality reduction using the NCA classifier was determined during model optimization (see below). Subsequently, decoding analyses using support vector machine (SVM) and logistic regression classifiers were performed. We also report analyses using long short-term memory (LSTM), dense neural networks, and a majority-class classifier in the Supplementary Material. An SVM model is a representation of the samples as points in space, mapped so that the samples of the separate categories are divided by a clear gap that is as wide as possible. New samples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall (Raschka, 2018; Tuckute et al., 2019). SVM classifiers have been previously used successfully on single-trial and trial-average EEG data (e.g., Tuckute et al., 2019). We used SVM classifiers implemented in scikit-learn (SVC class) that employ a radial basis function kernel, which is well suited for finding non-linear decision functions, by using a value of 1 for the regularization parameter “C” and a value of 0.03 for the kernel width parameter “γ” (similar to Tuckute et al., 2019) or using parameter optimization using grid search in scikit-learn (GridSearchCV class). Parameter optimization was based on 10-fold cross-validation of the training set using grid search across the hyperparameters: number of NCA components [10, 20, 40, 80, 160, 320, 640, 1280], regularization parameter C [ten values in the range: 1×10^{-7}; 1×10^{2}], and kernel-width parameter γ [ten values in the range: 1×10^{-7}; 1×10^{2}]. The model achieving the highest average accuracy on validation sets across 10 cross-validation folds was subsequently used for model testing. We observed a median number of $k = 320$ features across data splits for the final model used for testing. The optimal parameters were subsequently used to train an SVM classifier on the entire training set. The SVM classifier was subsequently used for predicting the class labels (i.e., word sentiment) of the test set samples. For the logistic regression classifier, we also used the scikit-learn implementation (LogisticRegression class) and followed the same training and parameter optimization procedure as for SVM classifiers, except for the difference that no γ parameter is required for logistic regression and for the use of an L2-norm.

Decoding performance for SVM and logistic regression classifiers was evaluated based on the comparison between predicted labels (i.e., positive or negative sentiment) and true labels of the test set resulting in a number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) across the classified trials. The following decoding performance metrics were computed:

- Accuracy = (TP + TN)/(TP + FP + FN + TN)
- Precision = TP/(TP + FP); Recall = TP/(TP + FN)
- F1 Score = 2*(Recall × Precision)/(Recall + Precision)

We report mean classification performance and 95% confidence intervals of the mean across 100 random splits for training and test set (Table 1).

2.8. Control analysis

We conducted additional control analysis of eye-tracker data and linguistic features to exclude that systematic differences in eye-movement behavior and linguistic material selected for the positive and negative sentiment condition confounded our EEG analyses. Eye-tracker data were analyzed by extracting horizontal and vertical eye velocity, computed as the speed of gaze position-changes along the horizontal and vertical axis of the screen over time. We computed the first derivative on gaze position (in screen coordinates) and segmented the resulting eye velocity data in −600 to 1000 ms peri-fixation epochs (no baseline correction). Statistical analysis was performed separately for horizontal and vertical eye velocity epochs using timewise GLMMs (model formula and parameters identical to those used for FRP Analysis, see below) across the 12,789 fixation trials from all subjects. Correction for multiple comparisons was performed using a minimum 40-ms duration criterion (>5 sampling points, Lehmann and Skrandies, 1980).

Next, we extracted fixation onset probabilities, which are the probability distributions across time of the n ± 3 fixations preceding or following the current fixation (n). Fixation onset probabilities provide information about the proportion of trials out of the total number of trials showing a fixation onset, and thus additional saccadic and visual-evoked electrical responses time-locked to these fixations, within a given time interval relative to the current fixation (see Dimigen et al., 2011 for a similar approach). Any difference in fixation onset probabilities between the positive and negative sentiment condition would indicate differences in the amount of visual and motor artifacts between conditions and could, therefore, confound the FRP analysis (especially for data not cleaned by deconvolution modeling) in identifying FRP differences related to lexico-semantic processing. We extracted fixation onset probabilities for both no-word fixations and word fixations of any sentiment occurring in the −600 to 1000 ms peri-fixation interval for all trials of interest (i.e., positive and negative word fixation trials from event selection, see below). First, we computed fixation onset times (in ms) of all preceding and subsequent fixations relative to current fixation onset. We then computed for each condition the probability of fixation onset in consecutive and non-overlapping 60-ms bins as the number of fixations per bin divided by the number of trials per condition across the entire −600 to 1000 ms peri-fixation interval. The data were statistically analyzed using bin-wise paired samples t-tests ($p < 0.05$; >1-consecutive-bins criterion for multiple comparison correction).

In addition, we compared word length and fixation duration between positive and negative sentiment trials using the same GLMM model as used for FRP Analysis. Finally, we compared word frequencies between words with positive and negative sentiment from the Sentiment Reading task. Among the 7129 words, from 400 sentences of the Sentiment Reading task, there were 2475 unique words, of which 275 (11%) had a negative, 400 (16%) had a positive, and 1800 (73%) had a neutral sentiment. We note that this imbalance in the number of unique words for the positive and negative sentiment condition is comparable to the imbalance in fixation trials before event matching (see Fig. 1B). However, after event matching the number of trials used for FRP analysis was matched between the positive and negative sentiment condition. We then extracted word frequencies for unique words by counting the number of times each word occurred in the text material from the Sentiment Reading task (7129 words). We statistically compared the frequencies of words with a positive sentiment and words with a negative sentiment using a two-samples *t*-test ($p < 0.05$).

3. Results

In this study, we assessed the electrical neural correlates of word sentiment processing during naturalistic reading by testing for differences between negative and positive sentiment in reading-related FRPs (FRP analysis) and by predictive modeling aimed at predicting the sentiment of the word from FRP data (decoding analysis). Finally, we compared eye movement and linguistic features between sentiment condition to exclude their confounding contribution the EEG analyses (control analysis).

3.1. FRP analysis

We performed the FRP analysis on two independent aspects of the global electrical field: response strength and response topography (Brunet et al., 2011; Murray et al., 2008; Tzovara et al., 2012). Response strength was assessed by time-wise GFP analysis using a GLMM to control for random subject effects. We complemented this analysis with an electrode-by-time GLMM and a nonparametric cluster-based permutation test to identify physiologically plausible electrode clusters of response strength differences between sentiment conditions (Maris and Oostenveld, 2007). Finally, we tested for differences in response topography between the positive and negative sentiment conditions using the TCT (König and Melie-García, 2010). We will first present the results for FRPs, which contain oculomotor, visual, and lexico-semantic processing related

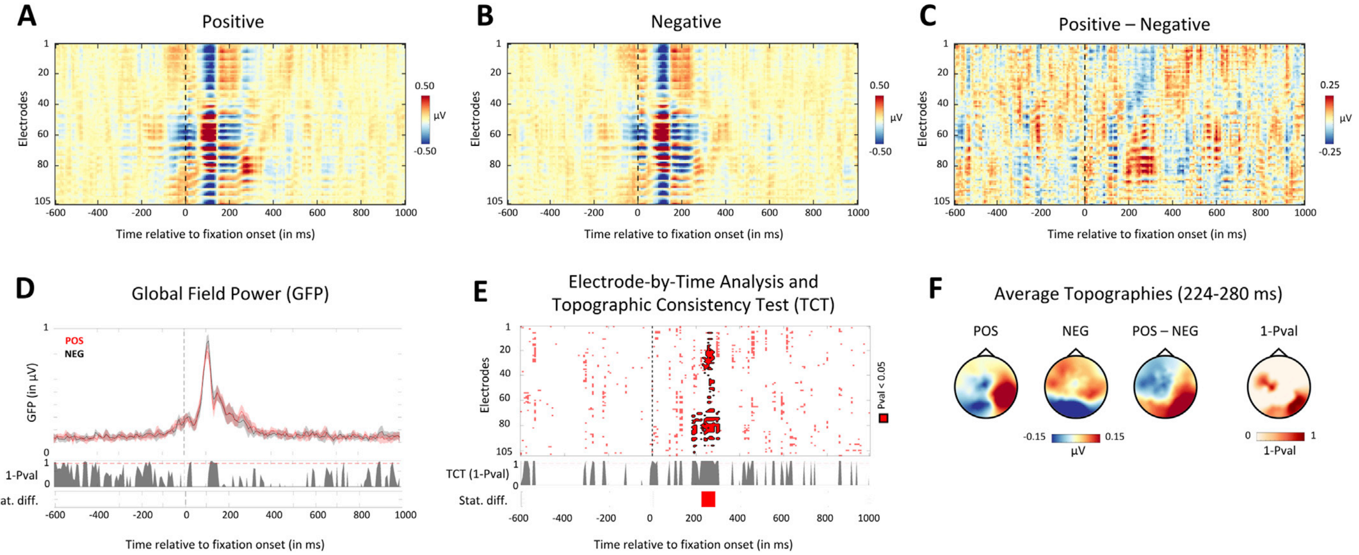

Fig. 2. Fixation-related potential (FRP) results. A-C. FRP grand averages for word fixations with positive sentiment, negative sentiment, and the difference between positive and negative sentiment. D. Global field power and statistical results from timewise general linear mixed model (GLMM) analysis. E. Top panel: Statistical results of electrode-by-time GLMM analysis showing significant time points ($p < 0.05$) in red and significant electrode clusters (cluster-level $p < 0.05$) highlighted with black contours. Bottom panels: Statistical results of the topographic consistency test (TCT). F. Average topographies in the period in which statistical differences were observed. Statistical differences: $p < 0.05$ for >40 ms and >5 electrodes.

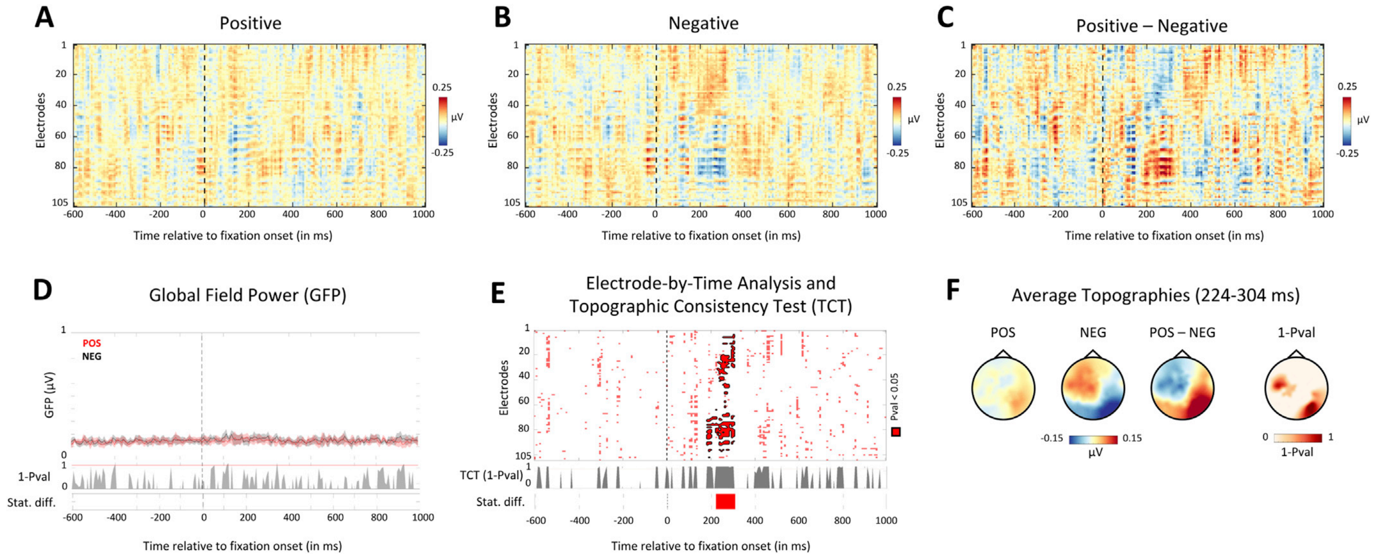

Fig. 3. Results for cleaned fixation-related potentials (FRPs), i.e., FRPs from which high-amplitude oculomotor and visual processing related activity was removed by deconvolution modeling (see Methods). A-C. Cleaned FRP grand averages for word fixations with positive sentiment, negative sentiment, and the difference between positive and negative sentiment. D. Global field power (GFP) and statistical results from timewise general linear mixed model (GLMM) analysis. E. Top panel: Statistical results of electrode-by-time GLMM analysis showing significant time points ($p < 0.05$) in red and significant electrode clusters (cluster-level $p < 0.05$) highlighted with black contours. Bottom panels: Statistical results of the topographic consistency test. F. Average topographies in the period in which statistical differences were observed. Statistical differences: $p < 0.05$ for >40 ms and >5 electrodes.

signals, followed by the results for cleaned FRPs, which are FRPs from which high-amplitude oculomotor and visual processing related signals were removed by deconvolution modeling.

3.1.1. FRP results

The timewise GLMM analysis of GFP of FRPs showed no statistical differences between the positive and negative condition (Fig. 2D). The electrode-by-time GLMM analysis of FRPs showed a statistical difference between the positive and negative condition in the 224–280 ms interval after fixation onset (condition fixed effect: $p < 0.05$, >5 electrodes, >40 ms; Fig. 2E). The electrodes that showed significantly different activations in this interval were located in left frontocentral electrodes—they presented higher activation for the negative compared to the positive sentiment condition—and in a right-posterior electrode cluster, which exhibited higher activation for the positive compared to the negative sentiment condition (Fig. 2F). Similarly, the cluster permutation test identified two significant electrode clusters (Fig. 2E–F). The first cluster was located in left frontocentral electrodes at 232–280 ms after fixation onset and showed a negative activation difference between the positive and negative sentiment conditions (cluster-level $p = 0.04$). The second cluster was located in right-posterior electrodes at 184–304 ms after fixation onset and showed a positive activation difference between

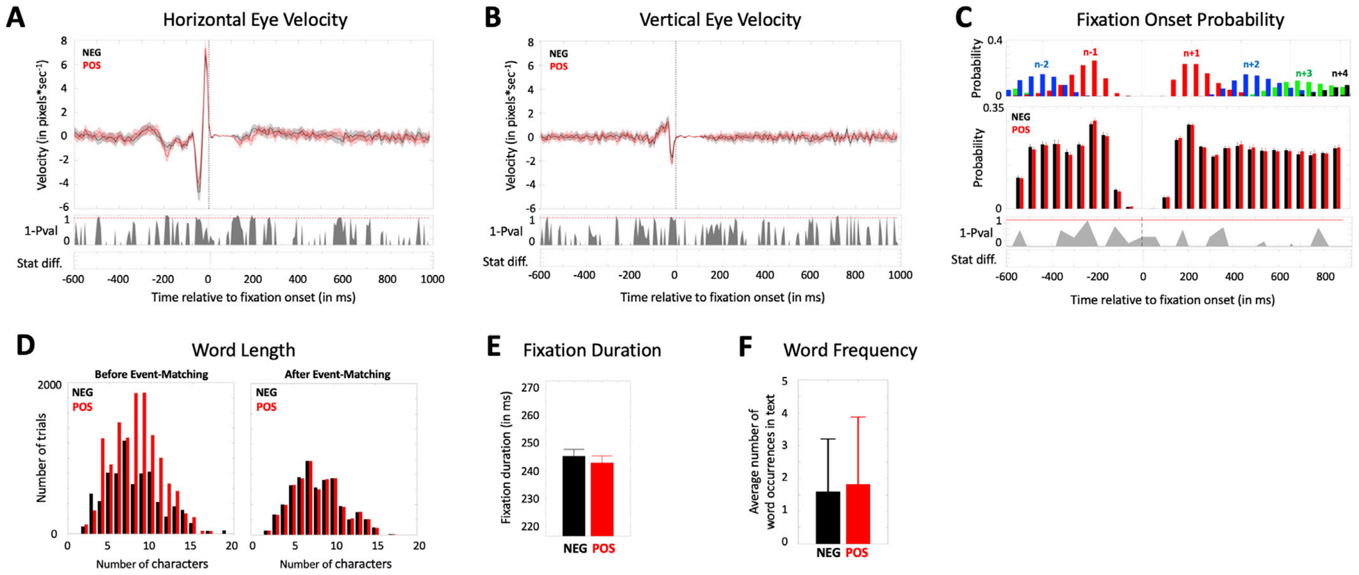

Fig. 4. Control analysis results. A-B. Horizontal and vertical eye velocity statistical comparison between the positive and negative sentiment condition. C. Fixation onset probabilities for ±3 fixations relative to current fixation (n) and statistical comparison of fixation onset probabilities within the −600 to 1000 ms peri-fixation interval (60-ms bins) between the positive and negative sentiment condition. D. Word length distributions before and after event matching for negative (black) and positive (red) sentiment conditions. E. Average fixation duration for the positive and negative sentiment conditions. F. Average word frequency for words with positive and negative sentiment.

Table 1
Sentiment decoding performance for cleaned FRP single trials and multi-trial averages for different classifiers. The first value represents the mean score across 100 test set classifications, and values in brackets refer to the 95% confidence interval. Differences from chance are highlighted bold.

	Single trials	*10-trial averages*	*20-trial averages*	*40-trial averages*
Support vector machines				
Accuracy	0.50 [0.49, 0.51]	0.51 [0.50, 0.52]	**0.54 [0.52, 0.55]**	**0.57 [0.55, 0.59]**
F1 Score	0.50 [0.49, 0.51]	**0.52 [0.51, 0.53]**	0.51 [0.49, 0.53]	**0.56 [0.54, 0.58]**
Precision	0.50 [0.49, 0.51]	0.50 [0.49, 0.52]	**0.54 [0.52, 0.55]**	**0.57 [0.55, 0.59]**
Recall	0.50 [0.49, 0.51]	**0.55 [0.53, 0.57]**	0.50 [0.47, 0.53]	**0.57 [0.54, 0.60]**
Logistic regression				
Accuracy	0.50 [0.50, 0.51]	**0.53 [0.52, 0.53]**	**0.55 [0.54, 0.56]**	**0.60 [0.58, 0.61]**
F1 Score	0.50 [0.50, 0.51]	0.52 [0.50, 0.53]	**0.54 [0.52, 0.55]**	**0.58 [0.56, 0.60]**
Precision	0.50 [0.50, 0.51]	**0.52 [0.51, 0.53]**	**0.55 [0.54, 0.56]**	**0.60 [0.58, 0.62]**
Recall	0.50 [0.50, 0.51]	0.52 [0.50, 0.54]	**0.54 [0.52, 0.56]**	**0.58 [0.55, 0.60]**

the positive and negative sentiment condition (cluster-level $p < 0.001$, Fig. 2F). Both clusters overlapped with the significant electrode clusters identified by the electrode-by-time GLMM analysis (Fig. 2E). The TCT identified a consistent topographical difference between FRPs for the positive and negative conditions in the 224–304 ms interval after fixation onset ($p < 0.05$, >40 ms, Fig. 2F).

3.1.2. Cleaned FRP results

The timewise analysis of GFP of cleaned FRPs showed no statistical differences between the positive and negative sentiment conditions (Fig. 3D). However, in line with the results observed for FRPs, the electrode-by-time GLMM analysis of cleaned FRPs showed statistical differences between the positive and negative conditions in the 224–304 ms interval after fixation onset in two electrode clusters. The first cluster was located in left-central scalp locations and showed higher activation for the positive compared to the negative sentiment condition (cluster-level $p = 0.009$). The second cluster was located at the right-posterior scalp location and showed lower activation for the positive compared to the negative sentiment condition (cluster-level $p < 0.0001$; Fig. 3E). The TCT identified a consistent topographical difference between cleaned FRPs for the positive and negative condition in the 224–304 ms interval after fixation onset ($p < 0.05$, >40 ms; Fig. 3D). Taken together, these results indicate topographical but not amplitude (GFP) differences in cleaned FRPs between the positive and negative sentiment condition.

In summary, the results from FRPs and cleaned FRPs were highly similar. These findings indicate that the presence of the high-amplitude visual-motor activation in FRPs did not affect the statistical results of our analysis. At the same time, these results rule out that our cleaning procedure artificially induced statistical differences between the positive and negative condition.

3.2. Decoding analysis

The results of decoding analysis of word sentiment from cleaned FRPs using the SVM and logistic regression classifiers are shown in Table 1. This analysis showed for single trials a chance-level performance (mean accuracy = 0.50, 95% CI: [0.49, 0.51]), possibly related to the low SNR in single-trial EEG (e.g., Tuckute et al., 2019). However, decoding analysis of trial-average data of 20 or more trials improved decoding performance to an above-chance level (0.60 mean accuracy; 95%CI: [0.56, 0.60]; Table 1). This finding indicates that increasing SNR improved sentiment decoding from cleaned FRPs. We report extensive analysis comparing different parameters for feature dimensionality reduction, classifiers, and tuning parameters in the Supplementary Material.

3.3. Control analysis

All control analyses were carried out on the final selection of events after fixation-event selection was performed (see Fig. 4B). The timewise analysis of horizontal and vertical eye velocity showed no statistical differences between the positive and negative condition in the entire −600 to 1000 ms peri-fixation interval (p values are shown in Fig. 4A–B). Likewise, bin-wise analysis (i.e., 60-ms bins) of fixation onset probabilities showed no significant differences between the positive and negative condition (p values shown in Fig. 4C). Finally, comparison of word length, fixation duration, and word frequency did not significantly differ between the positive and negative condition (Fig. 4D–F). In summary, the results of the control analysis showed no differences between the positive and negative sentiment condition in eye movements or linguistic features.

4. Discussion

This study used synchronized EEG and eye-tracking data to investigate the neural dynamics of word-level sentiment processing in humans reading naturalistic English sentences. Our results showed differences in the electrical neural responses to words with positive vs. negative sentiment that were reflected in an FRP topographical differences at 224–304 ms after fixation onset. Decoding analysis showed a consistent above-chance level decoding of the word sentiment based on cleaned FRP data (mean accuracy of 0.60). Our control analyses ruled out that these results were based on differences in eye movements or linguistic features between the positive and negative sentiment condition. In the following section, we will discuss these results with respect to previous research and add a methodological examination of advantages and limitations of our methods for naturalistic neuroimaging research.

4.1. Neural dynamics of word sentiment processing during naturalistic reading

Naturalistic reading is a complex multicomponent process that involves a temporal sequence of oculomotor, visual-perceptual, and cognitive processes for converting visual information into semantic information embedded into contextual memory. Thus, the brain processes of reading involve orthography, phonology, and semantic processing of single words, as well as processes relating this word-level information to grammar and lexico-semantic information of phrases and entire sentences (Citron, 2012; Hasson and Honey, 2012).

Neurophysiological studies of reading using word-by-word presentation demonstrated that approximately 100 ms after word presentation the visual input reaches the visual cortex. Around 50–100 ms later, the word is processed as strings of letters in a specialized region of the left visual cortex, and between 200 and 600 ms after a word is presented, its semantic properties are processed (Citron, 2012; Grainger and Holcomb, 2010; Salmelin, 2007). These findings have been substantiated by research showing sustained activity when reading words vs. non-words (Salmelin, 2007). Other studies have found a mismatch negativity response called N400 during this time period; this response is stronger for words that are semantically incongruent with previously presented words (Hillyard and Kutas, 2002). More recently, it has been shown that the N400 is a continuously graded response that depends on how surprising the word is (Frank et al., 2013). In contrast to single-word reading, naturalistic reading is characterized by a reader's spatiotemporal control. Readers move their eyes actively through text in a series of fixations and saccades (Dodge, 1901; Rayner and Clifton, 2009). Previous studies have suggested that the majority of word encoding and semantic language processing steps occur during word fixation (Clifton et al., 2016; Rayner and Clifton, 2009). In our study, the participants had an average reading speed of 5.5 words per second (Hollenstein et al., 2018) during self-paced naturalistic reading; this speed allowed them to extract semantic meaning from text (e.g., sentiment).

In the present study, we were specifically interested in spatiotemporal dynamics of sentiment word processing during naturalistic reading. A large body of literature has studied the neural dynamics of written sentiment word processing. The vast majority of existing findings are derived from controlled experiments (serial single word presentation and fixed presentation time), which may not generalize beyond the experimental setting (Hasson and Honey, 2012; see also our discussion of methodological considerations below). Here we primarily focus on the electrophysiological findings. For an overview of the hemodynamic neuroimaging (fMRI) studies, please refer to the review of Citron (2012).

Within the EEG literature, two event-related potential components, the early posterior negativity (EPN) and the late posterior positivity (LPP), have been repeatedly reported in the context of sentiment processing during reading (for reviews, see Citron, 2012; Kissler et al., 2006). The EPN has an occipital-temporal scalp distribution that peaks between 200 and 300 ms after word presentation. The EPN has been linked to attentional mechanisms during access to sentiment information (Schupp et al., 2004); this phenomenon suggests that this component is involved in implicit processing of emotional content. The EPN amplitude is reportedly increased for emotionally connotated words compared to neutral words during reading (Kissler et al., 2009, 2007), word recognition (Hinojosa et al., 2010), and lexical decision making (Citron, 2011; Schacht and Sommer, 2009; Scott et al., 2009). Although the EPN has mostly been examined with written verbal material, some studies have identified the EPN also in response to emotional pictures and faces (Martín-Loeches, 2007). Source localization of the EPN component has revealed that the EPN originates in the fusiform gyrus (Schacht and Sommer, 2009), or the visual word-form area (Hinojosa et al., 2010). These data support the hypothesis that a word's emotional connotation can be processed in parallel with the representation of its visual form (Kissler et al., 2006).

Our FRP analyses showed a topographical difference between the positive and negative sentiment conditions in an identical time window (224–304 ms after fixation onset), with a similar spatial scalp distribution as the EPN. Positive sentiment words exhibited higher activation compared to the negative words in a left temporal electrode cluster, as well as lower activations for the positive compared to the negative sentiment condition in a right occipital electrode cluster (Figs. 2F and 3F). As described earlier, this time period after fixation onset (200–300 ms) is generally associated with the processing of semantic properties (e.g., sentiment), such as integrating the visual stimulus with its corresponding lexical representation (e.g., Abdullaev and Posner, 1998). We hypothesize that spatiotemporal differences between the positive and negative sentiment word processing in the present study are closely related to the reported EPN component. Palazova et al. (2011) observed a topographical difference in FRPs similar to our study. Using single-word presentations, these authors found that at 300 ms after word onset there was a FRP topographical difference between adjectives with positive vs. negative sentiment with a centro-posterior topographical pattern similar to our study. However, their voltage pattern was different from our study (Palazova et al., 2011, Fig. 2). These results are interesting in that Palazova et al. (2011) and our study both found word-level sentiment differences in FRPs in a similar time period. However, the results of the two studies should be compared with caution because there are many methodological differences between the studies that may have affected group-level results. Those include the subject sample (native German vs. English speakers), the text material (German adjectives vs. English adjectives, verbs, and nouns), the experimental task (timed single-word presentation vs. self-paced naturalistic sentence reading), and the EEG recording settings (left mastoid reference and frontal ground electrode vs. vertex reference and posterior-central ground). In order to directly compare the paradigms of Palazova et al. (2011) and our study, both should be carried out in the same subjects, using similar linguistic material and EEG recording settings. Future work should investigate the relationship between emotional valence and linguistic features (such as word class) during naturalistic sentences reading. The timing of our results in the range of 200–300 ms after fixation onset are also compatible with an alternative explanation, namely cognitive-processing-related P300–N400 components. However, the timing of the sentiment effect in the present study and the topographical pattern (more posterior and more asymmetric) differed from N400 effects of word frequency or predictability (e.g., Dimigen et al., 2011), an outcome that suggests different underlying neural dynamics.

Another frequently reported event-related potential component connected to emotional word processing is the LPP (sometimes also called the late positive complex). The LPP has a centro-parietal scalp distribution and peaks between 500 and 800 ms. The LPP has been associated with sustained processing of emotional content of verbal stimuli as it has shown larger amplitude in emotional words compared to neutral words (Carretié et al., 2008; Hinojosa et al., 2010; Kanske and Kotz, 2007; Schacht and Sommer, 2009). Some studies have reported LPP amplitude differences between stimuli with positive vs. negative valence (Herbert et al., 2008, 2006; Kissler et al., 2009; Palazova et al., 2011) and have suggested that the LPP is involved when more controlled, explicit cognitive processes occur. In our study, the electrode-by-time analysis showed an electrode cluster at 400-500 ms—somewhat close to the time period of the LPP—that did not show statistical differences between sentiment condition (see Figs. 2E and 3E).

Regarding the relationship between EPN and LPP, Citron (2012) speculated that the early EPN component rather reflects the processing of arousal, while the LPP is involved in the processing of valence. However, the clear distinction between arousal and valence has been a source of debate. Lang et al. (1997) considered valence and arousal intrinsically associated. For example, emotionally valanced and neutral stimuli do not only differ along the arousal dimension, but also in terms of valence. Therefore, it has been suggested that the EPN effect can be seen as a more general "emotionality" effect, in which valence and arousal are integrated (Citron, 2012). Furthermore, our focus was to compare the processing of words with positive vs. negative sentiment. Although we cannot entirely exclude the possibility of small disparities in arousal related to reading words with positive vs. negative sentiment, these differences were likely to be small when compared to the strong differences in valence associated to words with positive vs. negative sentiment. Our study investigated differences in word sentiment for dichotomous categories (positive or negative). It is interesting to consider potential gradual differences between negative and positive sentiment. However, in our stimulus material there was an imbalance between the number of strongly positive/negative (19%) and moderately positive/negative (81%) trials. Hence, we cannot reliably model such gradual differences.

Our results indicated topographical differences in FRPs related to word-level sentiment. Other linguistic features may affect FRPs, for instance, word length and word frequency, which we addressed by our stimulus selection procedure and control analysis. Other properties of the linguistic material or reading behavior may have contributed to our results, such as the linguistic context at the phrase and sentence level or the word order (Hasson et al., 2015). Such behavioral-linguistic features may play an important role in naturalistic sentence reading, and future work should investigate the embedding of word- into phrase- and sentence-level processing using study designs tailored to address this research question. Recent studies have embarked on unraveling the cognitive and neural processes underlying lexico-semantic processing that relate word to phrase and sentence level processing (e.g., Yeshurun et al., 2017). For instance, in fMRI, Lerner et al. (2011) presented auditory stories scrambled at the sentence, phrase, or word level and found evidence for a hierarchical involvement of early sensory regions to more upstream areas such as the temporal and frontoparietal regions for phrase and sentence level processing. These results suggest that language processing involves multiple levels of processing at different temporal scales (Hasson and Honey, 2012). We could not perform such an analysis with our data because our participants always read meaningful sentences at their own pace. Nonetheless, we addressed this issue in supplementary analyses by using a GLMM including word-, phrase-, and sentence-level predictors of sentiments. This analysis showed no statistical differences for phrase and sentence predictors, while the same statistical difference between the positive and negative sentiment conditions was observed for the word sentiment predictor. These results should not be taken to imply that no such processing occurred. Rather, by focusing our analysis on FRPs for single words carefully selected to match for low-level oculomotor and linguistic properties, we focused our analysis mainly on the difference between positive and negative sentiment at the word level. Future studies should investigate the phrase and sentence level processing of linguistic sentiment.

In summary, our results provide evidence for the existence of a specific temporal window (224–304 ms after fixation onset) and a topographical difference (i.e., different underlying neural generators) for processing positive vs. negative sentiment of words during naturalistic reading. These results from FRP analyses were supported by independent decoding analyses of word sentiment from FRPs (discussed in the next section).

4.2. Sentiment decoding from FRPs

The second aim of our study was to assess whether there are any differences in FRPs for reading words with positive vs. negative sentiment that can be decoded from unseen (hold-out) FRP trials. It is important to note that the decoding analysis was independent of the FRP analysis and did not aim at developing algorithms for brain computer interfaces. That endeavor would require the capability of single-trial decoding in noisy environments, an application that is beyond the scope of our study. Instead, we conducted the decoding analysis on carefully selected fixation events, on cleaned FRPs, and on trial-average data, which served to increase the sensitivity of our analysis for detecting sentiment-related differences in FRPs. Our results showed a chance-level decoding performance for single-trial data. These results indicate that classifiers were unable to decode the word sentiment from unseen single-trial FRPs. These findings may be related to the low SNR generally observed in single-trial EEG and to the fact that participants performed sentence reading, which involves phrase- and sentence-level semantic processing that possibly interfere with the ability of classifiers to decode word-level sentiment from FRPs (Hasson et al., 2015). However, we observed an above-chance level decoding performance when decoding analysis was based on 20- or 40-trial averages across the same word-level sentiment. These results indicate that by increasing SNR via trial-averaging, the ability to decode the word sentiment from unseen FRP data was improved. These results are similar to the study by Tuckute et al. (2019), who observed an improvement of decoding performance from single-trial to 30-trial averages. These results highlight the importance of SNR for brain-based prediction of semantic processing. We note that in our study the level of decoding accuracy was low (0.60 mean accuracy for 40-trial averages) and based on large amounts of data (trial averages across 12,789 trials from 12 subjects). However, due to the limited number of 12 subjects in our study, future confirmatory studies are required to replicate and extend our results in larger subject samples. Given these limitations, naturalistic reading-related FRPs as used in our study may not be suited for single-trial brain-computer interface applications (Hebart and Baker, 2018). Our analyses focused on maximizing the sensitivity for detecting sentiment-related differences in FRPs by using carefully selected stimuli, cleaned FRPs, and trial-averages. This approach is very different from brain-computer interface research that aims at developing paradigms and classifiers that operate on single-trial data in noisy environments. Therefore, our classifiers are unlikely to succeed in brain-computer interface applications, and more work is needed to address these needs.

Previous research on sentiment analysis in natural language processing has traditionally been based on word-level and sentence-level linguistic features (e.g., Liu and Zhang, 2012; Pang and Lee, 2008). More recent work has used eye-movement data during reading as features to enhance sentiment decoding performance (Mishra et al., 2016; Tomanek et al., 2010; Xu et al., 2015). Only recently have EEG signals been considered for decoding sentiment polarity from words. For instance, Gu et al. (2014) recorded EEG from three humans during an experimentally-controlled word reading and mental imagery task. Classification analysis used the EEG features (electrode by time points) in a 1.5-sec temporal window after stimulus onset to predict the sentiment extracted from sentiment dictionaries. The classification performance ranged from 0.50 to 0.60 accuracy. Our results for sentiment decoding from single words are highly comparable to Gu et al. (2014) in that we observed an above-chance level predictive performance accuracy of 0.51–0.58, despite the fact that in our experiment sentiment processing occurred implicitly during natural reading; it was not related to an explicit task instruction as in Gu et al. (2014).

Sentiment decoding has also been performed on video material. For instance, Wang et al. (2014) and Nie et al. (2011) used EEG frequency components and different machine learning algorithms to decode the sentiment expressed in movie scenes. These studies focused on short movie clips and used as features the EEG segmented in 500-ms to 2-sec intervals. This approach provided classification accuracies between 0.50 and 0.81 across frequency bands and showed the best decoding performance in alpha (8–13 Hz) and beta band (13–30 Hz) frequency ranges. It is interesting that sentiment classification on word stimuli (Gu et al., 2014 and our study) showed lower classification accuracy than sentiment decoding from videos (Nie et al., 2011; Wang et al., 2014). This phenomenon may be related to the more engaging nature of a movie and the fixed timing of the visual stimuli as compared to linguistic stimuli processed in a self-paced fashion during naturalistic reading. This proposal is supported by previous studies showing that watching movie scenes evokes a high inter-subject correlation of electrical and functional neural activity (Gravens et al., 2011; Kauppi and Kauppi, 2010).

Data-driven classification analysis is sensitive to the amount of noise in the data (Delorme et al., 2007). Thus, computing trial averages leads to an increase of SNR that can substantially improve classification performance. In line with this idea, we found an improvement in the decoding accuracy from single-trial to 40-trial averages for both the SVM and logistic regression classifiers (Table 1). Tuckute et al. (2019) observed similar results for EEG-based decoding of image animacy from visual-evoked potentials using an SVM classifiers. These authors found an improvement in classification accuracy from single trial (mean accuracy of 0.54–0.61) to trial averages (mean accuracy of 0.50–0.90), which was higher than observed in our study—probably related to using picture stimuli instead of word reading data (as discussed above). Moreover, we found that the linear classifiers, SVM, and logistic regression performed better on our data than more complex architectures LSTM and DNN (see Supplementary Material). This result may be related to the comparably low number of trials in our data (12,789 trials) relative to the number of features (105 electrodes x 63 sampling points = 6615 features). For such data, SVMs have previously been shown to perform better than LSTMs (Arora et al., 2018; Güler and Koçer, 2005; Subasi, 2013). More complex architectures have been successfully applied to EEG data in other contexts (e.g., Gupta et al., 2019; Khurana et al., 2018).

4.3. Methodological considerations of naturalistic imaging for reading and sentiment processing

Traditional electrophysiological and functional neuroimaging studies typically consist of highly controlled experiments that vary among a few conditions. Controlled experiments are necessary in order to make accurate inferences; they enable the researcher to isolate a specific task while controlling for all other confounding variables. However, the stimuli for these conditions are artificially designed and, therefore, might result in conclusions that are not generalizable to how the brain works in real life (Wehbe, 2015). While controlled experiments allow the experimenter to make precise testable conclusions about the involved brain regions, they are not sufficient for understanding how complex cognitive tasks (e.g., naturalistic reading) are processed (Hasson and Honey, 2012). When studying language processing, for example, very few experiments have presented subjects with text encountered in everyday life. Instead, they have presented carefully designed stimuli. Based on these studies, it remains challenging to conclude how the multiple processes involved in reading work together and integrate, specifically when isolating one process at a time and keeping everything else constant. This issue might contribute to the current situation, where there is no convergence on a single model of how the brain extracts meaning from language (Friederici, 2012; Hagoort, 2013; Hickok and Pöppel, 2007). This deficit can be at least partly attributed to the difficulty of knowing how such a complex multicomponent process operates by isolating one of its subprocesses at a time (Wehbe, 2015).

In 1973, Newell had already highlighted the difficulty of combining the findings of a series of cognitive science experiments and advocated to select "a single complex task and do all of it" (Newell, 1973). This idea has also been highlighted in vivid detail by the fable of the six blind men and an elephant, in which blind men fail to come to an agreement on a perception of the elephant after each of them perceives only one body part of the elephant (Goldstein, 2009, p. 492). Thus, there is increasing interest in research of naturalistic human behavior because these conditions are more ecologically valid compared to traditional experimental research paradigms that use highly constrained stimulus material and frequent repetitions. Recent studies have subjects watching videos (Nishimoto et al., 2011; Petroni et al., 2018), solving math problems (Anderson et al., 2014), and listening to stories (Brennan et al., 2012; Broderick et al., 2018). Naturalistic experiments promise to deliver insights into human perceptual-cognitive decision making that not only provide a better approximation for identifying the cognitive and neural processes related to real-world human behavior, but can also be used to improve decision making of computer algorithms (e.g., for natural language processing).

In order to derive a more complete picture of the underlying neural processes of reading, increased research effort has been made to study naturalistic reading. In a series of simultaneous eye-tracking and fMRI naturalistic reading studies, neural correlates of the effects of word length, frequency, and predictability on brain responses during naturalistic reading have been identified (Desai et al., 2018; Henderson et al., 2016, 2015; Henderson and Choi, 2015; Schuster et al., 2016). There are several challenges associated with naturalistic reading. First, the common fMRI sequences typically acquire an image only every 2 s and measure a delayed smooth hemodynamic response. This function might be too slow for the dynamic process of reading at a natural pace and unable to identify the contributions of individual words or concepts to brain activity. However, new sophisticated methodological approaches (e.g., co-registering eye-movement or fast fMRI) have the ability to overcome these difficulties (e.g., Desai et al., 2018; Jones et al., 2001; Schuster et al., 2016; Yarkoni et al., 2008). Another issue is the fact that both fMRI and EEG are noisy imaging tools. Multiple repetitions are often necessary to produce a reliable representation of a cognitive process. Repetitions reduce stimulus diversity and decrease the ecological validity. In the present study, we avoided repetitions by presenting each sentence only once; we simultaneously achieved an increased SNR by computing averages across the word sentiment condition. Finally, another difficulty of naturalistic reading experiments is that various processes occur concomitantly during reading, including oculomotor behavior, visual processing, and sentiment processing. This fact makes it more difficult to separate activity patterns mediating linguistic information processing from areas mediating other co-occurring processes. In the present study, we chose to control and correct for various confounding parameters, such as linguistic properties as well uncontrolled eye movement, visual stimulation, and the embedding of currently read text into the lexico-semantic context of previous memory (e.g., Hasson et al., 2015). A particular concern for our study was that temporal overlap of high-amplitude visual-motor components precludes the detection of sentiment-related differences in FRPs (Dimigen et al., 2011; Ehinger and Dimigen, 2019; Smith and Kutas, 2015b). We therefore conducted analyses for cleaned and uncleaned FRPs by deconvolution modeling and found highly comparable statistical results. The absence of differences between the analysis may be related to the fact that we employed an event matching procedure where word fixation duration and therefore the temporal onset of subsequent fixations were matched across conditions. If one were to model the EEG by deconvolution modeling for unmatched conditions, or fewer trials from which to select the data, EEG deconvolution modeling may be of greater benefit. Alternatively, using deconvolution modeling to directly compare betas between different conditions may be applicable for between-subjects statistical analyses (Ehinger and Dimigen, 2019). This procedure has been previously used for modeling source-level activation from EEG (Brodbeck et al., 2018). Our results indicated that the FRP differences between positive and negative sentiment processing are not merely a function of high-amplitude visuo-motor activity. Rather, they reflect lexico-semantic processing of the word content. We note that the small sample size ($N = 12$) in our study limits the generalization ability of our results and did not allow us to investigate between-subjects random effects. A replication study based on our methods in a larger subject sample is desirable.

5. Conclusion

This study successfully identified the spatiotemporal neural dynamics of sentiment processing during naturalistic reading of English sentences. Combining high-density EEG and eye-tracking data, we showed that individual words of positive vs. negative sentiment evoke a consistent topographical difference in the FRPs at 224–304 ms after fixation onset. The FRP signal in this time period allowed decoding the word sentiment with an above chance-level performance, when considering FRP averages of 20 or more trials. Our results provide a proof of concept that the combination of state-of-the-art electrical neuroimaging and decoding-based analysis can serve to identify the neural dynamics of naturalistic stimulus processing in humans, which in turn can help to improve computer algorithms for natural language processing. This endeavor will advance our understanding of how the human brain extracts the meaning from written text under ecologically valid conditions.

Funding

This work was supported by the Swiss National Science Foundation [100014_175875].

Declaration of competing interest

The authors declare no competing interests.

CRediT authorship contribution statement

Christian Pfeiffer: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Nora Hollenstein:** Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing. **Ce Zhang:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition. **Nicolas Langer:** Conceptualization, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition.

Acknowledgments

We thank Marius Tröndle for help with data preparation and preprocessing, Benedikt Ehinger for feedback on deconvolution modeling, and Laura Wiles for proofreading.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2020.116934.

References

Manning, C.D., Manning, C.D., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press.

Abdullaev, Y.G., Posner, M.I., 1998. Event-related brain potential imaging of semantic encoding during processing single words. Neuroimage 7, 1–13.

Agarwal, B., Mittal, N., Bansal, P., Garg, S., 2015. Sentiment analysis using common-sense and context information. Comput. Intell. Neurosci. 2015, 715730.

Anderson, J.R., Lee, H.S., Fincham, J.M., 2014. Discovering the structure of mathematical problem solving. Neuroimage 97, 163–177.

Arora, A., Lin, J.-J., Gasperian, A., Maldjian, J., Stein, J., Kahana, M., Lega, B., 2018. Comparison of logistic regression, support vector machines, and deep learning classifiers for predicting memory encoding success using human intracranial EEG recordings. J. Neural. Eng. https://doi.org/10.1088/1741-2552/aae131.

Artoni, F., Delorme, A., Makeig, S., 2018. Applying dimension reduction to EEG data by Principal Component Analysis reduces the quality of its subsequent Independent Component decomposition. Neuroimage 175, 176–187.

Beymer, D., Russell, D.M., 2005. WebGazeAnalyzer: a system for capturing and analyzing web reading behavior using eye gaze. In: CHI '05 Extended Abstracts on Human Factors in Computing Systems, CHI EA '05. ACM, New York, NY, USA, pp. 1913–1916.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., Robbins, K.A., 2015. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. Front. Neuroinf. 9, 16.

Breiman, L., 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat. Sci. 16, 199–231.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D.J., Pylkkänen, L., 2012. Syntactic structure building in the anterior temporal lobe during natural story listening. Brain Lang. 120, 163–173.

Brodbeck, C., Presacco, A., Simon, J.Z., 2018. Neural source dynamics of brain responses to continuous stimuli: speech processing from acoustics to comprehension. Neuroimage 172, 162–174.

Broderick, M.P., Anderson, A.J., Di Liberto, G.M., Crosse, M.J., Lalor, E.C., 2018. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr. Biol. 28, 803–809 e3.

Brunet, D., Murray, M.M., Michel, C.M., 2011. Spatiotemporal analysis of multichannel EEG: CARTOOL. Comput. Intell. Neurosci. 2011, 813870.

Cambria, E., Schuller, B., Xia, Y., Havasi, C., 2013. New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. https://doi.org/10.1109/mis.2013.30.

Carretié, L., Hinojosa, J.A., Albert, J., López-Martín, S., De La Gándara, B.S., Igoa, J.M., Sotillo, M., 2008. Modulation of ongoing cognitive processes by emotionally intense words. Psychophysiology 45, 188–196.

Chanel, G., Kronegg, J., Grandjean, D., Pun, T., 2006. Emotion assessment: arousal evaluation using EEG's and peripheral physiological signals. Multimedia Content Representation, Classification and Security. https://doi.org/10.1007/11848035_70.

Citron, F.M.M., 2011. Neural Correlates of Emotion Word Processing: the Interaction between Emotional Valence and Arousal. University of Sussex.

Citron, F.M.M., 2012. Neural correlates of written emotion word processing: a review of recent electrophysiological and hemodynamic neuroimaging studies. Brain Lang. 122, 211–226.

Clifton, C., Ferreira, F., Henderson, J.M., Inhoff, A.W., Liversedge, S.P., Reichle, E.D., Schotter, E.R., 2016. Eye movements in reading and information processing: keith Rayner's 40 year legacy. J. Mem. Lang. https://doi.org/10.1016/j.jml.2015.07.004.

Currie, G., 1990. The nature of fiction. https://doi.org/10.1017/cbo9780511897498.

Delorme, A., Sejnowski, T., Makeig, S., 2007. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. Neuroimage 34, 1443–1449.

Desai, R.H., Choi, W., Henderson, J.M., 2018. Word frequency effects in naturalistic reading. Language. Cognit. Neurosci. https://doi.org/10.1080/23273798.2018.1527376.

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A.M., Kliegl, R., 2011. Coregistration of eye movements and EEG in natural reading: analyses and review. J. Exp. Psychol. Gen. 140, 552–572.

Dodge, R., 1901. The psychology of reading. Psychological Review. https://doi.org/10.1037/h0074611.

Ehinger, B.V., Dimigen, O., 2019. Unfold: an integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. https://doi.org/10.1101/360156.

Ehinger, B.V., Groß, K., Ibs, I., König, P., 2019. A new comprehensive eye-tracking test battery concurrently evaluating the Pupil Labs glasses and the EyeLink 1000. PeerJ 7, e7086.

Finke, A., Essig, K., Marchioro, G., Ritter, H., 2016. Toward FRP-based brain-machine interfaces—single-trial classification of fixation-related potentials. PloS One. https://doi.org/10.1371/journal.pone.0146848.

Frank, S.L., Otten, L.J., Galli, G., Vigliocco, G., 2013. Word Surprisal Predicts N400 Amplitude during Reading.

Frey, A., Lemaire, B., Vercueil, L., Guérin-Dugué, A., 2018. An eye fixation-related potential study in two reading tasks: reading to memorize and reading to make a decision. Brain Topogr. 31, 640–660.

Friederici, A.D., 2012. The cortical language circuit: from auditory perception to sentence comprehension. Trends Cognit. Sci. 16, 262–268.

Goldberger, J., Hinton, G.E., Roweis, S.T., Salakhutdinov, R.R., 2005. Neighbourhood components analysis. In: Saul, L.K., Weiss, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems 17. MIT Press, pp. 513–520.

Goldstein, E.B., 2009. Encyclopedia of Perception. SAGE Publications.

Grainger, J., Holcomb, P., 2010. Neural constraints on a functional architecture for word recognition. The Neural Basis of Reading. https://doi.org/10.1093/acprof:oso/9780195300369.003.0001.

Gravens, L., Price, T., Harmon-Jones, E., 2011. Contralateral Motor Cortex EEG Mirror Neuron Activity to Watching Motor Behavior. PsycEXTRA Dataset. https://doi.org/10.1037/e634112013-534.

Gu, Y., Celli, F., Steinberger, J., Anderson, A.J., Poesio, M., Strapparava, C., Murphy, B., 2014. Using brain data for sentiment analysis. JLCL 29, 79–94.

Guérin-Dugué, A., Roy, R.N., Kristensen, E., Rivet, B., Vercueil, L., Tcherkassof, A., 2018. Temporal dynamics of natural static emotional facial expressions decoding: a study using event- and eye fixation-related potentials. Front. Psychol. 9, 1190.

Güler, N.F., Koçer, S., 2005. Classification of EMG signals using PCA and FFT. J. Med. Syst. https://doi.org/10.1007/s10916-005-5184-7.

Gupta, A., Sahu, H., Nanecha, N., Kumar, P., Roy, P.P., Chang, V., 2019. Enhancing text using emotion detected from EEG signals. Int. J. Grid Util. Comput. 17, 325–340.

Hagoort, P., 2013. MUC (memory, unification, control) and beyond. Front. Psychol. 4, 416.

Hara, T., Mochihashi, D., Kano, Y., Aizawa, A., 2012. Predicting word fixations in text with a CRF model for capturing general reading strategies among readers. In: Proceedings of the First Workshop on Eye-Tracking and Natural Language Processing, pp. 55–70.

Hasson, U., Honey, C.J., 2012. Future trends in Neuroimaging: neural processes as expressed within real-life contexts. Neuroimage 62, 1272–1278.

Hasson, U., Chen, J., Honey, C.J., 2015. Hierarchical process memory: memory as an integral component of information processing. Trends Cognit. Sci. 19, 304–313.

Hauk, O., Pulvermüller, F., 2004. Effects of word length and frequency on the human event-related potential. Clin. Neurophysiol. 115, 1090–1103.

Hebart, M.N., Baker, C.I., 2018. Deconstructing multivariate decoding for the study of brain function. Neuroimage 180, 4–18.

Henderson, J.M., Choi, W., 2015. Neural correlates of fixation duration during real-world scene viewing: evidence from fixation-related (FIRE) fMRI. J. Cognit. Neurosci. 27, 1137–1145.

Henderson, J.M., Choi, W., Luke, S.G., Desai, R.H., 2015. Neural correlates of fixation duration in natural reading: evidence from fixation-related fMRI. Neuroimage 119, 390–397.

Henderson, J.M., Choi, W., Lowder, M.W., Ferreira, F., 2016. Language structure in the brain: a fixation-related fMRI study of syntactic surprisal in reading. Neuroimage. https://doi.org/10.1016/j.neuroimage.2016.02.050.

Herbert, C., Kissler, J., Junghöfer, M., Peyk, P., Rockstroh, B., 2006. Processing of emotional adjectives: evidence from startle EMG and ERPs. Psychophysiology 43, 197–206.

Herbert, C., Junghofer, M., Kissler, J., 2008. Event related potentials to emotional adjectives during reading. Psychophysiology 45, 487–498.

Hickok, G., Pöppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402.

Hillyard, S.A., Kutas, M., 2002. Event-related potentials and magnetic fields in the human brain. In: Neuropsychopharmacology: the Fifth Generation of Progress. Lippincott, Williams, and Wilkins, Baltimore.

Hinojosa, J.A., Méndez-Bértolo, C., Pozo, M.A., 2010. Looking at emotional words is not the same as reading emotional words: behavioral and neural correlates. Psychophysiology 47, 748–757.

Hollenstein, N., Rotsztejn, J., Tröndle, M., Pedroni, A., Zhang, C., Langer, N., 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. Sci Data 5, 180291.

Jones, R.A., Brookes, J.A., Moonen, C.T.W., 2001. Ultra-fast fMRI. Functional Magnetic Resonance Imaging. https://doi.org/10.1093/acprof:oso/9780192630711.003.0004.

Kanske, P., Kotz, S.A., 2007. Concreteness in emotional words: ERP evidence from a hemifield study. Brain Res. 1148, 138–148.

Kauppi, Kauppi, 2010. Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. Front. Neuroinf. https://doi.org/10.3389/fninf.2010.00005.

Khurana, V., Kumar, P., Saini, R., Roy, P.P., 2018. EEG based word familiarity using features and frequency bands combination. Cognit. Syst. Res. 49, 33–48.

Kissler, J., Assadollahi, R., Herbert, C., 2006. Emotional and semantic networks in visual word processing: insights from ERP studies. Prog. Brain Res. 156, 147–183.

Kissler, J., Herbert, C., Peyk, P., Junghofer, M., 2007. Buzzwords: early cortical responses to emotional words during reading. Psychol. Sci. 18, 475–480.

Kissler, J., Herbert, C., Winkler, I., Junghofer, M., 2009. Emotion and attention in visual word processing: an ERP study. Biol. Psychol. 80, 75–83.

Kliegl, R., Dambacher, M., Dimigen, O., Jacobs, A.M., Sommer, W., 2012. Eye movements and brain electric potentials during reading. Psychol. Res. 76, 145–158.

König, T., Melie-García, L., 2010. A method to determine the presence of averaged event-related fields using randomization tests. Brain Topogr. 23, 233–242.

Kononenko, I., Kukar, M., 2007. Machine learning basics. Machine learning and data mining. https://doi.org/10.1533/9780857099440.59.

Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu. Rev. Psychol. 62, 621–647.

Lang, P.J., Bradley, M.M., Cuthbert, B.N., Others, 1997. Motivated attention: affect, activation, and action. Attention and orienting: Sensory and motivational processes 97, 135.

Langer, N., Pedroni, A., Gianotti, L.R.R., Hänggi, J., Knoch, D., Jäncke, L., 2012. Functional brain network efficiency predicts intelligence. Hum. Brain Mapp. 33, 1393–1406.

Lehmann, D., Skrandies, W., 1980. Reference-free identification of components of checkerboard-evoked multichannel potential fields. Electroencephalogr. Clin. Neurophysiol. 48, 609–621.

Lemhöfer, K., Broersma, M., 2012. Introducing LexTALE: a quick and valid lexical test for advanced Learners of English. Behav. Res. Methods 44, 325–343.

Lerner, Y., Honey, C.J., Silbert, L.J., Hasson, U., 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci. 31, 2906–2915.

Li, M., Lu, B.-L., 2009. Emotion classification based on gamma-band EEG. Conf. Proc. IEEE Eng. Med. Biol. Soc. 2009, 1323–1326.

Liu, B., Zhang, L., 2012. A Survey of Opinion Mining and Sentiment Analysis. Mining Text Data. https://doi.org/10.1007/978-1-4614-3223-4_13.

Loberg, O., Hautala, J., Hämäläinen, J.A., Leppänen, P.H.T., 2018. Semantic anomaly detection in school-aged children during natural sentence reading - a study of fixation-related brain potentials. PloS One 13, e0209741.

Mamidi, R., Miller, M., Banerjee, T., Romine, W., Sheth, A., 2019. Identifying key topics bearing negative sentiment on twitter: insights concerning the 2015-2016 zika epidemic. JMIR Public Health Surveill 5, e11036.

Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164, 177–190.

Martín-Loeches, M., 2007. The gate for reading: reflections on the recognition potential. Brain Res. Rev. 53, 89–97.

Mason, R.A., Just, M.A., 2006. Neuroimaging contributions to the understanding of discourse processes. Handbook of Psycholinguistics. https://doi.org/10.1016/b978-012369374-7/50020-1.

Matusz, P.J., Thelen, A., Amrein, S., Geiser, E., Anken, J., Murray, M.M., 2015. The role of auditory cortices in the retrieval of single-trial auditory-visual object memories. Eur. J. Neurosci. 41, 699–708.

Mensen, A., Khatami, R., 2013. Advanced EEG analysis using threshold-free cluster-enhancement and non-parametric statistics. Neuroimage 67, 111–118.

Mishra, A., Kanojia, D., Nagar, S., Dey, K., Bhattacharyya, P., 2016. Leveraging cognitive features for sentiment analysis. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. https://doi.org/10.18653/v1/k16-1016.

Mullen, T., 2012. CleanLine EEGLAB Plugin. Neuroimaging Informatics Toolsand Resources Clearinghouse (NITRC), San Diego, CA.

Murray, M.M., Brunet, D., Michel, C.M., 2008. Topographic ERP analyses: a step-by-step tutorial review. Brain Topogr. 20, 249–264.

Newell, A., 1973. YOU CAN'T play 20 questions with nature and WIN: projective comments ON the papers OF this symposium. Visual Information Processing. https://doi.org/10.1016/b978-0-12-170150-5.50012-3.

Nie, D., Wang, X.-W., Shi, L.-C., Lu, B.-L., 2011. EEG-based emotion recognition during watching movies. In: 2011 5th International IEEE/EMBS Conference on Neural Engineering. https://doi.org/10.1109/ner.2011.5910636.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. 21, 1641–1646.

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.-M., 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. 2011, 156869.

Palazova, M., Mantwill, K., Sommer, W., Schacht, A., 2011. Are effects of emotion in single words non-lexical? Evidence from event-related brain potentials. Neuropsychologia 49, 2766–2775.

Pang, B., Lee, L., 2008. Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval. https://doi.org/10.1561/1500000011.

Pedroni, A., Bahreini, A., Langer, N., 2019. AUTOMAGIC: standardized preprocessing of big EEG data. https://doi.org/10.1101/460469.

Petroni, A., Cohen, S.S., Ai, L., Langer, N., Henin, S., Vanderwal, T., Milham, M.P., Parra, L.C., 2018. The variability of neural responses to naturalistic videos change with age and sex. eNeuro 5. https://doi.org/10.1523/ENEURO.0244-17.2017.

Rämä, P., Baccino, T., 2010. Eye fixation-related potentials (EFRPs) during object identification. Vis. Neurosci. 27, 187–192.

Raudonis, V., Dervinis, G., Vilkauskas, A., Paulauskaite, A., Kersulyte, G., 2013. Evaluation of human emotion from eye motions. Int. J. Adv. Comput. Sci. Appl. https://doi.org/10.14569/ijacsa.2013.040812.

Rayner, K., Clifton Jr., C., 2009. Language processing in reading and speech perception is fast and incremental: implications for event-related potential research. Biol. Psychol. 80, 4–9.

Salmelin, R., 2007. Clinical neurophysiology of language: the MEG approach. Clin. Neurophysiol. 118, 237–254.

Sato, N., Mizuhara, H., 2018. Successful encoding during natural reading is associated with fixation-related potentials and large-scale network deactivation. eNeuro 5. https://doi.org/10.1523/ENEURO.0122-18.2018.

Schacht, A., Sommer, W., 2009. Time course and task dependence of emotion effects in word processing. Cognit. Affect Behav. Neurosci. 9, 28–43.

Schupp, H.T., Junghöfer, M., Weike, A.I., Hamm, A.O., 2004. The selective processing of briefly presented affective pictures: an ERP analysis. Psychophysiology 41, 441–449.

Schuster, S., Hawelka, S., Hutzler, F., Kronbichler, M., Richlan, F., 2016. Words in context: the effects of length, frequency, and predictability on brain responses during natural reading. Cerebr. Cortex 26, 3889–3904.

Scott, G.G., O'Donnell, P.J., Leuthold, H., Sereno, S.C., 2009. Early emotion word processing: evidence from event-related potentials. Biol. Psychol. 80, 95–104.

Sereno, S.C., Rayner, K., 2003. Measuring word recognition in reading: eye movements and event-related potentials. Trends Cognit. Sci. 7, 489–493.

Simola, J., Torniainen, J., Moisala, M., Kivikangas, M., Krause, C.M., 2013. Eye movement related brain responses to emotional scenes during free viewing. Front. Syst. Neurosci. 7, 41.

Singmann, H., Kellen, D., 2017. An introduction to mixed models for experimental psychology. In: New Methods in Neuroscience and Cognitive Psychology. Psychology Press Hove.

Smith, N.J., Kutas, M., 2015a. Regression-based estimation of ERP waveforms: I. The rERP framework. Psychophysiology 52, 157–168.

Smith, N.J., Kutas, M., 2015b. Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. Psychophysiology 52, 169–181.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C., 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631–1642.

Subasi, A., 2013. Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders. Comput. Biol. Med. 43, 576–586.

Tateosian, L.G., Glatz, M., Shukunobe, M., Chopra, P., 2015. GazeGIS: a gaze-based reading and dynamic geographic information system. In: Workshop on Eye Tracking and Visualization. Springer, pp. 129–147.

The Oxford Handbook of Neurolinguistics, 2019. https://doi.org/10.1093/oxfordhb/9780190672027.001.0001.

Thibadeau, R., Just, M., Carpenter, P., 1980. Real reading behavior. In: Proceedings of the 18th Annual Meeting on Association for Computational Linguistics. https://doi.org/10.3115/981436.981486.

Tomanek, K., Hahn, U., Lohmann, S., Ziegler, J., 2010. A cognitive cost model of annotations based on eye-tracking data. In: Proceedings of the 48th.

Tuckute, G., Hansen, S.T., Pedersen, N., Steenstrup, D., Hansen, L.K., 2019. Single-trial decoding of scalp EEG under natural conditions. Comput. Intell. Neurosci. 2019, 9210785.

Tzovara, A., Murray, M.M., Michel, C.M., De Lucia, M., 2012. A tutorial review of electrical neuroimaging from group-average to single-trial event-related potentials. Dev. Neuropsychol. 37, 518–544.

Wang, X.-W., Nie, D., Lu, B.-L., 2014. Emotional state classification from EEG data using machine learning approach. Neurocomputing. https://doi.org/10.1016/j.neucom.2013.06.046.

Wehbe, L., 2015. The Time and Location of Natural Reading Processes in the Brain.

Winkler, I., Haufe, S., Tangermann, M., 2011. Automatic classification of artifactual ICA-components for artifact removal in EEG signals. Behav. Brain Funct. 7, 30.

Winkler, I., Brandl, S., Horn, F., Waldburger, E., Allefeld, C., Tangermann, M., 2014. Robust artifactual independent component classification for BCI practitioners. J. Neural. Eng. 11, 035013.

Xu, P., Ehinger, K.A., Zhang, Y., Finkelstein, A., Kulkarni, S.R., Xiao, J., 2015. TurkerGaze: Crowdsourcing Saliency with Webcam Based Eye Tracking (arXiv [cs.CV]).

Yang, H., Willis, A., de Roeck, A., Nuseibeh, B., 2012. A hybrid model for automatic emotion recognition in suicide notes. Biomed. Inf. Insights 5, 17–30.

Yarkoni, T., Westfall, J., 2017. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect. Psychol. Sci. 12, 1100–1122.

Yarkoni, T., Speer, N.K., Balota, D.A., McAvoy, M.P., Zacks, J.M., 2008. Pictures of a thousand words: investigating the neural mechanisms of reading with extremely rapid event-related fMRI. Neuroimage 42, 973–987.

Yeshurun, Y., Nguyen, M., Hasson, U., 2017. Amplification of local changes along the timescale processing hierarchy. Proc. Natl. Acad. Sci. U.S.A. 114, 9475–9480.