Language Resources and Evaluation (2024) 58:641–657
https://doi.org/10.1007/s10579-023-09684-x

ORIGINAL PAPER

An eye-tracking-with-EEG coregistration corpus of narrative sentences

Stefan L. Frank[1] · **Anna Aumeistere**[2]

Accepted: 2 August 2023 / Published online: 29 August 2023

Abstract
We present the Radboud Coregistration Corpus of Narrative Sentences (RaCCooNS), the first freely available corpus of eye-tracking-with-EEG data collected while participants read narrative sentences in Dutch. The corpus is intended for studying human sentence comprehension and for evaluating the cognitive validity of computational language models. RaCCooNS contains data from 37 participants (3 of which eye tracking only) reading 200 Dutch sentences each. Less predictable words resulted in significantly longer reading times and larger N400 sizes, replicating well-known surprisal effects in eye tracking and EEG simultaneously. We release the raw eye-tracking data, the preprocessed eye-tracking data at the fixation, word, and trial levels, the raw EEG after merger with eye-tracking data, and the preprocessed EEG data both before and after ICA-based ocular artifact correction.

Keywords Narrative sentence reading · Eye tracking · Electroencephalography · Fixated-related potentials · Dutch · Surprisal effects

1 Introduction

Psycholinguistic studies of sentence and discourse comprehension traditionally rely on experiments with a small number of hand-crafted stimuli in which just one or two factors are manipulated and that contain only a single critical word or clause. More recently, however, it has become increasingly common to collect measures of human cognitive or neural processing over *all* words of a collection of (semi-)*natural*

✉ Stefan L. Frank
stefan.frank@ru.nl

Anna Aumeistere
anna.aumeistere@ru.nl

[1] Centre for Language Studies, Radboud University, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands

[2] Donders Centre for Cognitive Neuroimaging, Radboud University, Kapittelweg 29, 6525 EN Nijmegen, The Netherlands

sentences or texts. In many cases, these measures are then analysed by comparing them to the quantitative predictions of computational language models that process the same items; or conversely (from a computational linguistic perspective) the human processing measures are used to evaluate and compare models' cognitive validity.

The Dundee Corpus (Kennedy & Pynte, 2005) was possibly the first corpus of this kind. It comprises British newspaper editorials that were read by ten participants while their eye movements were tracked. The word-reading times from this corpus were used for model evaluation by Demberg and Keller (2008) and Frank and Bod (2011), among several others. Other examples of sentence or text corpora with human behavioural data are the Potsdam Sentence Corpus (Kliegl et al., 2006), which consists of isolated German sentences that were designed to include various syntactic structures, and comes with eye-tracking data; the UCL Corpus (Frank et al., 2013) of English sentences that were sampled from novels, and has both eye-tracking data and self-paced reading times; the Natural Stories Corpus (Futrell et al., 2021) of self-paced reading times on English narratives that were adapted to cover a range of syntactic complexities; and the Ghent Eye-Tracking Corpus (GECO; Cop et al., 2017) with eye-movements recorded while participants read an entire Agatha Christie novel in either English or Dutch. The recently published Multilingual Eye-Movements Corpus (MECO; Siegelman et al., 2022) is particularly interesting because it includes materials in 13 different languages.

There are also corpora that come with neuroimaging (rather than behavioural) data collected during language comprehension. For example, in studies by Wehbe and colleagues, participants read one chapter of a Harry Potter book while their brain activity was recorded using functional magnetic resonance imaging (fMRI; Wehbe et al., 2014) or magnetoencephalography (MEG; Wehbe et al., 2014). Another series of studies had participants listen to spoken Dutch narratives (audio book excerpts) while in an fMRI (Lopopolo et al., 2018) or MEG (Armeni et al., 2019, 2022) scanner.

The sentences of the above-mentioned Potsdam and UCL eye-tracking corpora were also presented to participants in electroencephalography (EEG) studies (Dambacher et al., 2006; Frank et al., 2015) using the standard 'rapid serial visual presentation' (RSVP) approach where a sentence's words are flashed sequentially at a fixed location in order to prevent eye movements. Thus, EEG and eye-movement data are available on the same set of sentences, allowing for direct comparison between language models' ability to predict the behavioural and electrophysiological measures. Unfortunately, such a comparison is hampered by the fact that the two data types were recorded from different sets of participants and in different experimental sessions.

This problem can be solved by *coregistration*, that is, the simultaneous recording of eye movements and EEG, thereby allowing participants to read naturally during EEG recording. Natural reading and RSVP differ in early orthographic processing (Kornrumpf et al., 2016; Nárai et al., 2022) and evoke different cognitive processes (Metzner et al., 2015), highlighting the importance of natural reading in

EEG studies.[1] Although eye movements cause strong ocular artifacts in the EEG signal, techniques are available for removing these to obtain a fairly clean EEG signal (Dimigen et al., 2011).

As far as we are aware, the only currently available corpus of eye-tracking-with-EEG data is the Zurich Corpus (ZuCo; Hollenstein et al., 2018). The ZuCo stimuli were isolated sentences extracted from expository texts, namely movie reviews and Wikipedia pages about well-known individuals, in English. A potential drawback of using such items is that they can be hard to interpret out of context, for example because of a pronoun that refers to a specific but unnamed individual (as in the ZuCo item 'He won a Nobel Prize in Chemistry in 1928'). Moreover, when the specific movie or individual is mentioned, readers' understanding and appreciation of the statement will surely depend on their knowledge and opinion about the movie/individual. Such knowledge- and opinion-dependent processing factors are difficult to capture in a computational language model and increase by-participant variance in the data.

In the eye-tracking-with-EEG data set we present here, we took a slightly different approach and followed the UCL Corpus method of extracting from narrative texts individual sentences that are comprehensible out of context. From the eye-movement corrected EEG signal, we derive the voltages time-locked to fixation onset, known as fixation-related potentials (FRPs). These FRPs show positive- and negative-going deflections, called components, the earliest of which are indicative of the fixation itself and of early orthographic processing. Our main interest here is in the so-called N400 component, a negative-going deflection on central electrodes that tends to peak at around 400 ms after the appearance of a word and is larger for words that are semantically implausible (Kutas & Hillyard, 1980) or less predictable (Kutas & Hillyard, 1984).

We analyse and validate the data by comparing word-reading times and N400 sizes to word-surprisal values estimated by a simple computational language model on the same sentence stimuli. A word's surprisal is the negative log-transformed probability of its occurrence given the sentence so far (Hale 2001; Levy 2008) and earlier research has repeatedly shown that words with higher surprisal take longer to read (Goodkind & Bicknell, 2018; Monsalve et al., 2012; Smith & Levy, 2013) and yield larger N400 (Frank et al., 2015; Michaelov & Bergen, 2020). We find the same effects in our data, after correcting for word frequency, length, and position in the sentence.

We call our data set the Radboud Coregistration Corpus on Narrative Sentences, or *RaCCooNS* for short. In the remainder of this paper, we will discuss stimuli selection, participant properties, procedural details, and data recording, preprocessing, and analysis. Next, we show how word surprisal relates to word-reading time, N400 size, and the shapes of the FRPs. Finally, we discuss the quality of ocular artifact correction and possible avenues for future research. The RaCCooNS data set is freely downloadable (see Supplementary information section).

[1] Weiss et al. (2016), however, found that 'the time course of orthographic processing during natural reading might be remarkably similar to that found during word reading with fixed gaze.' (p. 12).

2 Methods

2.1 Stimuli

All 200 sentence stimuli were taken from the 'books' section (excluding biblical texts) of the SONAR-500 Dutch corpus (Oostdijk et al., 2014), according to the following procedure:

1. The 20,000 most frequent word types were selected from the first slice of the NLCOW2014 Dutch web corpus (Schäfer, 2015; Schäfer & Bildhauer, 2012).
2. From the SONAR-500 texts, all 266,132 sentences were selected that:
   - Contain only word types from the 20,000 most frequent word list, or any of the following punctuation markers: period, comma, question mark, exclamation mark;
   - Are at least 5 words long (not counting punctuation);
   - Are at most 30 tokens (i.e., words or punctuation markers) long.
3. From these sentences, the subset containing at least three content words (not including the last word, i.e., the pre-final token under the assumption that the last token is punctuation) was selected. Content words are adjectives, verbs, and nouns, as determined by the Frog part-of-speech tagger (Van der Sloot et al., 2018)
4. From the remaining 175,948 sentences, 400 items were identified that
   - Form a grammatical sentence;
   - Are comprehensible outside of their context without requiring uncommon world knowledge;
   - Do not contain possibly offensive words nor describe potentially upsetting events (e.g., extreme violence).
5. From these 400 items, 200 were selected to represent a wide range of words and constructions. Five other sentences were chosen as practice items.
6. The occasional exclamation mark was replaced by a period.

The 200 stimuli sentences comprised 2783 word tokens of 1015 word types (excluding punctuation). One hundred of the stimuli sentences and three of the practice items were paired with a yes/no comprehension question. Correct answers were divided equally between 'yes' and 'no'.

All stimuli, questions, and correct answers can be found in the shared folder `Stimuli`, which also contains the list of 20,000 most frequent words and a data frame (tab-separated values text file) with information about each word in the sentence stimuli (for details, see the `README.md` file in the same folder).

Table 1 Demographic information of included participants

	Eye-tracking	EEG
N	37	34
Age		
Range	18–49	18–49
Mean	26.2	25.8
Gender		
Male	11	10
Female	26	24
Handedness		
Left	5	5
Right	31	28
Unclear	1	1
Dominant eye		
Left	10	9
Right	27	25

All participants' EEG data has corresponding eye-tracking data, but there are three participants with eye-tracking data only

2.2 Participants

Participants were recruited via Radboud University's Research Participation System. Forty-three native Dutch speaking participants were tested, none of which reported experiencing reading difficulties (e.g., due to visual impairment or dyslexia). The eye-tracking and EEG data from six participants were discarded because of technical issues. The EEG data from an additional three participants was discarded because of an extremely noisy signal. Table 1 lists basic demographic information for the remaining participants. Information about individual participants is in the tab-separated text table `ET_Participants.tsv` in the shared folder `eyetracking` (for details, see the `README.md` file in the same folder).

2.3 Procedure

All experimental protocols were approved by Radboud University's Ethics Assessment Committee Humanities (application nr. 1036). All procedures were carried out in accordance with the relevant guidelines and regulations. Informed consent was obtained from all participants.

Participants were tested individually in a soundproof booth at the Centre for Language Studies experiment lab (Radboud University, Nijmegen). Their dominant eye was determined using the Miles test and they were fitted with an EEG cap and four electrooculography (EOG) electrodes (see Sect. 2.5 for electrode locations). A conduction gel was injected into the electrodes to ensure high

conductivity between the electrode and the scalp. The impedance was considered sufficient if it was below 20 kΩ.

Participants were then asked to put their chin on the eye-tracker chin rest and to find a comfortable position to remain in for the duration of the experiment. Their eyes were at a distance of 105.5 cm from the top of the monitor and 108 cm from the bottom.[2] The experiment description, which appeared on the monitor, instructed participants to read the sentences like they would normally read a book.

After the participants were given the opportunity to ask questions, a nine-point eye-tracking calibration was performed and the five practice sentences were presented, followed by a repetition of the calibration. Next, the 200 experimental sentences were presented in randomized order.

Participants were encouraged to take a break after every 33 or 34 sentences, which was followed by another nine-point calibration. Upon finishing the reading task, participants were asked about their gender, age, and handedness, and received a €20 gift voucher or course credit. The duration of an experiment session was between 100 and 120 min.

2.4 Stimulus presentation

Sentences were presented left aligned with a 40-pixel margin on the right-hand side. The 19 longest sentences were split into two left-aligned lines[3] with a 76-pixel distance between the two lines. Stimuli were presented in Courier font, with a 16-pixel character width (0.26° visual angle). The screen resolution was set to 1920 × 1018 and the top-left coordinates of each stimulus sentence were (62, 352) so that it appeared at approximately one-third from the top of the screen. Because we used a fixed-width font, the character width, start coordinates, and (if applicable) line break location suffice to reconstruct the screen position of each character.

Each sentence was preceded by a fixation cross located where the beginning of the first word would appear. After the fixation cross had been visible for 500 ms, the eye-tracker waited for the occurrence of a fixation in a 25 × 25 pixel square centered on the fixation cross. As soon as such a fixation was automatically detected by the built-in function of SR Research Experiment Builder, the fixation cross was replaced by the stimulus sentence. Failure to detect a fixation on the fixation cross would have resulted in recalibration, but in practice this was never required.

2.5 Data recording

Eye movements were recorded at a time resolution of 1000 Hz, using an SR Research EyeLink 1000+ desk-mounted eye tracker.

[2] The monitor was a BenQ type XL2430T with a pixel dot pitch value of 0.276 mm.

[3] Line-break locations are indicated by `\n` in `stimuli_with_linebreaks.txt` in the shared folder `Stimuli`.

We used an ActiCHamp EEG system with 28 EEG channels (locations Fp2, Fz, F3, F7, F4, F8, FT9, T7, T8, FC5, FC1, FC6, FC2, C3, C4, CP5, CP1, CP6, CP2, Pz, P3, P7, P4, P8, Oz, O1, O2; and Cz as reference). EOG electrodes were located under and above the left eye, and at left and right outer canthi. EEG and EOG were recorded at 500 Hz using Brain Vision Recorder acquisition software.

Synchronization between the eye-tracker and EEG signals was established by sending simultaneous EEG triggers and eye-tracker messages at each stimulus sentence onset and at the start and end of recording.

2.6 Preprocessing

2.6.1 Eye tracking

A research assistant inspected all trials for monotonous change in vertical fixation location resulting in fixations consistently falling below or above the area-of-interests that SR Research Data Viewer automatically assigned to the words. The research assistant then used Data Viewer to correct these vertical drifts by vertically re-assigning fixations to words. Fixations were never moved horizontally. A trial was marked for rejection when fixations could not reliable be attributed to words (according to the subjective opinion of the research assistant) or if no fixation on the sentence was recorded due to track loss. In total, 309 of 7209 trials (4.29%) were rejected.

We make available the raw eye-tracking data (converted from edf to ascii format), as well as pre-processed participant/session-level, trial-level, word-level, and fixation-level data as tab-separated text tables. Word-level data includes four reading-time measures (Rayner, 1998): the duration of the first fixation on the word, the first-pass duration (amount of time from the onset of the first fixation until the offset of the last consecutive fixation on the word), the regression-path duration (amount of time from the onset of the first fixation on the word until the offset of the last fixation that is followed by a fixation on a later word in the sentence), and the total reading time (sum of durations of all fixations on the word). A list of rejected trials is provided in `bad_data.txt` in the main shared folder. For further details of the shared eye-tracking data, see the `README.md` file in the shared folder `eyetracking`.

2.6.2 EEG

2.6.2.1 Merger with eye-tracking data We used the MATLAB EYE-EEG toolbox (v0.85) by Dimigen et al. (2011) for EEGLAB (v2021.0; Delorme & Makeig, 2004) to merge the eye-tracker and EEG data. The toolbox function `synchronize` first performs a linear interpolation between the start- and end-of-recording eye-tracker messages and EEG triggers, in order to match the number of samples between the two signals. Next, it detects the sentence-onset events shared between the eye-tracker and EEG, and measures the latency difference between the shared events. All stimuli onsets were found to be well synchronised between the EEG and eye-tracker signals,

with no more than 1 sample (2 ms) difference, except for the single occurrence of a 2-sample asynchrony.

EEG data was rejected when the tracker coordinates fell outside of a rectangular window with top left pixel coordinates (40, 300) and bottom right coordinates (1880, 500); these mostly correspond to track losses, often due to blinks. Data recorded during the 100 ms leading up to or following out-of-bound gaze coordinates was also rejected.

2.6.2.2 Ocular artifact correction Data from an individual electrode was excluded when visual inspection revealed it to be faulty (not transmitting a signal) or to show nothing but very strong noise; this was the case for one electrode for 7 participants, and for 3 electrodes for a single participant (see `bad_data.txt` for details). We did not use interpolation to estimate the missing electrode data. EEG data were re-referenced to the average over the remaining EEG channels, turning the original Cz reference into a proper EEG channel.

We then performed ICA-based ocular artifacts removal, following Dimigen (2020) as closely as possible. First, a 4 Hz passband edge high-pass filter (but no low-pass filter) was applied. Stretches of data from 0.2 s before until 2.9 s after every even-numbered stimulus appearance event were selected as ICA training data. Spike potentials caused by saccades are overweighted in the training data by copying (appending) the data 20 ms before until 10 ms after a saccade; this was repeated until the total amount of data increased by 50%. Next, we ran fastICA with components estimated in parallel, using the `FastICA` package for MATLAB.[4] We then removed ICA components whose variance during saccades was over 10% higher than variance during fixations (a method originally proposed by Plöchl et al. 2012). The results are stored in a single EEGLAB struct variable per participant.

2.6.2.3 Fixation-related potentials FRPs were extracted for all non-excluded EEG electrodes and for all first fixations on each word. Following Dimigen et al. (2011), we excluded fixations that start within 700 ms from trial onset and fixations rejected in the eye-tracking data.

The N400 size in response to the first fixation on a word was defined as the average FRP over electrodes Cz, C3, C4, CP1, CP2, Pz, P3, and P4 (based on Fig. 3 of Dimigen et al., 2011) in a time window between 250 ms and 450 ms from the start of fixation. This time window was based on Dimigen et al. (2011) who report slightly earlier N400 effects than the standard 300–500 ms window in EEG reading studies using RSVP. The N400 baseline was taken to be the average voltage over the same set of electrodes in the 100 ms leading up to the fixation.

We make the following EEG data available (for details, see the `README.md` file in the shared folder `EEG`):

[4] https://research.ics.aalto.fi/ica/fastica/.

Table 2 Fixed effects of interest from regression model fitted to log-transformed first-pass reading times

Predictor	b (× 1000)	SE (× 1000)	z	p
Surprisal	9.9	1.4	6.89	< .00001
Log wordfreq	3.0	4.3	0.70	0.48
Word length	16.4	2.8	5.87	< .00001

- The merged raw EEG and eye-tracker data: one MATLAB file per participant in folder `EEG/Merged`.
- The preprocessed EEG data: one MATLAB file per participant in folder `EEG/Preprocessed`.
- The FRPs for all non-rejected fixations, including information about channel locations (in EEGLAB format), fixated words, and fixation onset times: a single MATLAB file `FRP.mat` in folder `EEG`.
- N400 sizes for all words and participants, including baseline level: a single tab-separated text table `N400.tsv` in folder `EEG`.

2.7 Analysis

A 5-gram language model was trained on the first slice of the NLCOW14 corpus, using the The SRI Language Modeling Toolkit (Stolcke, 2002) with modified Kneser-Ney smoothing (Chen & Goodman, 1999). This model then generated a surprisal estimate for each word in the stimuli sentences.

These surprisals served as the main predictor of interest in linear mixed-effects models fitted to either the log-transformed first-pass reading times or the N400 sizes; the latter on content words only because surprisal effects on the N400 are known to be driven almost exclusively by content words (Frank et al., 2015). Other predictors were the log-transformed word frequency in the SUBTLEX-NL corpus (Keuleers et al., 2010), the length of the word (number of characters), and the position of the word in the sentence, which was included as a covariate of no interest. For the N400 analysis, the N400 baseline is also a covariate of no interest. By-word-token and by-participant random intercepts were also included, as were by-participant random slopes of all fixed effects.

Data from trials marked for rejection (see Sect. 2.6.1) was excluded, as was data on sentence-initial words and words attached to punctuation. This left 56,876 data points for the reading time analysis and 16,277 data points for N400 analysis. The regression models were fit using the MixedModels package (v4.8.0; Bates et al., 2022) in Julia (v1.8.3; Bezanson et al., 2017).

3 Results

Due to a programming error, the responses to comprehension questions were not recorded for the first nine included participants. The mean error rate across the other 28 participants was 7.0% (range: 2.0% to 23.3%).

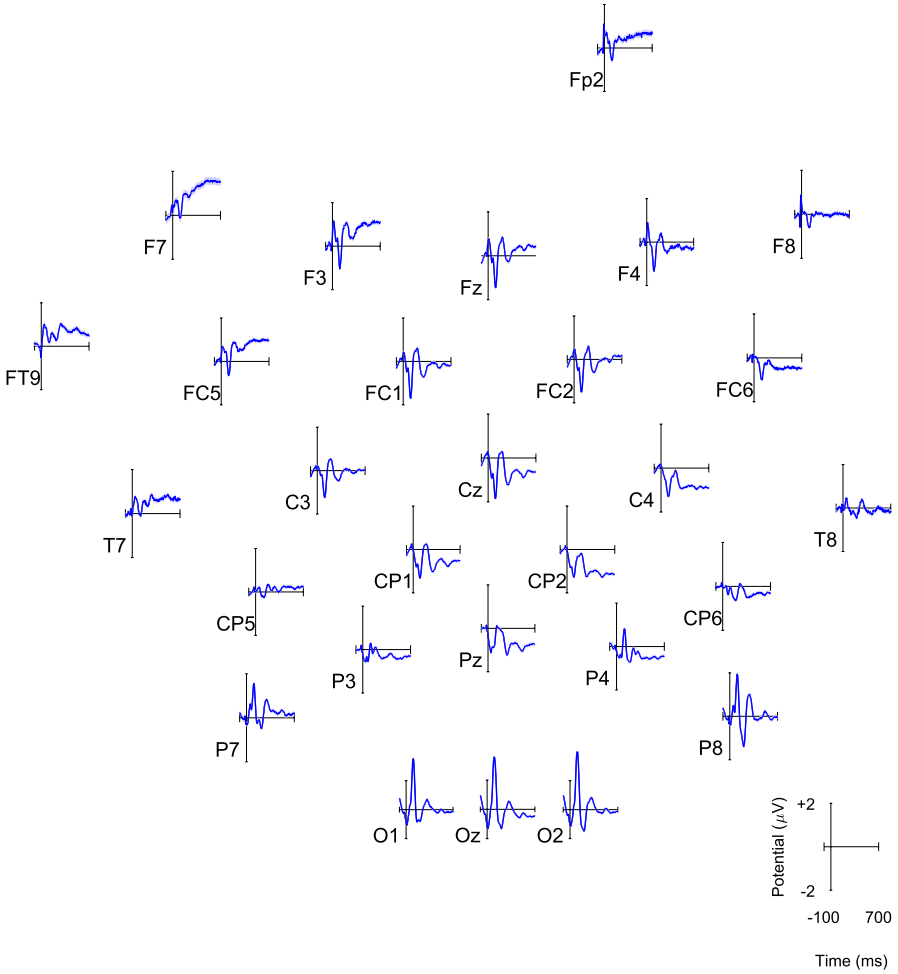

Fig. 1 Topographically plotted fixation-related potentials (average voltage time-locked to first fixation on each word) after ocular artifact correction. Shaded areas are 95% confidence intervals

3.1 Eye tracking

Table 2 shows the estimated regression coefficients of the most important predictors for the analysis of log-transformed first-pass reading times. For the full regression output and results on non-transformed reading times, see the file `analysis_RT.jl` in shared data folder `eyetracking`. As expected, there is a highly significant positive effect of word surprisal: less predictable words take longer to read. All Variance Inflation Factors (VIFs) were below 3.5, indicating that there were no multicollinearity issues.

Table 3 Fixed effects of interest from regression model fitted to N400 size after first fixations on content words

Predictor	*b*	SE	*z*	*p*
Surprisal	− 0.0384	0.0122	− 3.15	.002
Log wordfreq	0.0359	0.0523	0.69	.493
Word length	− 0.0153	0.0199	− 0.77	.440

3.2 EEG

3.2.1 Fixation-related potentials

Figure 1 shows the baseline-corrected FRPs, averaged over all non-rejected fixations from all included participants. Fixations on sentence-initial words, function words, and words attached to punctuation were excluded. No clear N400 is visible, but the shapes of the FRP curves are similar to those reported by Dimigen et al. (2011).

3.2.2 N400 effect

Table 3 shows the estimated regression coefficients of the most important predictors for the N400 analysis, excluding words if (part of) the corresponding eye-tracking signal in the N400 time window was rejected. For the full regression output, see the file `analysis_N400.jl` in shared data folder `EEG`. The significant, negative effect of surprisal shows that fixating on a less predictable word results in a stronger (more negative-going) N400 FRP component. All VIFs were below 1.9, indicating that there were no multicollinearity issues.

3.2.3 N400 localisation

The definition of N400 size was based on earlier literature rather than the current data because making the definition dependent on the data itself constitutes statistical 'double dipping', which results in invalid *p*-values. Having established that an N400-effect of surprisal is visible in the FRPs, we can further localize the effect in space and time by computing 'regression FRPs', similar to the 'regression ERP' approach by Smith and Kutas (2015). We fit a linear mixed-effects regression model at each sample point (from 0 ms to 600 ms from fixation onset) and at each electrode included in the N400 plus six neighbouring electrodes (P7, P8, CP5, CP6, FC1, and FC2).[5] This regression model has the same fixed and random factors as in the N400 analysis above, but the dependent variable is the ocular artifact-corrected EEG voltage rather than the N400 average. We then plot the time-series of surprisal coefficients as if they are FRPs. The results in Fig. 2 clearly show that the N400 effect is indeed limited to approximately the set of electrodes included in the N400-size

[5] See the supplementary materials for results on all electrodes.

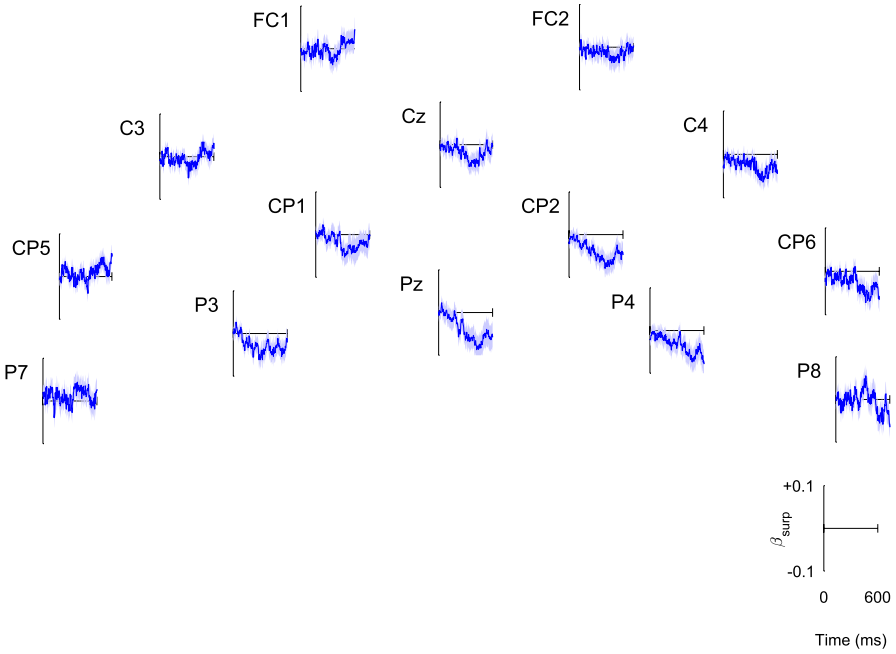

Fig. 2 Topographically plotted regression coefficients of surprisal, time-locked to first fixation on each word. Shaded areas are standard errors

definition, but the timing is similar to what is commonly found in RSVP studies with a peak at about 400 ms from fixation onset.

4 Discussion and conclusion

We have presented RaCCooNS: a new corpus of simultaneously recorded eye-tracking and EEG data during naturalistic sentence reading. Co-registration of eye movements and EEG allows for an analysis of EEG patterns during natural reading behaviour (as opposed to the traditional RSVP method) as well as a rigorous comparison between effects on behaviour (e.g., reading times, skipping rates, and regression probabilities) and neurophysiology (e.g., size of N400 and possibly earlier FRP components).

Many eye-tracking corpora already exist, often much larger than RaCCooNS, and several corpora with neural activation during language comprehension are also available. However, the number of data sets that combine synchronous eye-movement and EEG data is still very small. To the best of our knowledge, RaCCooNS is the only such corpus with narrative sentences, as well as the only one in Dutch.

Using MATLAB with the free EEGLAB and EYE-EEG toolboxes, our EEG preprocessing script (`EEG/preprocess.m`) can fully recreate the FRPs from the merged raw EEG and raw eye-tracking data. In contrast, preprocessing of the eye-tracking data cannot be rerun automatically as this involved hand-correction

of vertical drift. A similar limitation is that to-be-rejected participants, trials, and EEG electrodes were subjectively identified rather than by an algorithm that can be inspected and adjusted. However, the shared file `bad_data.txt` that lists all data rejections can be edited to investigate how these rejections affect the results.

RaCCooNS's eye-movement and EEG measures are specifically intended for the evaluation of predictions by computational linguistic models. Word surprisal is the most popular (and, arguably, successful) linking hypotheses between probabilistic language models and psycholinguistic measures of human sentence processing such as reading time and N400 size. Hence, the corpus's usability for language model evaluation is validated by the robust effects of surprisal on both first-pass reading times and the size of the N400 FRP component.

One possible weakness of RaCCooNS is that it is limited to individual sentences so it is not suitable for evaluating models that are sensitive to (discourse) relations across sentence boundaries. Indeed, we used surprisal values that were estimated by a sentence-bounded language model: a rather simplistic 5-gram model trained on individual sentences, which is both cognitively and linguistically very unrealistic. It stands to reason that more accurate surprisal values, which more accurately predict reading time and N400 size, can be obtained from more realistic models, such as (recurrent) neural networks and probabilistic grammars (Armeni et al., 2017). In principle, the RaCCooNS data can then be used to qualitatively compare the cognitive validity of surprisal estimates from different model architectures or model variants. Other model-based measures can also be investigated, for example, Frank and Willems (2017) used a distributional semantics model to quantify the semantic relatedness between content words in the UCL corpus sentences and showed that weaker relatedness results in larger N400 size during sentence reading.

Another potentially fruitful avenue for further research would be to go beyond FRPs and analyse effects on neural oscillations. Oscillatory power in different frequency bands can be predicted by surprisal (and related measures) from a trigram language model (Armeni et al., 2019), and Vignali et al. (2016) showed that the oscillatory dynamics from EEG reading studies are similar for RSVP and natural reading.

Two remaining questions are to what extent our ocular artifact removal procedure (based on Dimigen, 2020) was successful, and whether it was required to reveal a surprisal effect. Figure 3 presents the FRPs before and after artifact correction, immediately revealing that the correction procedure indeed resulted in substantial artifact reduction. Nevertheless, some ocular artifacts remain visible, in particular during saccades (i.e., in the baseline period) and shortly after fixation onset.

Interestingly, a post-hoc analysis of N400 sizes extracted from the FRPs *before* ocular artifact correction revealed an equally strong and reliable effect of surprisal ($b = -0.0388, z = -3.07, p = .002$). Nevertheless, artifact correction did meaningfully affect N400 sizes: Pre-correction, there was a highly significant effect of word position in the sentence ($b = -0.0425, z = -3.17, p = .0015$) which was no longer present post-correction ($b = -0.0104, z = -1.07, p = 0.28$). Apparently, the effect of word position on N400 size was in fact merely an ocular artifact that was successfully removed by the correction procedure. These conclusions are corroborated by the pre- and post-correction regression FRPs presented

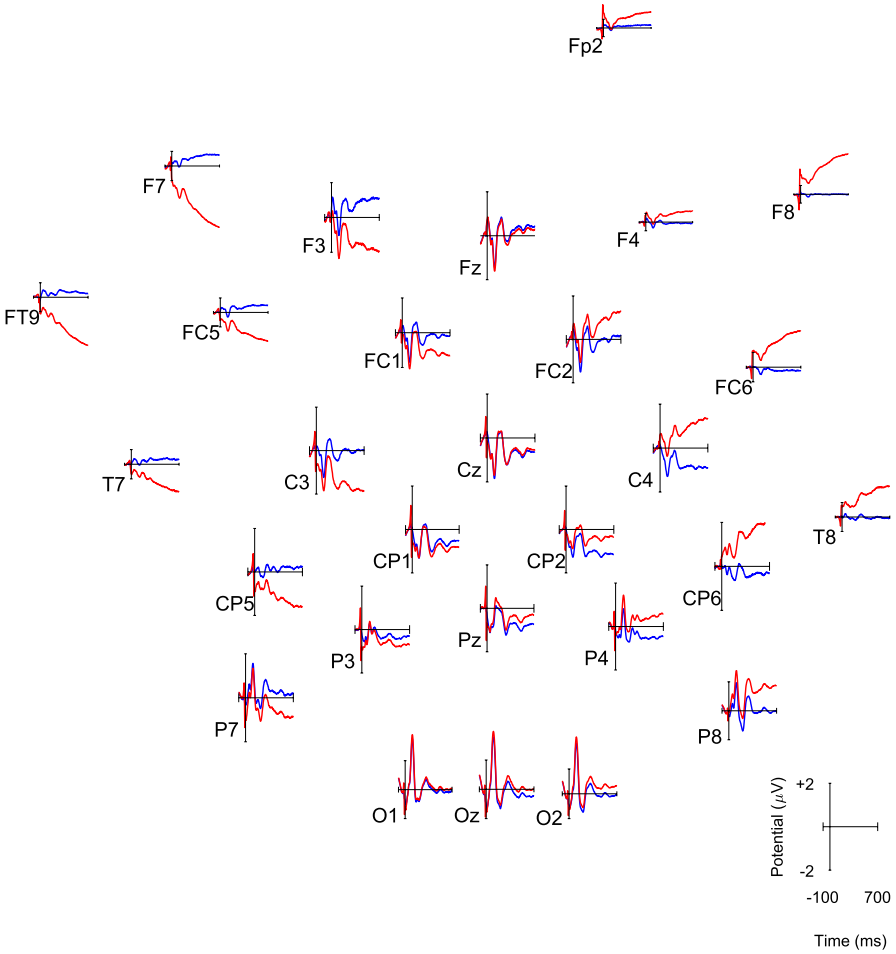

Fig. 3 Topographically plotted fixation-related potentials (average voltage time-locked to first fixation on each word) before (red) and after (blue) ocular artifact correction. Note the large y-axis scaling differences between channels. (Color figure online)

in the supplementary materials. Future work could include comparing the results of our artifact-correction pipeline to others (e.g., Henderson et al., 2013; Weiss et al., 2016). We also expect that better results can be obtained by applying deconvolution to correct for overlap between EEG responses to consecutive fixations (Ehinger & Dimigen, 2019; Shain & Schuler, 2021).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10579-023-09684-x.

Acknowledgements The work presented here was funded by NWO Gravitation Grant 024.001.006 awarded to the Language in Interaction Consortium.

Funding Funding was provided by Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Grant No. 024.001.006).

References

Armeni, K., Frank, S.L., Willems, R.M. (2017). Probabilistic language models in cognitive neuroscience: Promises and pitfalls. Neuroscience & Biobehavioral Reviews, 83, 579–588.

Armeni, K., Güçlü, U., van Gerven, M., Schoffelen, J.-M. (2022). A 10- hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. Scientific Data, 9, 278.

Armeni, K., Willems, R.M., van den Bosch, A., Schoffelen, J.-M. (2019). Frequency-specific brain dynamics related to prediction during language comprehension. NeuroImage, 198, 283–295.

Bates, D., Alday, P., Kleinschmidt, D., Calderón, J.B.S., Zhan, L., Noack, A., Arslan, A., Bouchet-Valat, M.,Kelman, T., Baldassari, A., Ehinger, B., Karrasch, D., Saba, E., Quinn, J., Hatherly, M., Piibeleht, M., Mogensen, P.K., Babayan, S., Gagnon, Y.L. (2022). JuliaStats/MixedModels.jl: v4.6.0. https://doi.org/10.5281/zenodo.5825693

Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B. (2017). Julia: A fresh approach to numerical computing. SIAM Review, 59, 65–98.

Chen, S.F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. Computer Speech and Language, 13, 359–394.

Cop, U., Dirix, N., Drieghe, D., Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. Behavior Research Methods, 49 (2), 602–615.

Dambacher, M., Kliegl, R., Hofmann, M., Jacobs, A.M. (2006). Frequency and predictability effect on event-related potentials during reading. Brain Research, 1084, 89–103.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open-source toolbox for analysis of single-trial EEG dynamics. Journal of Neuroscience Methods, 134, 9–21.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition, 109, 193–210.

Dimigen, O. (2020). Optimizing the ICA-based removal of ocular EEG artifacts from free viewing experiments. NeuroImage, 140 (4), 552–572.

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

Ehinger, B., & Dimigen, O. (2019). Unfold: an integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. PeerJ, 7, e7838.

Frank, S.L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. Psychological Science, 22, 829–834.

Frank, S.L., Monsalve, I., Thompson, R.L., Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. Behavior Research Methods, 45, 1182–1190.

Frank, S.L., Otten, L.J., Galli, G., Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. Brain and Language, 140, 1–11.

Frank, S.L., & Willems, R.M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. Language, Cognition and Neuroscience, 32, 1192–1203.

Futrell, R., Gibson, E., Tily, H.J., Blank, I., Vishnevetsky, A., Piantadosi, S.T., Fedorenko, E. (2021). The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. Language Resources and Evaluation, 55 (1), 63–77.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018), pp. 10–18.

Hale, J.T. (2001). A probabilistic Early parser as a psycholinguistic model. Proceedings of the 2nd conference of the North American chapter of the association for computational linguistics (Vol. 2, pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Henderson, J.M., Luke, S.G., Schmidt, J., Richards, J.E. (2013). Coregistration of eye movements and event-related potentials in connectedtext paragraph reading. Frontiers in Systems Neuroscience, 7, 28.

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

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. Vision Research, 45, 153–168.

Keuleers, E., Brysbaert, M., New, B. (2010). Subtlex-nl: A new frequency measure for Dutch words based on film subtitles. Behavior Research Methods, 42, 643–650.

Kliegl, R., Nuthmann, A., Engbert, R. (2006). Tracking the mind during reading: the influence of past, present, and future words on fixation durations. Journal of Experimental Psychology: General, 135 (1), 12–35.

Kornrumpf, B., Niefind, F., Sommer, W., Dimigen, O. (2016). Neural correlates of word recognition: a systematic comparison of natural reading and rapid serial visual presentation. Journal of Cognitive Neuroscience, 28, 1374–1391.

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

Kutas, M., & Hillyard, S. (1984). Brain potentials during reading reflect word expectancy and semantic association. Nature, 307, 161–163.

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

Lopopolo, A., Frank, S., Van den Bosch, A., Nijhof, A., Willems, R. (2018). The Narrative Brain Dataset: An fMRI dataset for the study of natural language processing in the brain. B. Devereux, E. Shutova, & C.- R. Huang (Eds.), Proceedings of the LREC 2018 workshop linguistic and neuro-cognitive resources (LiNCR), pp. 8–11.

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

Michaelov, J.A., & Bergen, B.K. (2020). How well does surprisal explain N400 amplitude under different experimental conditions? Proceedings of the 24th conference on computational natural language learning (CoNLL 2020). Association for Computational Linguistics.

Monsalve, I.F., Frank, S.L., Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. Proceedings of the 13th conference of the European chapter of the association for computational linguistics (pp. 398–408). Avignon, France: Association for Computational Linguistics.

Nárai, A., Nemecz, Z., Vidnyánszky, Z., Weiss, B. (2022). Lateralization of orthographic processing in fixed-gaze and natural reading conditions. Cortex, 157, 99–116.

Oostdijk, N., Hoste, V., de Jong, F., Reynaert, M. W. C., De Clercq, O., Desmet, B., & van den Heuvel, H. (2014).SoNaR-500. Database, Centrale voor Taal- en Spraaktechnologie.

Plöchl, M., Ossandón, J.P., König, P. (2012). Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. Frontiers in Human Neuroscience, 6, 278.

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

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. Proceedings of challenges in the management of large corpora (CMLC-3), pp. 28–34.

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. Proceedings of the eighth international conference on language resources and evaluation (LREC'12), pp. 486–493.

Shain, C., & Schuler, W. (2021). Continuous-time deconvolutional regression for psycholinguistic modeling. Cognition, 215, 104735.

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H.-D., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D. and Da Fonseca, S.M., Kuperman, V. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). Behavior Research Methods. 54:2843–2863

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

Smith, N.J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. Cognition, 128, 302–319.

Stolcke, A. (2002). SRILM–an extensible language modeling toolkit. Proceedings of the international conference on spoken language processing (pp. 901–904). Denver, Colorado.

Van der Sloot, K., Hendrickx, I., Van Gompel, M., Van Den Bosch, A., Daelemans, W. (2018). Frog, a natural language processing suite for Dutch, reference guide. Radboud University, Nijmegen.

Vignali, L., Himmelstoss, N., Hawelka, S., Richlan, F., Hutzler, F. (2016). Oscillatory brain dynamics during sentence reading: a fixation-related spectral perturbation analysis. Frontiers in Human Neuroscience, 10, 191.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. PLoS ONE, 9, e112575.

Wehbe, L., Vaswani, A., Knight, K., Mitchell, T. (2014). Aligning contextbased statistical models of language with brain activity during reading. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 233–243.

Weiss, B., Knakker, B., Vidnyánszky, Z. (2016). Visual processing during natural reading. Scientific Reports, 6, 26902.