A Corpus of Joint EEG and Self-Paced Reading of Natural Dutch Texts

Sara Møller Østergaard, Bruno Nicenboim
Department of Computational Cognitive Science, Tilburg University
s.m.ostergaard@tilburguniversity.edu

Background: There is an increasing interest in natural stimuli and experimental setups in psycholinguistic research, e.g., studies of reading times on natural texts [1], electroencephalographic (EEG) studies using all words in the sentence rather than single target words in hand-crafted contexts [2], and coregistration of EEG and eye-tracking [3, 4]. We present an open-access corpus of joint EEG and self-paced reading (SPR) of natural, medium-length, Dutch texts. The corpus contributes to a small collection of corpora with simultaneous recording of reading times and EEG. While this is often achieved using eye-tracking, the novel combination of SPR and EEG offers methodological advantages, particularly in aligning neural signals with word-level processing. Using SPR instead of eye-tracking eliminates parafoveal effects and allows EEG signals to be time-locked to word onsets, making it compatible with classical event-related potential (ERP) analyses. Furthermore, the inclusion of longer, natural texts enables novel analyses of reading and sentence processing.

Method: The corpus contains joint SPR and EEG recordings from 71 participants (47 female, mean age = 20.31) reading eight Dutch texts of approx. 800 words each. The linguistic stimuli are naturally occurring texts of different genres that were chosen based on overall fluency and comprehensibility. Every participant read seven texts using a SPR paradigm with central presentation and a single text in rapid serial visual presentation (RSVP). Bayesian hierarchical models were fit to validate the corpus by replicating surprisal effects on reading times and the N400 ERP component using log-probability and frequency of the given word as the predictors. The N400 was quantified as the average amplitude in centroparietal channels between 300 ms and 500 ms after the word onset. Log-probabilities of the words were obtained using unidirectional language models provided with the entire context of the texts preceding the word. The probabilities were extracted from four distinct models (see Table 1), and the average log-probability was used as the independent variable in the models. Word frequencies were obtained from the SUBTLEX-NL corpus [5]. All words were included in the statistical analysis.

Results: Both log-probability and word frequency had a negative effect on reading times, meaning that more probable and more frequent words elicited a faster reading pace, replicating previous findings [1, 4]. As expected [2, 4, 6], a positive effect of log-probability on N400 amplitude was found, indicating that less probable words elicit a more negative amplitude in the N400 time window. Previous studies have found either a positive or no N400 frequency effect [4, 6]. In contrast, we observed a negative word frequency effect on the N400 amplitude, with more frequent words eliciting a more negative N400 amplitude.

Use of the Corpus: Our corpus enables a range of future analyses. The simultaneous recording of reading times and EEG offers possibilities for exploring mechanisms underlying both behavioral and neurological responses during reading. The corpus allows for validation and exploration of EEG signal and ERPs for texts with longer linguistic dependencies. Additionally, the nature of the corpus makes it a useful asset in the development of computational cognitive models of reading. The corpus is available at DataverseNL [7].[1]

[1] Available from February 2026 at https://doi.org/10.34894/0O5XQ7

Table 1: Overview of Hugging Face models used for the analysis.

Hugging Face Reference
GroNLP/gpt2-small-dutch [8]
GroNLP/gpt2-medium-dutch-embeddings [8]
yhavinga/gpt2-large-dutch
yhavinga/gpt-neo-125M-dutch

Figure 1: Average SPR ERPs in centroparietal channels for content words with a log-probability <25% quantile, >75%, and between 25%-75%.

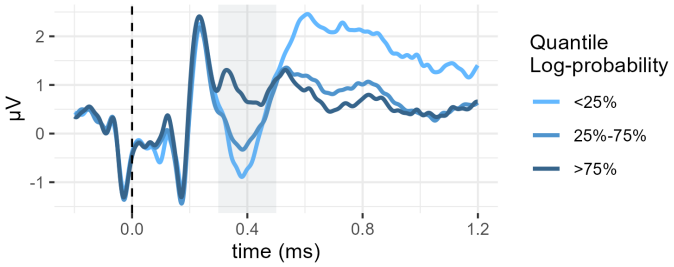

Table 2: Regression coefficients for models of reading times (only SPR), N400 amplitude (SPR and RSVP separately). All predictors were standardized. 95% credible intervals are reported in parentheses.

Coefficient	Reading Time	N400 (SPR)	N400 (RSVP)
Log-probability	$-0.02\ (-0.03, -0.01)$	$0.56\ (0.35, 0.77)$	$0.52\ (0.19, 0.85)$
Zipf Frequency	$-0.07\ (-0.08, -0.06)$	$-0.93\ (-1.22, -0.64)$	$-0.91\ (-1.20, -0.62)$

References

[1] 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. https://doi.org/10.1007/s10579-020-09503-7

[2] 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. https://doi.org/10.1016/j.bandl.2014.10.006

[3] Hollenstein, N., Troendle, M., Zhang, C., & Langer, N. (2020). ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 138–146. https://aclanthology.org/2020.lrec-1.18/

[4] Frank, S. L., & Aumeistere, A. (2024). An eye-tracking-with-EEG coregistration corpus of narrative sentences. *Language Resources and Evaluation*, *58*(2), 641–657. https://doi.org/10.1007/s10579-023-09684-x

[5] Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643

[6] Kretzschmar, F., Schlesewsky, M., & Staub, A. (2015). Dissociating word frequency and predictability effects in reading: Evidence from coregistration of eye movements and EEG. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1648–1662. https://doi.org/10.1037/xlm0000128 frequency effects.

[7] Østergaard, S., Lichtenberg, L., Boon, L., & Nicenboim, B. (2025). *EEG and Self-Paced Reading of Natural, Dutch Texts*. https://doi.org/10.34894/0O5XQ7

[8] de Vries, W., & Nissim, M. (2020). As good as new. how to successfully recycle english gpt-2 to make models for other languages.