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

Sara Møller Østergaard*, Lenneke Doris Lichtenberg†, Laura Boon‡, and Bruno Nicenboim*

*Department of Computational Cognitive Science, †Department of Communication and Cognition, ‡Department of Medical and Clinical Psychology, Tilburg University
s.m.ostergaard@tilburguniversity.edu

Abstract

We present the Tilburg corpus of Natural Dutch Texts (TiNT): A corpus of joint electroencephalography (EEG) and self-paced reading (SPR) of natural, medium-length, Dutch texts. The corpus contains recordings from 71 native Dutch speakers reading eight naturally occurring texts of around 600 words each. The texts are of varying genres and were chosen based on overall fluency and comprehensibility. To assess the quality of the corpus, we examined participant responses to comprehension questions, self-reported familiarity with the texts, and whether well-established effects replicated for both reading times and event-related potentials (ERPs) (N400 and P600). 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 use of SPR offers methodological advantages, particularly in aligning neural signals with word-level processing. In addition, the use of natural texts with longer dependencies makes the corpus a unique resource for psycholinguistic research. The corpus enables research into the relationship between neural and behavioral responses in naturalistic reading contexts.

Keywords: electroencephalography (EEG), self-paced reading (SPR), event-related potential (ERP), sentence processing, Dutch, surprisal effects

1. Introduction

Psycholinguistic studies have traditionally relied on hand-crafted sentences and, for electroencephalography (EEG) studies, fixed presentation rates. However, there is an increasing interest in both more naturalistic stimuli and experimental setups (Futrell et al., 2021; Frank and Aumeistere, 2024; Hollenstein et al., 2020; Cop et al., 2017; Siegelman et al., 2022; Jakobi et al., 2024; Ditman et al., 2007). Recent studies have replicated established effects for reading times and event-related potentials (ERPs) using natural texts where all words are of interest (Futrell et al., 2021; Frank and Aumeistere, 2024; Frank et al., 2015; Dimigen et al., 2011), highlighting the promise of more naturalistic approaches. We present the Tilburg corpus of Natural Dutch Texts (TiNT): A corpus of joint EEG and self-paced reading (SPR) of natural, medium-length, Dutch texts. TiNT offers a resource for studying real-time sentence processing in natural contexts through simultaneous behavioral and neural responses.

EEG studies of reading commonly use rapid serial visual presentation (RSVP) as the word presentation rate, in which words are presented one at a time at a predefined pace. While RSVP allows for a high degree of experimental control, the paradigm fails to reflect important aspects of natural reading as it removes active pacing (Kornrumpf et al., 2016). Studies have found that word presentation rate affects neural correlates of sentence processing (Bulkes et al., 2020; Tanner, 2019; Kornrumpf et al., 2016; Kuperberg et al., 2020), emphasizing the need for corpora collected using a more natural reading pace. Additionally, relying on RSVP makes it difficult to reconcile effects from reading times and EEG. Combining EEG with a SPR paradigm, where participants advance through a sentence one word at a time at their own pace, addresses these issues. It allows for natural individual variation in word presentation rates and enables the joint analysis of behavioral and neural measures during sentence processing. While deploying SPR provides advantages, it also introduces challenges. Advancing to the next word requires a motor response, which may introduce artifacts into the EEG signal. Moreover, as the presentation rate is not controlled, and people generally read faster than the fixed presentation rate in RSVP studies, ERPs might overlap with the onset of a new word. This is particularly relevant for later ERP components, such as the P600, but also the N400 may be affected, as typical self-paced reading times are around 300 ms (Futrell et al., 2021; Frank et al., 2013). Nonetheless, previous studies combining EEG and SPR have been able to detect common psycholinguistic effects (Ditman et al., 2007; Tanner, 2019; Payne and Federmeier, 2017).

Coregistration of eye-tracking and EEG provides an alternative paradigm for joint recordings of behavioral and neural data (Dimigen et al., 2011; Hollenstein et al., 2018, 2020; Frank and Aumeistere, 2024). Here, fixation-related potentials (FRPs) are

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11260–11271
11–16 May 2026.

used to time-lock the EEG signal to the onset of the first fixation. While eye-tracking arguably offers a more natural reading task compared to SPR, fixations are not clearly aligned with words, making it difficult to exclude the influence of the parafoveal preview of upcoming words. SPR provides an underexplored middle ground, allowing EEG signals to be time-locked to word onsets and eliminating parafoveal effects, making it compatible with classical ERP analyses from RSVP experiments.

The corpus presented in this paper not only aims to elicit a more natural reading task by relying on SPR rather than RSVP, but also by using natural texts extending beyond the sentence level. Natural reading comprehension often unfolds over longer contexts spanning multiple paragraphs. Using stimuli with extended discourse may reveal context-dependent processing effects that are unobservable with sentence-level stimuli. The importance of context for real-time sentence processing is evident in the effects of word predictability and frequency on both behavioral and neural responses during reading.

The predictability of a word, i.e., how unexpected a word is given its preceding context, has been shown to strongly affect reading comprehension (Hale, 2001; Levy, 2008). Predictability has been found to modulate ERPs, such as the N400 and P600, and reading times, where less predictable words amplifies the ERPs (i.e., a more negative N400 and more positive P600) and cause longer reading times (Ehrlich and Rayner, 1981; Kutas and Federmeier, 2011; Linzen and Jaeger, 2016; Pimentel et al., 2023; Delong et al., 2011; Kuperberg et al., 2020; Federmeier et al., 2007; Szewczyk and Schriefers, 2013; Dimigen et al., 2011; Aurnhammer et al., 2023).

Furthermore, word frequency also affects reading comprehension, where more frequent words have been found to elicit longer reading times and a more negative-going N400 amplitude (Sereno et al., 2020; Dambacher et al., 2006; Dambacher and Kliegl, 2007; Futrell et al., 2021; Shain, 2024). However, the position of the word within the sentence has been found to modulate frequency effects on the N400. The N400 frequency effect is primarily found when the word is presented in a list or at the beginning of sentences, whereas with context, the effect is either diminished or completely unobserved (Kretzschmar et al., 2015; Van Petten and Kutas, 1990; Wong et al., 2024).

TiNT complements and extends existing work relying on simultaneous recordings of EEG and behavioral data, such as the eye-tracking and EEG corpora, ZuCo (Hollenstein et al., 2018, 2020), RaCCooNS (Frank and Aumeistere, 2024), and Dimigen et al. (2011). Additionally, TiNT is inspired by recent corpora of behavioral signals deploying multi-sentence stimuli (Futrell et al., 2021; Cop et al., 2017; Siegelman et al., 2022; Hollenstein et al., 2022; Jakobi et al., 2024) and consists of medium-length, Dutch texts. The design of the corpus supports a range of future analyses, e.g., joint analysis of ERPs and reading times, exploration of the influence of discourse extended beyond sentence level on neural and behavioral signals, and could be used for the development and expansion of computational models of reading. The corpus is available at DataverseNL.

2. Methods

2.1. Linguistic Stimuli

The corpus consists of eight Dutch texts of around 600 words each and contains a total of 4,786 words across all eight documents. The texts were obtained from online open-access resources and are all naturally occurring texts. Documents were selected mainly based on overall fluency and comprehensibility as assessed by the two native Dutch-speaking co-authors (L.L. and L.B.). Additionally, documents with higher rates of specific syntactic features (i.e., idioms, ambiguity, metaphors, and rare words) were prioritized. The corpus consists of different types of texts to encompass multiple topics and reading difficulties, intending to elicit a varied sample of reading experiences. The texts were edited to ensure approximately an equal number of words in each text and to improve readability by correcting old spelling and grammatical errors. The edits were kept minimal, maintaining the structure of the original documents. A single document was translated from English to Dutch by the two native Dutch-speaking co-authors (L.L. and L.B.). Table 1 provides an overview of titles and sources of the texts. A summary of the documents by length is given in Table 2. Figure 1 shows histograms of word length and sentence length across the corpus.

2.2. Participants

71 healthy, adult participants (47 female, mean age = 20.31) partook in the SPR-EEG study. All participants were native Dutch speakers, had normal or corrected-to-normal vision, had never had any traumatic head injury, had not been diagnosed with dyslexia or any learning disability, and were right-handed. The participants were either recruited through a university participant pool for students from Humanities and Digital Sciences and rewarded in course points (N = 66) or recruited through online social media channels and received a monetary reward (N = 5). Two participants were excluded from the EEG analysis due to technical issues. A third participant was discarded from both reading time and EEG analyses

ID	Title	Author	Type of text	Source	Retrieval
1	Mijn Heer Zak met Rijst	F. Hadland Davis	Fairy tale	Gutenberg[a]	06-10-2023
2	Waarom de reuzen in Limburg zijn uitgestorven	Josef Cohen	Fairy Tale	Gutenberg[b]	10-10-2023
3	Aspasia	Wikipedia	History	Wikipedia[c]	10-10-2023
4	De zilveren schaatsen	P. J. Andriessen & Mary Mapes Dodge	Novel	Gutenberg[d]	09-10-2023
5	Permafrost	Wikipedia	Technical text	Wikipedia[e]	08-10-2023
6	Nomadisch pastoralisme	Wikipedia	Technical text	Wikipedia[f]	09-10-2023
7	Vleermuizen	Wikipedia	Technical text	Wikipedia[g]	09-10-2023
8	Violetta[*]	Wikipedia	Synopsis show	Wikipedia[h]	21-11-2023

[a] https://www.gutenberg.org/cache/epub/16043/pg16043-images.html#xd0e1307
[b] https://www.gutenberg.org/cache/epub/3455/pg3455-images.html#d0e5089
[c] https://nl.wikipedia.org/wiki/Aspasia
[d] https://www.gutenberg.org/cache/epub/60777/pg60777-images.html
[e] https://nl.wikipedia.org/wiki/Permafrost
[f] https://nl.wikipedia.org/wiki/Nomadisch_pastoralisme
[g] https://nl.wikipedia.org/wiki/Vleermuizen
[h] https://en.wikipedia.org/wiki/Violetta_(TV_series)
[*] The original text is in English. The experimental text was translated into Dutch by two native Dutch speakers.

Table 1: Document ID, title, and sources

ID	# Words	# Sentences	# Paragraphs	Word Length	Sentence Length
1	600	33	10	4.71 (2.75)	18.2 (7.16)
2	594	51	18	4.82 (2.80)	11.6 (7.39)
3	600	31	7	5.52 (3.08)	19.4 (7.72)
4	598	40	19	4.59 (2.28)	15.0 (13.8)
5	597	34	7	5.47 (3.40)	17.6 (7.70)
6	597	28	6	6.27 (4.09)	21.3 (8.84)
7	600	39	9	3.52 (3.52)	15.4 (7.24)
8	600	32	5	4.95 (2.47)	18.8 (10.3)

Table 2: Summary of documents by length. Word length and sentence length are summarized by their mean and standard deviation.

because of problems during data collection. The study was approved by the Research Ethics and Data Management Committee of Tilburg School of Humanities and Digital Sciences (Identification code: REDC2023.30a), and all participants gave informed consent prior to starting the experiment.

2.3. Procedure

The participants were placed in front of the presentation screen in a dimly lit room. Before starting the SPR-EEG experiment, they filled out a questionnaire on the experimental computer on demographics and literary proficiency. The experimental presentation script was created using PsychoPy (version 2022.2.5; Peirce et al., 2019). The texts were presented word-for-word at the center of the monitor in a white font on a grey full-screen background.

The eight texts were presented in a random order. One of the texts (*"Mijn Heer Zak met Rijst"* for half the participants and *"Permafrost"* for the other half) was presented using RSVP and the remaining seven texts using SPR. This setup was chosen to enable validation of the experimental procedure across both paradigms. Before the beginning of every text, the title and progress (e.g. "text 5 out of 10") were presented. Preceding the beginning of a new paragraph, a centrally located fixation cross appeared in grey for 500 ms, after which it turned white, indicating the possibility of a keypress that would result in the first word of the paragraph being presented. The experiment included two different presentation paradigms, SPR and RSVP, which defined the word presentation rate. For SPR, the word was presented until a keypress was registered. For RSVP, the word presentation duration was fixed and defined as $c \cdot 190$ms, where c is the number of

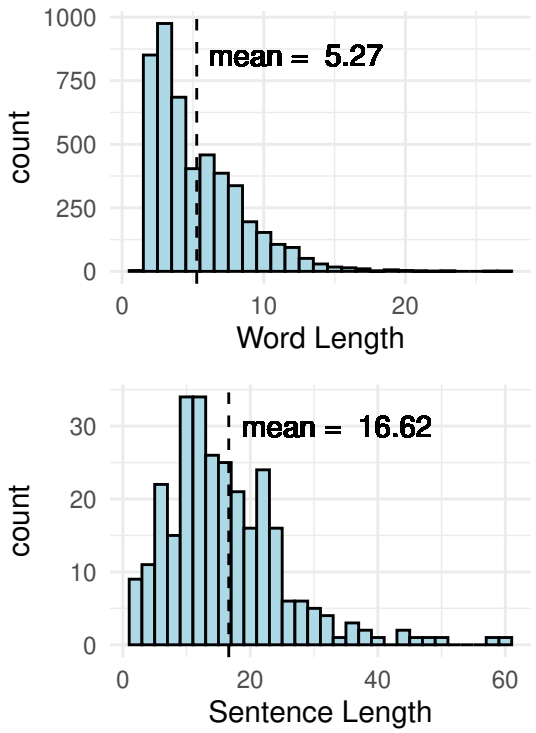


Figure 1: Histograms of word length and sentence length in the corpus

characters in the word, including punctuation, with a minimum duration of 250 ms, following prior literature (Frank et al., 2015). For both paradigms, each word presentation was followed by a 200 ms blank screen preceding the onset of the subsequent word. The blank screen after the final word was slightly longer, with a duration of 750 ms. The first time the participants were to read the text in a new presentation paradigm (i.e., SPR or RSVP), a training session preceded the beginning of the experimental text. The training sessions consisted of a short text of 74 to 98 words.

Every text was paired with seven questions: one familiarity question and six comprehension questions. These were presented right after reading the text that the questions belonged to. The wording of the familiarity questions depended on the type of text; for the fairy tales, the participants were asked if they were familiar with the specific story, for the novel and the synopsis, they were asked how familiar they were with the show or novel, and for the history and technical texts, the participant had to indicate their familiarity with the topic presented in texts. The answers to the familiarity questions were given on a 5-point scale (1 = I've never heard of it; 5 = I am very familiar with it). The texts were also paired with six comprehension questions on the content of the text of three types inspired by the Multipleye project (Jakobi et al., 2025). The question types reflect different levels of cognitive processing, following models from large-scale reading assessments (e.g., Artelt et al., 2001; Gehrer et al., 2013): local questions, bridging questions, and global questions. Local questions target the retrieval of explicitly stated information within a single sentence (i.e., textbase representation), bridging questions require integrating information across sentences or clauses to establish connections (e.g., causal or referential), and global questions involve reflecting on the overall meaning, purpose, or structure of the text (i.e., situation model) (Kintsch, 1998; Clark, 1975). All texts included three local questions, two bridging questions, and one global question (examples of the question types can be found in the appendices). The comprehension questions were multiple-choice questions with four possible answers. All questions can be found in the DataverseNL repository (see section 5). The texts in the training sessions were paired with two questions (one familiarity question and one comprehension question). The entire experiment took approximately one hour.

2.4. EEG recording and Preprocessing

The EEG signal was recorded at a rate of 512 Hz using a Biosemi system with a 64 scalp electrode cap mounted in accordance with the 10-20 montage. Additional electrodes were placed over each mastoid, as well as above and below the right eye and outer canthi of both eyes. Offline preprocessing of EEG signal was conducted using the R-package `eeguana` (Nicenboim, 2018). The scalp electrodes were re-referenced to an average of the two mastoids, and the eye electrodes were re-referenced to an above-to-below (horizontal electrooculogram; HEOG) and a left-to-right (vertical EOG; VEOG) bipolar montage, respectively. The continuous EEG signal was band-pass filtered between 0.1 and 30 Hz using zero-phase finite impulse response (FIR) filters. The width of the transition band at the low cut-off frequency was 0.01 Hz, and at the high cut-off frequency, 7.5 Hz. The raw EEG signal was visually inspected, and channels identified as faulty were excluded from further analysis. Additionally, if longer segments showed contamination, the signal of the entire text for the given recording was excluded.[1] Eye-blink artifacts were defined as peaks above 50μV in either Fp1 or Fp2 that occurred simultaneously with peaks below 100μV in the VEOG channel. A step function with a threshold of 50μV was used on the HEOG channel to identify horizontal eye movements. Additional artifacts in the EEG channels were marked

[1] A specification of excluded data can be found in `paper/data/exclude.xlsx` within the GitHub repository (see section 5).

Name	Time Window (ms)	Electrode sites (M10)	Electrode sites (10-20)
N400	300-500	1, 14, 24, 25, 26, 29, 30, 31, 41, 42, 44, 45	Cz, Pz, C4, CP6, P4, P3, CP5, C3, P8, PO4, PO3, P7
P600	500-700	1, 12, 14, 16, 24, 25, 26, 29, 30, 31, 39, 40, 41, 42, 44, 45, 46, 47	Cz, CP2, Pz, CP1, C4, CP6, P4, P3, CP5, C3, T8, TP8, P8, PO4, PO3, P7, TP7, T7

Table 3: Time windows and electrode sites related to N400 and P600 from Frank et al. (2015). Electrode sites were reported in an M10 montage; therefore, a manual translation to electrode sites in a 10-20 montage, used in the current analysis, is included.

using a sliding window of 200 ms where the min-max threshold exceeded 200μV for a minimum of three channels or where the min-max threshold exceeded 400μV in just a single channel. The signal was segmented into epochs from 200 ms before word onset to 1,200 ms after. Epochs that included a marked artifact were removed. If more than 30% of words within one text for one participant were excluded due to artifacts, then the entire story for the given participant would be excluded from the analysis. For 12 participants, the number of marked artifacts exceeded the threshold in all eight texts, meaning they were completely excluded from the analysis. The final EEG analysis included 145,194 epochs (18,369 with RSVP and 126,825 with SPR) from 56 participants.[2]

2.5. Event-Related Potentials

The ERP components N400 and P600 were extracted from the epochs. The N400 was defined as the average amplitude between 300-500 ms after the onset of the word in centroparietal electrodes. The P600 was similarly the average amplitude in selected electrodes in the time window 500-700 ms after onset of the word. The electrode sites and the time windows were adopted from Frank et al. (2015). Table 3 lists the exact electrode sites used to calculate the two ERPs. If any of the selected electrodes were marked as faulty, they would be excluded from the average; for three participants, this resulted in the exclusion of a single electrode site.

2.6. Analysis

The dependent variables, i.e., reading times and the N400 and P600 time- and spatial-window averages, were analyzed using Bayesian hierarchical models with predictability and word frequency as the independent variables. The models were fitted using the `brms` package (version 2.22.0; Bürkner, 2017) in R (R Core Team, 2024). All predictors were standardized with a zero mean and a standard deviation of one. The reading times were extracted from the time between the triggers of the current and the subsequent word, subtracting the time of the blank screen. Words with reading times lower than 100 ms or greater than 3000 ms were excluded from analysis.

For the reading time model, a log-normal likelihood was used, while for both the N400 and P600 models, a Gaussian likelihood was used (See Equations 1, 2, and 3). Group-level (also known as random) intercepts were estimated for each unique participant, document, and word. For each participant and document, group-level slopes for both predictability and frequency were estimated. Only a group-level slope for predictability was included for each word (as a single word would always have the same overall frequency regardless of context). In the models of ERPs from the RSVP reading condition, the group-level effects for document were excluded as only two unique documents were presented in the RSVP paradigm. The group-level effects were assumed to be uncorrelated. The group-level effects are denoted with u in Equations 1-3.

$$RT \sim LogNormal(\mu, \sigma) \tag{1}$$

$$ERP \sim Normal(\mu, \sigma) \tag{2}$$

$$\begin{aligned} \mu =& \alpha + u_{participant,0} + u_{document,0} + \\ & u_{word,0} + (\beta_1 + u_{participant,1} + \\ & u_{document,1} + u_{word,1}) \cdot lp + (\beta_2 + \\ & u_{participant,2} + u_{document,2}) \cdot freq \end{aligned} \tag{3}$$

Here, *RT* refers to reading time, *ERP* refers to the ERP components N400 and P600, *lp* refers to log-probability of the word, and *freq* refers to word frequency. Word predictability was quantified by the log-probability of the word given the entire preceding context of the story. The log-probability was extracted from four different GPT models from Hugging Face.[3] using the R-package `pangoling`

[2]The proportion of excluded epochs was approximately the same across the two reading paradigms.

[3]GroNLP/gpt2-small-dutch, GroNLP/gpt2-medium-dutch-embeddings, yhavinga/gpt2-large-dutch, and yhavinga/gpt-neo-125M-dutch. Revision can be found in appendices.

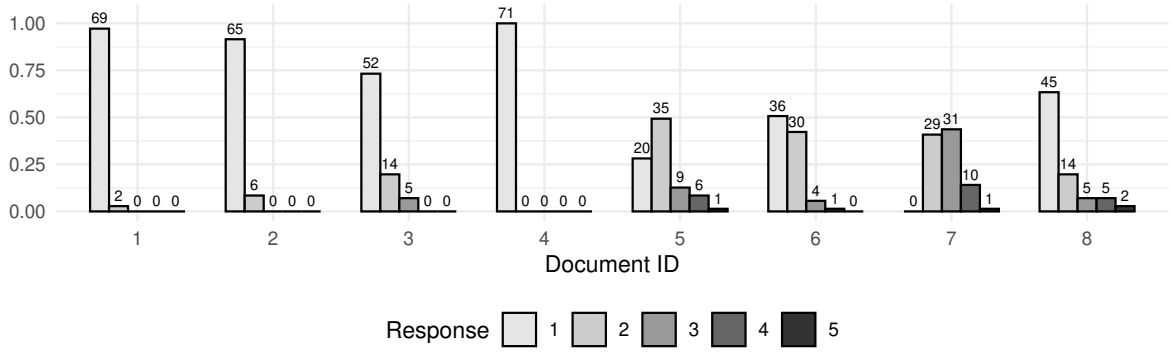


Figure 2: Familiarity ratings of the different texts on a 5-point scale (1 = I've never heard of it; 5 = I am very familiar with it). The y-axis indicates the proportion of participants giving that particular rating. The numbers above the bar are the counts.

(version 1.0.0; Nicenboim, 2023). Log-probabilities from the four models were highly correlated (Pearson's r > 0.8 for all comparisons; correlation matrix can be found in the appendices). The average log-probability score of all the models was used as the predictor variable in the statistical models. Word frequencies were obtained from the SUBTLEX-NL corpus (Keuleers et al., 2010). The Zipf frequency score was used, which is defined as $Zipf = log_{10}((frequency + 1)/44.106) + 3$. For words that were not represented in the SUBTLEX-NL corpus, a frequency score of $Zipf = log_{10}(1/44.106) + 3 = 1.3555$ was used.[4]

Different priors were used for the reading times and ERP models, as the scales of the dependent variables were different, i.e., reading times in ms and EEG components in μV. In all cases, regularizing priors were used to ensure stable and plausible estimates (Nicenboim et al., 2025). Additionally, a prior sensitivity analysis was carried out using the `priorsense` package (version 1.1.0; Kallioinen et al., 2023), confirming that the priors were not overly informative. The priors for the reading times model were as follows:

$$\begin{aligned}\alpha &\sim Normal(5.5, 1)\\ \beta &\sim Normal(0, .1)\\ u &\sim Normal(0, sd)\\ sd &\sim Normal_+(0, .5)\\ \sigma &\sim Normal_+(0, .5)\end{aligned}$$

The priors for the models of the ERP components were:

$$\begin{aligned}\alpha &\sim Normal(0, 20)\\ \beta &\sim Normal(0, 10)\\ u &\sim Normal(0, sd)\\ sd &\sim Normal_+(0, 10)\\ \sigma &\sim Normal_+(0, 10)\end{aligned}$$

A subsequent analysis was performed, taking word class into account. Part of speech (POS) tags for all words were identified using the *nl_core_news_sm* model from `SpaCy` (Honnibal et al., 2020). A boolean term indicating if the word was a content word (i.e., noun, verb, adjective, or adverb) was added as an interaction with all the predictors.

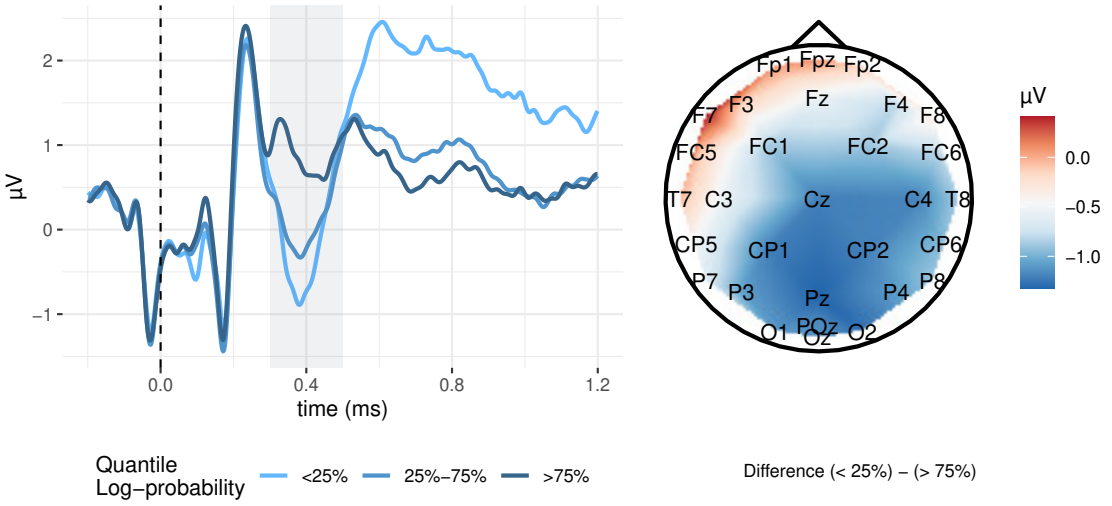
The models were fitted using four chains with 2,000 iterations, where half the iterations were warm-up samples. The models reported in this paper had no divergent transitions, $\hat{R}s \leq 1.04$, and the number of bulk and tail effective samples was at least 180, with an average of 1934.65.

3. Results

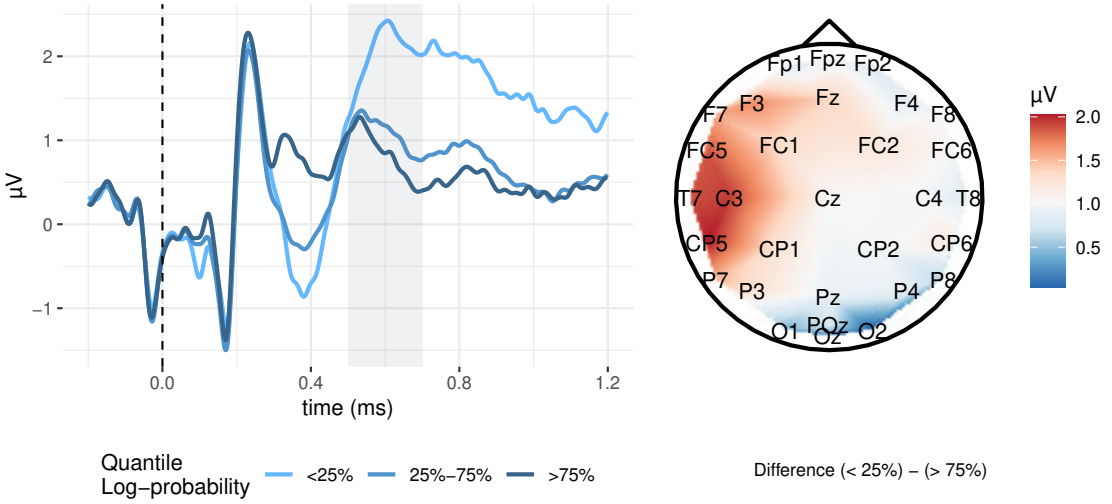
Participants generally reported low familiarity with the texts, with a mean rating of 1.56 (SD = 0.87) out of 5 across documents. The highest familiarity ratings were found for documents 5, 6, 7, and 8. For these documents, participants were asked about their familiarity with the topic or show referenced in the text, rather than the text itself. An overview of the familiarity rating of the different documents can be found in Figure 2

On average, participants answered 84.1% of comprehension questions correctly, with individual accuracy ranging from 45.8% to 97.9%. Across all individual questions, the proportion of correct responses ranged from 43.7% to 100%. There was no notable difference in accuracy across the

[4]Approximately 20% of the words in the TiNT corpus were not represented in the SUBTLEX-NL corpus, with similar proportions across all eight documents.



(a) N400. Average ERPs in centroparietal channels (see N400 in Table 3). Difference between ERP log-probability $< 25\%$ and $> 75\%$ in time window 300-500 ms.



(b) P600. Average ERPs in centroparietal channels (see P600 in Table 3). Difference between ERP log-probability $< 25\%$ and $> 75\%$ in time window 500-700 ms.

Figure 3: Average SPR ERPs for content words in centroparietal channels with a log-probability $< 25\%$ quantile, $> 75\%$, and between $25\% - 75\%$ and subtraction plots of the ERPs for content words. Difference between average amplitude for words with a log-probability $< 25\%$ quantile and words $> 75\%$ quantile in marked time windows.

question types. Local questions had a correct response 85.6% of the time, bridging 80.7%, and global 86.2%.

3.1. Regression Analysis

Figure 4 shows the estimated regression coefficients and 95% credible intervals (CrI) from regression models. The credible intervals represent the range within which the parameter lies with 95% probability.

Both log-probability and word frequency show clear negative effects on reading times ($\beta_1 = -0.02$, 95% CrI $= [-0.03, -0.01]$ and $\beta_2 = -0.07$, 95% CrI $= [-0.08, -0.06]$). This indicates that words with higher contextual probability and higher overall frequency were read more quickly, replicating previous findings in SPR and eye-tracking studies (Futrell et al., 2021; Frank and Aumeistere, 2024; Kennedy et al., 2013).

As expected (Frank et al., 2015; Ditman et al., 2007; Tanner, 2019; Frank and Aumeistere, 2024),

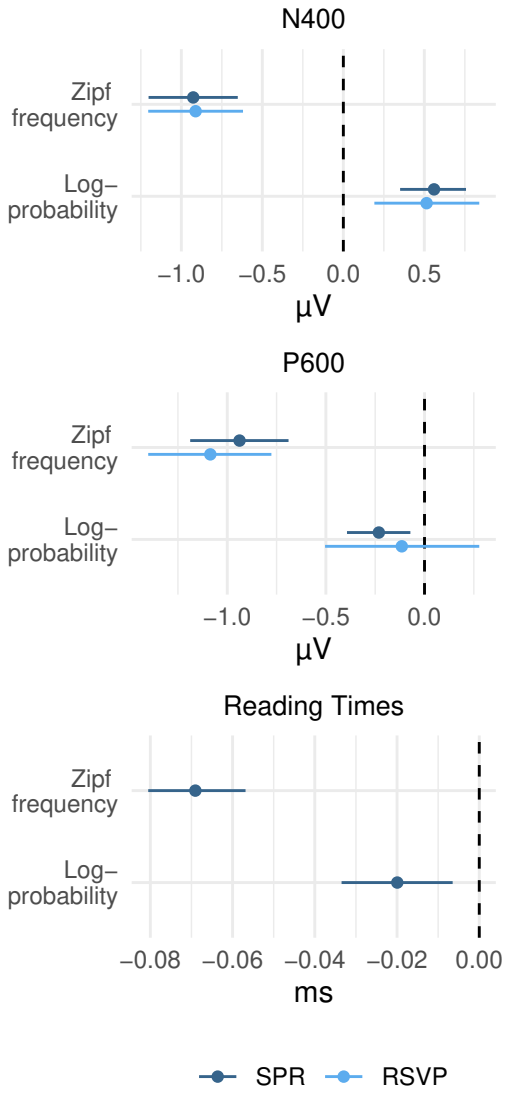


Figure 4: Regression coefficients and 95% credible intervals for models of reading times (only for SPR) and N400 and P600 amplitude (for SPR and RSVP separately). All predictors were standardized.

we find a positive effect of log-probability on the N400 amplitude ($\beta_1 = 0.56$, 95% CrI $= [0.35, 0.76]$). Thus, less probable words elicited a more negative-going N400. The model showed a negative effect of word frequency on N400 amplitude ($\beta_2 = -0.93$, 95% CrI $= [-1.20, -0.65]$), suggesting that more frequent words elicited stronger (more negative) N400 amplitude.[5] Results of previous studies looking at the N400 frequency effect are divided, with some finding positive effects (usually for words in lists or at the beginning of sentences) while others report no effect of word frequency (Kretzschmar et al., 2015; Wong et al., 2024; Shain, 2024). A subsequent model was run, which included an interaction effect between both predictors and word type (content vs. non-content). The effects of log-probability and word frequency were both amplified when only looking at content words. Figure 3a shows the ERPs for only content words in the SPR paradigm, and the difference in amplitude between high and low probability words in the N400 time window. The model using only data from the RSVP paradigm showed nearly identical results to those of the model using the SPR data.

For the model of the P600 in the SPR condition, negative effects were found for both log-probability and word frequency ($\beta_1 = -0.23$, 95% CrI $= [-0.39, -0.07]$ and $\beta_2 = -0.94$, 95% CrI $= [-1.19, -0.69]$). This suggests that less probable words and less frequent words elicited a more positive amplitude in the P600 time window. Figure 3b shows the ERPs for only content words in the SPR paradigm, as well as the difference between high and low probability words in the P600 time window. As for the N400 models, the estimated coefficients of the P600 models were similar between SPR and RSVP. However, the uncertainty of the estimated effect was greater in the RSVP condition.

4. Conclusion

We presented TiNT, a corpus of joined EEG and SPR of natural, medium-length, Dutch texts. The corpus contributes to a small number of open corpora of co-registered reading times and EEG, allowing for joint analyses of behavioral and neural signals during reading. The corpus was validated using regression analyses on reading times and ERPs. The analyses replicated well-established effects of predictability and word frequency on both the behavioral variable (reading times) and the ERPs (N400, P600). These findings align with results typically observed in RSVP-EEG and eye-tracking or SPR paradigms conducted separately. The results demonstrate the methodological validity of TiNT and highlight the corpus's suitability for future analyses.

5. Data and Code Availability

The Tilburg corpus of Natural Dutch Texts (TiNT) is available from DataverseNL at https://doi.org/10.34894/0O5XQ7 (Østergaard et al., 2025). The repository includes raw EEG and reading times as well as supplementary materials, such as preprocessed data to replicate the analysis, the experimental texts, the comprehension questions, and the questionnaires. Materials and reading

[5] The negative effect of word frequency was also found when excluding effects of log-probability from the model and when running the model only with words that had a word frequency in the SUBTLEX-NL corpus.

times data are openly shared under a CC-BY-NC-SA license. Raw EEG data and demographic information are shared upon request under the condition of academic use only and no redistribution of the raw data.

The experimental script and the code for analyzing the data are publicly available on GitHub at https://github.com/saraoe/natural-stories-dutch.

6. Ethics Statement

Ethics approval of the experimental study was given by the Research Ethics and Data Management Committee of Tilburg School of Humanities and Digital Sciences (Identification code: REDC2023.30a). Informed consent was obtained from all participants. The authors declare no competing interests.

7. Acknowledgements

This work was funded by NWO SSH Open Competition XS pilot - round 3 (project number: 406.XS.03.033). Hil Steketee contributed significantly to the data acquisition and data preprocessing.

8. Bibliographical References

Cordula Artelt, Petra Stanat, Wolfgang Schneider, and Ulrich Schiefele. 2001. Lesekompetenz: Testkonzeption und Ergebnisse. In Jürgen Baumert, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Petra Stanat, Klaus-Jürgen Tillmann, and Manfred Weiß, editors, *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*, pages 69–137. VS Verlag für Sozialwissenschaften, Wiesbaden.

Christoph Aurnhammer, Francesca Delogu, Harm Brouwer, and Matthew W. Crocker. 2023. The P600 as a continuous index of integration effort. *Psychophysiology*, 60(9):e14302.

Nyssa Z. Bulkes, Kiel Christianson, and Darren Tanner. 2020. Semantic constraint, reading control, and the granularity of form-based expectations during semantic processing: Evidence from ERPs. *Neuropsychologia*, 137:107294.

Paul-Christian Bürkner. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28.

Herbert H. Clark. 1975. Bridging. In *Theoretical Issues in Natural Language Processing*.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

Michael Dambacher and Reinhold Kliegl. 2007. Synchronizing timelines: Relations between fixation durations and N400 amplitudes during sentence reading. *Brain Research*, 1155:147–162.

Michael Dambacher, Reinhold Kliegl, Markus Hofmann, and Arthur M. Jacobs. 2006. Frequency and predictability effects on event-related potentials during reading. *Brain Research*, 1084(1):89–103.

Wietse de Vries and Malvina Nissim. 2020. As good as new. how to successfully recycle english gpt-2 to make models for other languages.

Katherine A. Delong, Thomas P. Urbach, David M. Groppe, and Marta Kutas. 2011. Overlapping dual ERP responses to low cloze probability sentence continuations. *Psychophysiology*, 48(9):1203–1207.

Olaf Dimigen, Werner Sommer, Annette Hohlfeld, Arthur M. Jacobs, and Reinhold Kliegl. 2011. Coregistration of eye movements and EEG in natural reading: Analyses and review. *Journal of Experimental Psychology. General*, 140(4):552–572.

Tali Ditman, Phillip J. Holcomb, and Gina R. Kuperberg. 2007. An investigation of concurrent ERP and self-paced reading methodologies. *Psychophysiology*, 44(6):927–935.

Susan F. Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

Kara D. Federmeier, Edward W. Wlotko, Esmeralda De Ochoa-Dewald, and Marta Kutas. 2007. Multiple effects of sentential constraint on word processing. *Brain research*, 1146:75–84.

Stefan L. Frank and Anna Aumeistere. 2024. An eye-tracking-with-EEG coregistration corpus of narrative sentences. *Language Resources and Evaluation*, 58(2):641–657.

Stefan L. Frank, Irene Fernandez Monsalve, Robin L. Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4):1182–1190.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to

the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2021. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55(1):63–77.

Karin Gehrer, Stefan Zimmermann, Cordula Artelt, and Sabine Weinert. 2013. NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for educational research online*, 5(2):50–79.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Nora Hollenstein, Maria Barrett, and Marina Björnsdóttir. 2022. The Copenhagen Corpus of Eye Tracking Recordings from Natural Reading of Danish Texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1712–1720, Marseille, France. European Language Resources Association.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5(1):180291.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 138–146, Marseille, France. European Language Resources Association.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Deborah N. Jakobi, Thomas Kern, David R. Reich, Patrick Haller, and Lena A. Jäger. 2024. PoTeC: A German Naturalistic Eye-tracking-while-reading Corpus.

Deborah Noemie Jakobi, Maja Stegenwallner-Schütz, Nora Hollenstein, Cui Ding, Ramune Kaspere, Ana Matić Škorić, Eva Pavlinusic Vilus, Stefan Frank, Marie-Luise Müller, Kristine M Jensen De López, Nik Kharlamov, Hanne B. Søndergaard Knudsen, Yevgeni Berzak, Ella Lion, Irina A. Sekerina, Cengiz Acarturk, Mohd Faizan Ansari, Katarzyna Harezlak, Pawel Kasprowski, Ana Bautista, Lisa Beinborn, Anna Bondar, Antonia Boznou, Leah Bradshaw, Jana Mara Hofmann, Thyra Krosness, Not Battesta Soliva, Anila Çepani, Kristina Cergol, Ana Došen, Marijan Palmovic, Adelina Çerpja, Dalí Chirino, Jan Chromý, Vera Demberg, Iza Škrjanec, Nazik Dinçtopal Deniz, Dr. Inmaculada Fajardo, Mariola Giménez-Salvador, Xavier Mínguez-López, Maroš Filip, Zigmunds Freibergs, Jéssica Gomes, Andreia Janeiro, Paula Luegi, João Veríssimo, Sasho Gramatikov, Jana Hasenäcker, Alba Haveriku, Nelda Kote, Muhammad M. Kamal, Hanna Kędzierska, Dorota Klimek-Jankowska, Sara Kosutar, Daniel G. Krakowczyk, Izabela Krejtz, Marta Łockiewicz, Kaidi Lõo, Jurgita Motiejūnienė, Jamal A. Nasir, Johanne Sofie Krog Nedergård, Ayşegül Özkan, Mikuláš Preininger, Loredana Pungă, David Robert Reich, Chiara Tschirner, Špela Rot, Andreas Säuberli, Jordi Solé-Casals, Ekaterina Strati, Igor Svoboda, Evis Trandafili, Spyridoula Varlokosta, Mila Vulchanova, and Lena A. Jäger. 2025. MultiplEYE: Creating a multilingual eye-tracking-while-reading corpus. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, pages 1–11, Tokyo Japan. ACM.

Noa Kallioinen, Topi Paananen, Paul-Christian Bürkner, and Aki Vehtari. 2023. Detecting and diagnosing prior and likelihood sensitivity with power-scaling. *Statistics and Computing*, 34(1):57.

Alan Kennedy, Joël Pynte, Wayne S. Murray, and Shirley-Anne Paul. 2013. Frequency and predictability effects in the Dundee Corpus: An eye movement analysis. *Quarterly Journal of Experimental Psychology*, 66(3):601–618.

Emmanuel Keuleers, Marc Brysbaert, and Boris New. 2010. SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, 42(3):643–650.

Walter Kintsch. 1998. *Comprehension: A Paradigm for Cognition*. Comprehension: A Paradigm for Cognition. Cambridge University Press, New York, NY, US.

Benthe Kornrumpf, Florian Niefind, Werner Sommer, and Olaf Dimigen. 2016. Neural Correlates of Word Recognition: A Systematic Comparison of Natural Reading and Rapid Serial Visual Presentation. *Journal of Cognitive Neuroscience*, 28(9):1374–1391.

Franziska Kretzschmar, Matthias Schlesewsky, and Adrian Staub. 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.

Gina R. Kuperberg, Trevor Brothers, and Edward W. Wlotko. 2020. A Tale of Two Positivities and the N400: Distinct Neural Signatures Are Evoked by Confirmed and Violated Predictions at Different Levels of Representation. *Journal of Cognitive Neuroscience*, 32(1):12–35.

Marta Kutas and Kara D. Federmeier. 2011. Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, 62(1):621–647.

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

Tal Linzen and T. Florian Jaeger. 2016. Uncertainty and Expectation in Sentence Processing: Evidence From Subcategorization Distributions. *Cognitive Science*, 40(6):1382–1411.

Bruno Nicenboim. 2018. *eeguana: A package for manipulating EEG data in R*.

Bruno Nicenboim. 2023. *pangoling: Access to language model predictions in R*.

Bruno Nicenboim, Daniel J. Schad, and Shravan Vasishth. 2025. The influence of priors: sensitivity analysis. In *Introduction to Bayesian Data Analysis for Cognitive Science*, 1st edn edition, chapter 3.4, page 634. Chapman & Hall.

Brennan R. Payne and Kara D. Federmeier. 2017. Pace Yourself: Intraindividual Variability in Context Use Revealed by Self-paced Event-related Brain Potentials. *Journal of Cognitive Neuroscience*, 29(5):837–854.

Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203.

Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger Levy, and Ryan Cotterell. 2023. On the Effect of Anticipation on Reading Times.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sara C. Sereno, Christopher J. Hand, Aisha Shahid, Ian G. Mackenzie, and Hartmut Leuthold. 2020. Early EEG correlates of word frequency and contextual predictability in reading. *Language, Cognition and Neuroscience*, 35(5):625–640.

Cory Shain. 2024. Word Frequency and Predictability Dissociate in Naturalistic Reading. *Open Mind*, 8:177–201.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Maria Da Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina A. Gattei, Areti Kalaitzi, Nayoung Kwon, Kaidi Lõo, Marco Marelli, Timothy C. Papadopoulos, Athanassios Protopapas, Satu Savo, Diego E. Shalom, Natalia Slioussar, Roni Stein, Longjiao Sui, Analí Taboh, Veronica Tønnesen, Kerem Alp Usal, and Victor Kuperman. 2022. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, 54(6):2843–2863.

Jakub M. Szewczyk and Herbert Schriefers. 2013. Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and Language*, 68(4):297–314.

Darren Tanner. 2019. Robust neurocognitive individual differences in grammatical agreement processing: A latent variable approach. *Cortex*, 111:210–237.

Cyma Van Petten and Marta Kutas. 1990. Interactions between sentence context and word frequencyinevent-related brainpotentials. *Memory & Cognition*, 18(4):380–393.

Roslyn Wong, Erik D. Reichle, and Aaron Veldre. 2024. Prediction in reading: A review of predictability effects, their theoretical implications, and beyond. *Psychonomic Bulletin & Review*.

Sara Møller Østergaard, Lenneke Lichtenberg, Laura Boon, and Bruno Nicenboim. 2025. EEG and Self-Paced Reading of Natural, Dutch Texts.

A. Appendices

A.1. Examples of Question Types

Type	Question	Translated Question
Local	Welke kleur had het haar van de drakenkoning?	What color was the dragon king's hair?
Bridging	Waarom veranderde de drakenkoning zijn uiterlijk in een vreselijk monster?	Why did the dragon king change his appearance into a terrible monster?
Global	Hoe zou je de relatie tussen Hidesato en de Drakenkoning omschrijven op basis van het verhaal?	How would you describe the relationship between Hidesato and the Dragon King based on the story?

Table 4: Examples of the three different types of comprehension questions for the story *Mijn Heer Zak met Rijst*. The questions were shown in Dutch in the experiment. Here, they are provided with an English translation.

A.2. Hugging Face References and Revision

Hugging Face Reference	Revision
GroNLP/gpt2-small-dutch (de Vries and Nissim, 2020)	d0e3f07a6e7cad045c45569bdaa08d318a275456
GroNLP/gpt2-medium-dutch-embeddings (de Vries and Nissim, 2020)	a7ea2d4a0dfc0a36b5fb11b93be9f63bf9cc89fb
yhavinga/gpt2-large-dutch	992e422249fbda8000b5e65fdb86a6fd7a690865
yhavinga/gpt-neo-125M-dutch	f7ba70ce7b62fbd1c29fd9012cf7b3b9bf0fd5d
spacy/nl_core_news_sm (Honnibal et al., 2020)	b9d28fe480eeacf9809fbd5ead5ef1ff27d9394e

Table 5: Overview of Hugging Face models used for the analysis.

A.3. Correlations of log-probability for GPT-models

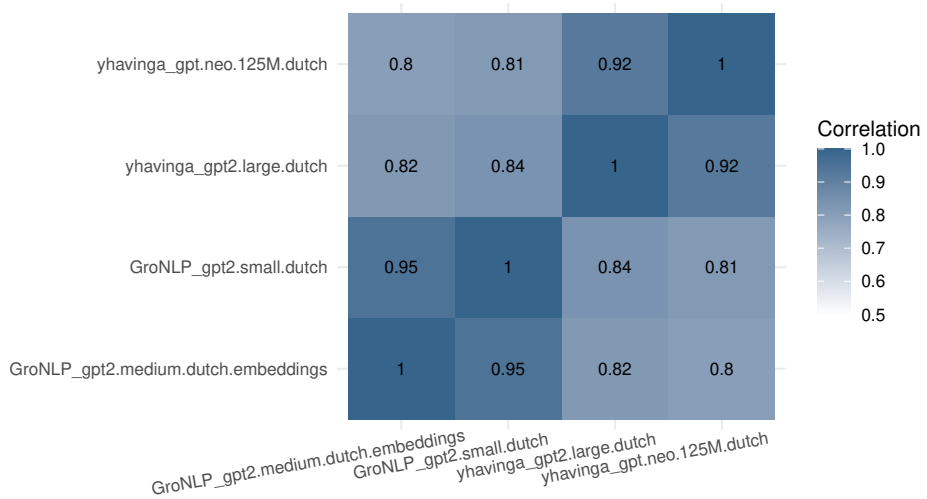


Figure 5: Pearson's correlation of log-probability extracted from the four GPT-models (GroNLP/gpt2-small-dutch, GroNLP/gpt2-medium-dutch-embeddings, yhavinga/gpt2-large-dutch, and yhavinga/gpt-neo-125M-dutch).