university of groningen / faculty of science and engineering / artificial intelligence

Word Surprisal effect on Cognitive Processes During Reading: An EEG and Eye-Tracking Study

Bachelor's Project Thesis

Luca Borgogno, s4776178, l.borgogno@student.rug.nl
Collaborator: Blanca San Millan Peralta, s4613902, b.san.millan.peralta@student.rug.nl
Supervisors: Jelmer Borst, j.p.borst@rug.nl
and Joshua Krause, j.krause@rug.nl

Abstract: This paper investigates the brain's prediction mechanisms in natural language processing by examining how word surprisal affects cognitive processes during naturalistic reading. Using EEG and eye-tracking technologies, this research measures brain activity and pupil dilation while participants read sentences from "Animal Farm" by George Orwell. This study quantified word surprisal using the GPT-2 language model, aiming to simulate the brain's predictive mechanisms. The EEG analysis focuses on the N400 and P600 components, associated with semantic and syntactic processing respectively. However, the results did not show significant differences in neural responses across varying levels of word surprisal. Pupil dilation data suggested a correlation between higher surprisal and increased cognitive load, although these results were not conclusive due to anomalies in the data indicating the need for further investigation. This study highlights the complexity of modelling the brain's predictive mechanisms and suggests improvements for future research to better understand how word predictability influences language comprehension during natural reading.

1 Introduction

Language processing is a fundamental cognitive ability that allows humans to seamlessly interpret linguistic information in real time. Language processing involves multiple brain regions and cognitive processes working together, enabling humans to understand and generate language efficiently (Friederici, 2011). Understanding the complex mechanisms behind natural language processing has been an area of focus of cognitive science. While previous studies have focused on single words presented in isolation or during listening tasks, this research aims to explore the entire reading process in a naturalistic setting similar to how people read at home. By examining how word surprisal affects cognitive processes during natural reading through the combined use of EEG and eye-tracking, we seek to provide a more comprehensive understanding of real world language comprehension.

Schrimpf et al. (2021) compared brain activity patterns with those of deep neural networks (transformers) trained for next-word prediction. These models simulate neural activation using layers of artificial neurons that respond to linguistic input. They discovered that the models showed activity patterns similar to those observed in the human brain during language processing tasks (Schrimpf et al., 2021). This similarity could be the result of the mechanisms observed in human brain activity that inspired the development of these AI predictive models. Alternatively, it could suggest that prediction is a key part of how the brain processes language in real time and mirrors the predictive mechanisms used in AI algorithms. This indicates that large language models can be used to generate predictions about the next word in a sentence based on the preceding context and can be used to simulate the brain's predictive process.

However, the role of predictive mechanisms in language comprehension is a topic of considerable debate among researchers. Nieuwland et al. (2018) examined the findings of DeLong et al. (2005) which initially reported evidence for probabilistic word

pre-activation, where the brain anticipates upcoming words during language comprehension. The study measured brain responses to sentences with varying levels of word predictability. Nieuwland et al. found minimal evidence for pre-activation and suggested the brain might not be as central to language comprehension as some researchers propose (Nieuwland et al., 2018). In contrast, Heilbron et al. (2022) demonstrated a hierarchical structure of linguistic predictions during natural language comprehension in a listening task. They suggested that the brain continuously generates and updates predictions at multiple levels of linguistic analysis. This study showed that higher level predictions (semantic and syntactic structures) influence lower level prediction (phonemes and words) supporting the idea of a top-down predictive mechanism in language processing. Adding to the debate, Falk Huettig (2015) concluded that prediction in language is facilitated by multiple mechanisms that work together to anticipate upcoming information. The paper emphasised that prediction is an important aspect of language processing but not fundamental (Huettig, 2015). The contrasting findings underscore the complexity of understanding predictive mechanisms in language processing and highlight the need for further research to analyze the conditions under which predictive processing occurs and its significance to language comprehension.

Given this ongoing debate, the application of natural language processing (NLP) provides a powerful tool for exploring these cognitive processes. NLP enables machines to understand, interpret and generate human language (Khurana et al., 2023). NLP models allow us to simulate and study how the brain processes language, providing valuable data that can be used to enhance the models themselves. This iterative process helps to refine NLP models, making them more accurate in replicating the brain's functions and improving our understanding of human cognition.

To analyse the brain predictive process, we chose to quantify contextual predictions using the large language model GPT-2, similar to the approach in Heilbron et al. (2022)'s experiment. GPT-2 is a transformer based neural network trained on large amounts of text data from the internet, learning to generate predictions about the next word in a sentence based on the preceding context through unsupervised learning. This is similar to how the brain uses context to inform predictions. However, GPT-2 is trained using data-driven methods without innate linguistic knowledge while the brain learns through experiences and sensory inputs. The brain uses more complex and dynamic predictive mechanisms but large language models can be used to simulate these processes in a less complex manner. The ability to generate predictions about which words are likely to come next in a given context provides a computational framework for studying prediction in the brain.

The probabilities calculated by the large language models can be used to analyse how unexpected a word was given the preceding context. If the brain is actively predicting the next words it suggests that it expects certain linguistic elements more than others based on the context. This expectation can be quantified using the concept of surprisal which measures the unexpectedness of a word given its context. Surprisal can be calculated by taking the inverse of the probability associated with a certain word so that lower probabilities result in higher surprisal. For example for the sentence: "He was twelve years old and had lately grown rather stout, but he...". The following word in this context is "was", with a probability of 0.44. The surprisal rating for "was" is approximately 2.27, which can be calculated by taking the inverse of this probability. Higher surprisal indicates that a word is less predictable and may affect the cognitive load and neural responses. This leads to the research question of this paper: How does the degree of word surprisal affect cognitive processes during reading?

To investigate how word surprisal affects cognitive processes while reading one can look at EEG and eye tracking data. EEG offers a high temporal resolution and captures activity in real-time which is crucial for studying dynamic cognitive processes. The ability of EEG to record event-related potentials (ERPs) makes it a good tool for examining how the brain responds to linguistic stimuli (Steinhauer et al., 2008). ERPs can be used to analyse the brain processes involved in reading after a word or sentence has been shown, revealing insights into the neural mechanisms underlying language comprehension. Vo & Gedeon (2011) used EEG to investigate the neural mechanisms behind reading comprehension . They found that EEG could effectively capture cognitive processes involved in reading and provide valuable insights into how the brain pro-

cesses written language Vo & Gedeon (2011). Using EEG one can analyse how the degree of surprisal affects cognitive load and brain processes during reading.

Research by Aurnhammer et al. (2021) explored how the brain processes language and showed that the N400 and P600 are particularly relevant in the context of language processing. In their study, participants read sentences presented one word at a time on a screen, with each sentence ending in a word that varied in predictability and semantic fit. The N400 component, which peaks 400 milliseconds after a stimulus, is a associated with semantic processing. This component is sensitive to the predictability of words within a given context, with higher surprisal words (less predictable) eliciting larger N400 responses indicating increased cognitive effort in integrating these words into the ongoing context. This finding aligns with previous research by Kutas and Federmeier (2011), who extensively reviewed the role of the N400 component in semantic processing and word predictability, and by Brown and Hagoort (1993), who demonstrated the N400's sensitivity to semantic priming effects. The P600 component, appearing around 600 milliseconds after a stimulus, is linked to syntactic reanalysis and integration processes. It is observed when readers encounter syntactic anomalies or complex sentence structures, reflecting the brain's effort to reprocess and integrate unexpected linguistic elements. This is consistent with findings by Brouwer and Crocker (2017), who discussed the relevance of the P600 in syntactic processing and reanalysis, and by Delogu et al. (2021), who examined the overlap and interaction between the N400 and P600 components in language comprehension.

We chose to focus on the central (Cz) and parietal (Pz) electrodes for a detailed analysis of our results based on findings from prior research. Šoškić et al. (2022) noted that Cz was the most commonly used electrode for measuring N400, appearing in 55.4% of studies. This prelevance underscores the importance of Cz in capturing semantic processing and word predictability effects. In the P600 time window, (Aurnhammer et al., 2021) stated that there was significant effect of word predictability. Additionally, the study by Yang et al. (2015) highlighted that the N400 effect was followed by the P600 effect, which was most prominently observed over central-parietal areas, including the Cz and Pz sites. This finding supports the relevance of these electrode locations for capturing ERP responses to linguistic processing.

Eye tracking technology complements EEG by keeping track of where the participants direct their gaze and provides context to the recorded brain signals. Eye tracking allows researchers to measure gaze direction, fixation duration and pupil size. Eye tracking data can provide detailed information about reading behaviour and by correlating specific eye movements and pupil responses with neural activity, researchers can obtain a comprehensive view of the cognitive processes involved in reading. Pupil dilation has been shown to be a reliable measure of cognitive load Krejtz et al. (2018) which could increase when reading more unexpected words. A 2019 study revealed that unconscious surprisal can trigger larger pupil responses (Alamia et al.). This physiological response reflects the brain's unconscious prediction mechanisms and the effort required to process unexpected linguistic input. Measuring pupil dilation for words with different surprisal ratings could reveal how cognitive load varies with predictability.

Combining GPT-2 with EEG and eye tracking allows for a multi-faceted investigation of how word surprisal impacts cognitive processes during reading. GPT-2 generates predictions about what the next word in a sentence based on the preceding context similar to how the brain uses context to inform predictions. By using GPT-2 we can quantify contextual predictions and calculate the surprisal ratings of different words. By examining ERP and pupil area associated with words of various surprisal, we could better understand the cognitive mechanisms underlying language processing. We hypothesize that higher surprisal ratings (less predictable words) will be associated with distinct ERP patterns, including more pronounced N400 and P600 responses, and increased pupil dilation, reflecting greater cognitive load.

2 Methods

In this experiment, participants engaged in a reading task involving the first five chapters of "Animal Farm", by George Orwell. Their brain activity was recorded using EEG and eye movements, gaze coordinates and pupil dilation were recorded with

an eye tracker.Participants are instructed to read sentences naturally, displayed one at a time on the screen, in a single row. To trigger the next sentence, they fixated on a fixation dot, always shown at the same coordinates on the right side of the screen. At the end of every chapter, the participants could take breaks. The time at which new sentences were displayed was recorded for further data processing and analysis.

2.1 Participants

Two participants performed the experiment. They were recruited through flyers and questionnaires in which information on potential candidates was collected. The participants were selected based on criteria including language proficiency, absence of visual impairments, and the number of times they read in a week. Both were native speakers of English to ensure that there were no language barriers or misunderstandings in the text that could alter the EEG data. In addition, they had no visual impairments to facilitate eye tracking and ensure the accuracy of the recording. Lastly, they read 3-4 times (n = 1) and 1-2 times a week (n = 2) for pleasure. This requirement was selected to ensure that the participant can read fluently and with better reading comprehension.

Therefore, the data consisted of 2 participants (both female, $N_{female} = 2$), aged 20 and 21 years old, both pursuing a university bachelor's degree. An informed consent form was signed in which the participants were informed to abort the experiment at any time. After the experiment's completion, the data was anonymized for later analysis. Both participants were compensated for their time.

2.2 Apparatus

2.2.1 Eye-tracking

To track where the participants were looking on the screen while reading the text, the EyeLink 1000 v.4.40 was used with PyGaze v.0.8a6. The former is a video-based eye-tracking system that supports high sampling rates (up to 2000Hz). It outputs the x and y coordinates and the pupil size, that we sampled at 1000hz to extract exact gaze position throughout the experiment. The left eye was recorded throughout the experiment to track eye movements and fixations.

2.2.2 Electroencephalography

To measure the electrical activity in the participant's brain while reading, we used EEG. We utilized an EEG cap equipped with 32 gel-based electrodes and 6 skin electrodes from BioSemi. The choice of the number of electrodes offers a compromise between data quality and manageability. The 32 electrodes cover the major areas of the scalp that are involved in the reading process while ensuring the ease of the setup and preprocessing of the data. The 6 skin electrodes were located above the left eye, underneath the left eye, next to the right and left eye's outer canthus (temple part of the face) and on the right and left mastoids. In addition, the EEG was configured to record at a sampling rate of 1024 Hz which provides a high temporal resolution and captures brain activity at the millisecond level.

2.3 Setup and Materials

Participants were asked to read sentences from the first 5 chapters of "Animal farm", displayed one at a time. The experiment was conducted in a quiet room to minimize distractions and ensure controlled environmental conditions. In addition, a chin rest was utilized to minimize the participant's head and neck movements, thus ensuring EEG data quality and the facilitation of eye tracking. The experiment was programmed in OpenSesame software version 4.0.23 with Python version 3.11.8 that integrated eye-tracking and EEG recording systems.

Animal farm was chosen for its engaging content from the start to immediately capture the participant's interest, avoiding lengthy and tedious descriptions. The aim was to minimise mind-wandering in order to reduce its disruptive effects on the EEG signal. In addition, it is in the public domain facilitating the replication of the study.

Figure 2.1 shows how the sentences were displayed on the screen. A fixation dot was placed on the right side of the screen, after the sentence and was encircled to indicate the area where participants should focus their gaze. This setup was programmed to turn the circle around the fixation dot green momentarily and trigger automatic progression to the next sentence once the eye tracker detected that the participant had fixated on the

dot for 500 milliseconds. The sentences all started from the same coordinates and the fixation dot was always in the same position, ensuring that participants knew where to look each time and maintained a consistent gaze pattern throughout the experiment. We also implemented a feature where if the participant did not advance to the next sentence within 15 seconds, a re-calibration of the eye tracker would be triggered. This time frame is more than sufficient to read a sentence and move on to the next one, so a delay likely indicates that the eye tracker did not capture the pupil accurately or the fixation conditions were not met.

The code split the text by prioritizing punctuation signs to facilitate natural reading and a maximum of 75 characters were shown at a time in order to display the sentences on a single line. To display the text, we used a monospaced font so that every character had the same width, which facilitated the decoding of which words the participants were looking at when analyzing the eye tracker data.

Once the participants fitted the EEG cap, placed the head on the chin rest and the eye tracker was calibrated for the fist time, they underwent a training phase to accustom the participants to the real experiment, including the multiple sentence splits of the story and reading interface.

The training phase involved a short story generated by ChatGPT, version 3.5. It was used to generate a randomized story and avoid any disturbance with the data collected during the experimental phase. The prompt given was "Create a short story of about 500 characters", to assure the multiple splits of sentences.

Upon completion of the training phase, the participants started the real experiment. The eye tracker and EEG data were recorded throughout the experiment. A calibration of the eye tracker was performed at the beginning of every new chapter to ensure the accuracy of the eye data throughout the experiment. If participants did not proceed to the next sentence within 15 seconds, the calibration was repeated. Besides, the participants were allowed a small break of 5 to 10 minutes at the end of each chapter to prevent fatigue and maintain engagement which can alter EEG signal.

Figure 2.1: Display of the experiment, including fixation point and an example sentence

2.4 Data Preprocessing

The study involved recording EEG and eye tracker data, which were subsequently preprocessed to derive the study's dependent variables.

The mastoid electrodes were used to re-reference the data to reduce noise and improve the signal quality. The rest of skin electrodes that extracted the horizontal and vertical eye movements served to create electrooculography (EOG) channels. This is an essential step to later identify and correct eye movements in the EEG data. The data presented a peak at 50Hz, probably caused by the electrical interference from the battery. Therefore, the data were band-pass filtered between 1Hz and 40Hz to remove low-frequency drifts and high-frequency noise. The artifacts were removed after performing a visual artefact inspection during which excessively noisy segments and periods of instability were removed. The last step of pre-processing before epoching encompassed independent component analysis (ICA). Components related to EOG were detected and removed from the EEG data.

The data collected with the eye tracker was also preprocessed. Firstly, the trigger values were converted into messages that inform on the beginning and end of a new sentence. This enables the possibility to convert data from ASC data into a dataframe that was filtered based on the messages. Subsequently, the data can be further preprocessed using the PupilPre package (v0.6.2.).

The trials are filtered, aligned to start messages and a time column is created. Thereafter, the left eye data is selected for which the blinks are removed

along with the saccades and the artifacts around them. This enabled the further analysis of the data collected.

2.5 Analysis

We examined the effect of surprisal on pupil and EEG data for the first word of each displayed sentence. Initially, the preprocessed data were divided into epochs.

For the EEG data these consisted of 1.5 seconds from stimulus onset, capturing the neural responses to the first 1.5 seconds of the text display that corresponded to the processing of the first word of the sentence. The pupil data was segmented into epochs of a duration of 3.5 seconds from stimulus onset, as the pupil data's response to surprising stimuli is still immediate than the neural response's captured with EEG. In trials in which the sentence was read in less than 3.5 seconds, the duration of the epochs was calculated as the difference between the start and end time of those trials.

Subsequently, predictions were made using the open-source transformer GPT-2 with a fixed context window of 1024 tokens. The text was tokenized, and the probabilities of the first word displayed for each sentence were generated. We decided to focus solely on the first word of each sentence. We assumed that when the screen transitioned from one sentence to the next, the participants' brains had more time to predict the upcoming word. Additionally, participants did not fixate on every word but always began reading from the first word. These probabilities were used to calculate the surprisal ratings of the words as the inverse of the predicted probabilities which are used as the different conditions to be later analysed.The probabilities were very small and had many similar values, meaning they were not normally distributed. To normalize the data, we applied the logarithm to the surprisal ratings to obtain more normally distributed values that we could divide into bins. The surprisal ratings then allowed us to categorize the epochs of data, for both participants combined, into different levels of surprisal bins. We divided the surprisal ratings into five bins, ensuring that each bin had roughly the same amount of data.

Event-Related Potentials (ERPs) were generated to analyze EEG data. ERPs were produced by condition, to allow for comparison. These plots allow for the recognition of ERP components that, based on the literature, can explain cognitive processes. Thereafter, the pupil data was averaged across epochs per condition. To account for pupil dilation differences between participants, we calculated the average change of pupil size with a baseline of 5 minutes per participant. By plotting the graphs, we expect to see a higher dilation change with higher levels of suprisal ratings.

3 Results

We focused on examining how word surprisal, as quantified using the GPT-2 model, impacts cognitive processes during reading. Specifically, we analysed ERPs and pupil dilation to understand the relationship between word predictability and cognitive load. The data of both participants was combined and we decided not to include data from chapter 1 as there were anomalies in the values recorded. The results may have been anomalous due to the less informed predictions by GPT-2, as there was little or no preceding context at the start of the book.

After calculating the surprisal ratings, we found that the mean surprisal rating was 10.73 with a standard deviation of 4.21. To analyse the impact of word predictability on cognitive processes, the surprisal ratings were categorised into specific bins based on their value ranges. The bins were set as follows: These bins were set as follows: Bin 0 ranged from 0.07745 to 8.07063, Bin 1 from 8.07063 to 10.97010, Bin 2 from 10.97010 to 12.63626, Bin 3 from 12.63626 to 14.60179, and Bin 4 from 14.60179 to 23.06128.

3.1 EEG analysis

The EEG data were analysed to identify ERP components for each bin of surprisal ratings. The ERPs were compared across different electrode sites (F7, Fz, F8, T7, Cz, P7, Pz, P8) to capture neural responses, as seen in Figure 3.1. The ERP plots display neural responses to words categorized into five surprisal bins, with bin 0 containing the most predictable words and bin 4 the most surprising words. The x-axis of each plot represents the time in milliseconds (ms) where the 0 ms mark indicates when the stimulus (sentence) appeared on the par-

ticipant's screen. The y-axis represents the amplitude of the ERP in microvolts (µV), indicating the strength of the neural response. In the ERP plots we focused on the N400 and P600 components as they reflect different aspects of cognitive processing during reading.

We focused on the central (Cz) and parietal (Pz) electrodes as these are the most commonly associated with the N400 and P600, respectively (Aurnhammer et al., 2021; Šoškić et al., 2022; Yang et al., 2015). Figure 3.2 provides a detailed comparing on evoked responses at the Cz and Pz electrodes. In the ERP data from the Cz electrode, we observe peaks strongest in magnitude at 200 ms (N200), 400 ms (P400) and 600 ms (N600). Similarly, at the Pz electrode, peaks strongest in magnitude can be seen at 100 ms (N100), 200 ms (N200), 400 ms (P400) and 600 ms (N600). In both plots the amplitude differences between the bins are minimal, suggesting that different levels of surprisal did not result in substantial differences in the brain's electrical responses. The data does not show the N400 and P600 effects as described in the literature. Both ERP plots show a negative component around 500-600 ms, which might be an N400 component occurring later than usual due to the experimental setup. We will provide a more detailed explanation of this phenomenon in the discussion section.

3.2 Eye-tracking analysis

The area of the left pupil was measured throughout the experiment as pupil dilation is an indicator of cognitive load (Krejtz et al. (2018) and unconscious surprisal Alamia et al. (2019)). We normalised the data by subtracting the average pupil size during the last five minutes from each sample. This baseline adjustment was done to account for individual differences between subjects and to normalize fluctuations in pupil size across trials. By using the average over the previous five minutes as a reference, we can more accurately assess how a word's surprisal impacts pupil dilation. This approach focuses on the effects at the sentence start, rather than continuous changes throughout the experiment. Figure 3.3 represents the average left pupil area over time for each surprisal bin. The x-axis represents time in milliseconds (ms) with the 0 ms mark indicating when the stimulus (sentence) appeared on the participant's screen. The y-axis represents the average left pupil area.

Figure 3.3 shows a pronounced peak at around 250ms for all bins which could be related to the initial pupil dilation caused by the switching of the sentence screen. Interestingly, if we exclude bin 0, bins 2,3 and 4 consistently show higher pupil dilation than bin 1, aligning with our hypothesis that higher surprisal correlates with greater cognitive load. This observation could suggest that the initial peak in dilation. seen in Figure 3.3. reflects the more than just the automatic response to sentence switching. It could be showing the increased cognitive load due to brain's prediction mechanism actively anticipating upcoming words. This observation is not conclusive because bin 0 (lowest surprisal), which should theoretically have the least pupil dilation, shows higher values than the other bins. This anomaly suggests that other factors could be influencing pupil dilation, or that the surprisal ratings for bin 0 were unusual.

4 Discussion

This study aimed to answer the research question: How does the degree of word surprisal affect cognitive processes during reading? To explore this, we quantified word surprisal using the GPT-2 model and collected EEG and eye tracking data to analyse the neural and physiological responses to varying levels of word predictability.

4.1 EEG analysis

The EEG analysis focused on the N400 and P600 components, which are linked to semantic and syntactic processing, respectively (Aurnhammer et al., 2021).Based on previous findings (Heilbron et al. (2022); Aurnhammer et al. (2021)), we hypothesised that higher surprisal ratings would result in more pronounced N400 and P600 effects.

The N400 component, typically peaking around 400 ms after stimulus onset, is sensitive to semantic discrepancies and predictability. We hypothesised that words with higher surprisal rating would trigger larger N400 amplitudes, reflecting the increased cognitive effort required to integrate these unexpected words into the ongoing context. Contrary to our hypothesis, the results shown in Figure 3.2 do not display a negative peak at 400 ms;

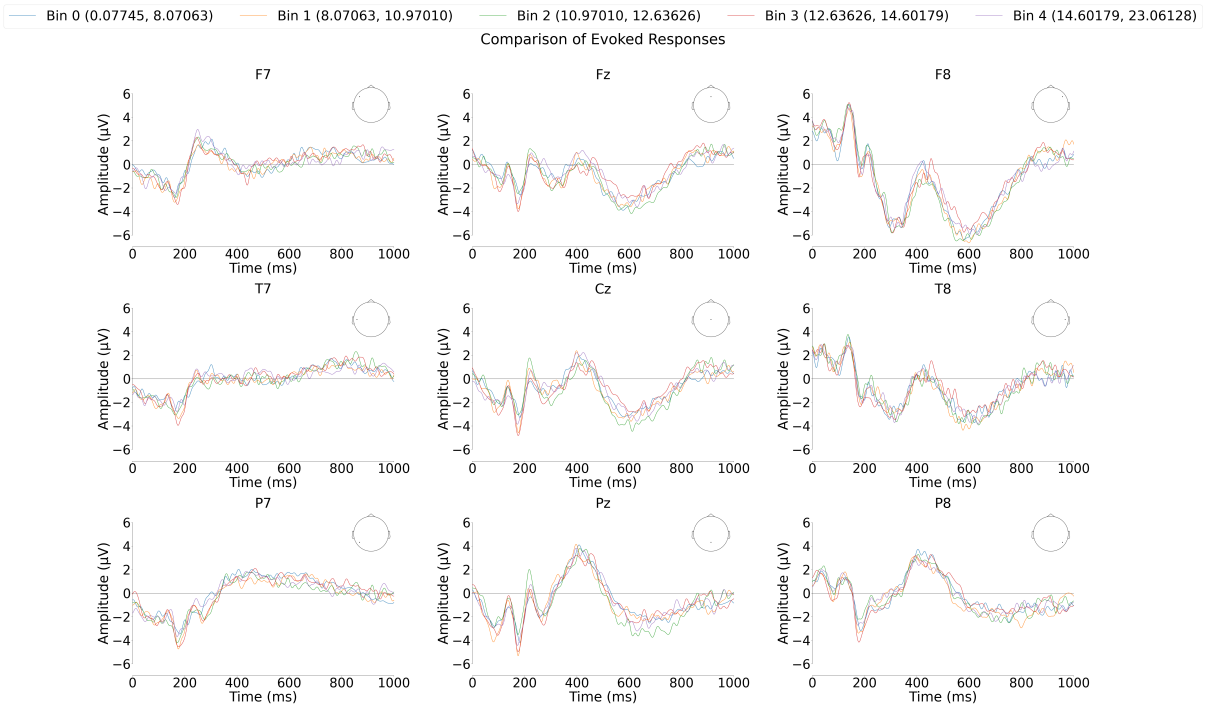

Figure 3.1: Comparison of evoked responses to words with varying surprisal ratings at multiple electrode sites

therefore, we cannot observe the N400 effect. Instead, significant peaks can be seen at 600 ms in the ERP plots for both Cz and Pz electrodes. This shift could be due to the experiment setup, where participants had to fixate on a dot on the right side on the screen to advance to the next sentence. This required them to move their gaze back to the start of the new sentence, which could have taken around 200 ms. This delay in gaze shift could explain the negative peak observed at 600 ms, which might be the N400 component shifted by 200 ms. Even if the peak at 600 ms represents a time-shifted N400 component, Figure 3.2 does not show significant differences between the surprisal levels. This lack of variation suggests that there is no observable effect of surprisal on the N400 component. This finding contradicts our initial hypothesis and expectations.

The P600 component, associated with syntactic reanalysis and integration, can be seen around 600 ms after stimulus onset. We hypothesized that words with higher surprisal ratings may also affect syntactic processing, requiring the reader to reanalyze and integrate the unexpected linguistic elements, which would trigger larger P600 amplitudes. Figure 3.2 shows a negative peak at 600 ms meaning that we do not see the P600 component in the plots. However, if we consider the same delay that shifted the N400 component, we should expect to observe the P600 component around 800 ms. No significant positive peaks were observed at 800 ms in the ERP plots for both electrodes, suggesting that we did not find evidence of the P600 component in our data.

As seen in Figure 3.2 there were no visible differences in the ERP plots between different surprisal bins for both the Cz and Pz electrodes. One possible reason for this lack of significant differences could be the nature of the probabilities calculated by GPT-2. As shown in Figure 4.1, the probabilities were very small and similar to each other. The mean probability calculated was 0.017 with a standard deviation of 0.091. These results reflect the inherent difficulty of predicting the next word given the context. This challenge arises due to the many potential words that could logically and semantically fit into any given sentence. This makes

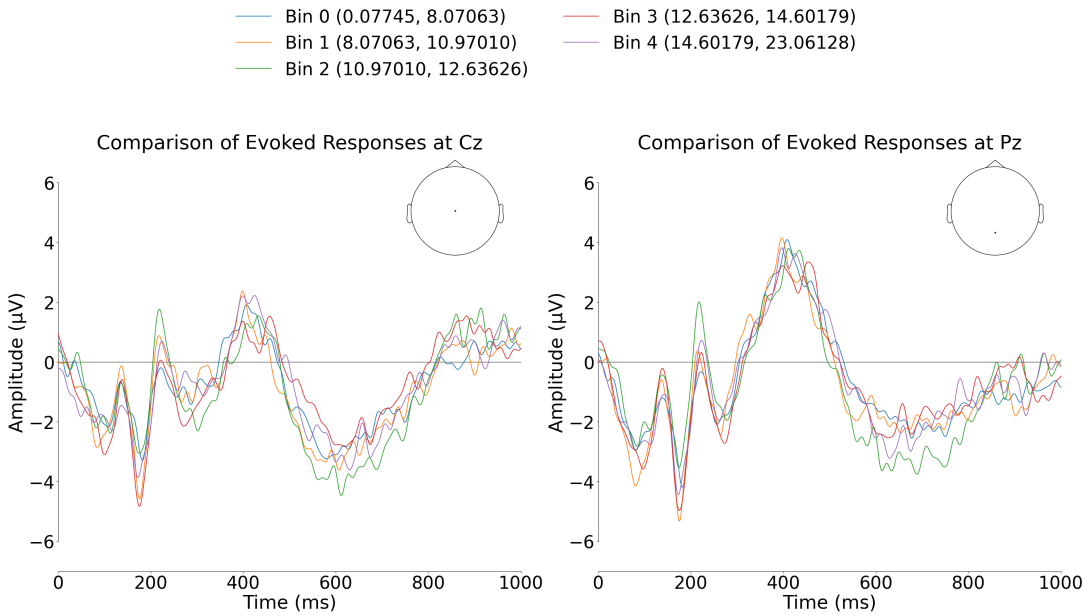

Figure 3.2: Comparison of evoked responses to words wit varying surprisal ratings at central (Cz) and parietal (Pz) electrode sites

all words somewhat unexpected and makes it very hard to calculate these probabilities. These probabilities were used to calculate the surprisal ratings, suggesting that the differences in surprisal were not significant enough to produce distinct neural responses. This similarity in the probabilities could have resulted in a lack of variation in the ERP amplitudes across different surprisal conditions.

4.2 Eye-tracking analysis

The analysis of pupil dilation was conducted to asses cognitive load during reading. Pupil size increases with cognitive effort, providing a physiological marker of cognitive load. We hypothesised that more surprising words would lead to greater pupil dilation, indicating increased cognitive load. As shown in Figure 3.3, we notice that bin 0 (lowest surprisal) showed unexpectedly high and anomalous pupil dilation compared to the other bins. If we exclude bin 0, we observe that bins 2, 3 and 4 consistently show higher pupil dilation that bin 1. This patterns matches our prediction that higher surprisal ratings (less predictable words) would result in greater cognitive load. Bin 0 starts at a higher level in the graph compared to the other bins, even though a baseline correction was applied. The fact that the other bins are closer together at the start further indicates that the pupil response for bin 0 is anomalous. This anomaly might be due to external factors or individual differences in baseline pupil size that were not fully controlled. Even though the general trend supports our hypothesis, these results are not conclusive, as the anomaly in bin 0 suggests that other factors could be influencing pupil dilation.

The differences observed in pupil dilation across conditions might be attributed to the slower response time of pupil dilation compared to EEG changes. Studies show that pupil dilation is influenced by previous trials, suggesting that the pupil's response integrates information form prior experiences rather than only responding to immediate stimuli (Van der Wel & Van Steenbergen, 2018). Pupil dilation tends to accumulate context over a longer period of time, more like GPT-2, which uses the context to calculate the probability of the first word. This could mean that the pupil dilation reflects a more integrated cognitive load over several words or sentences, while EEG reflects more immediate neural responses to individual words. This temporal difference might explain why we observed

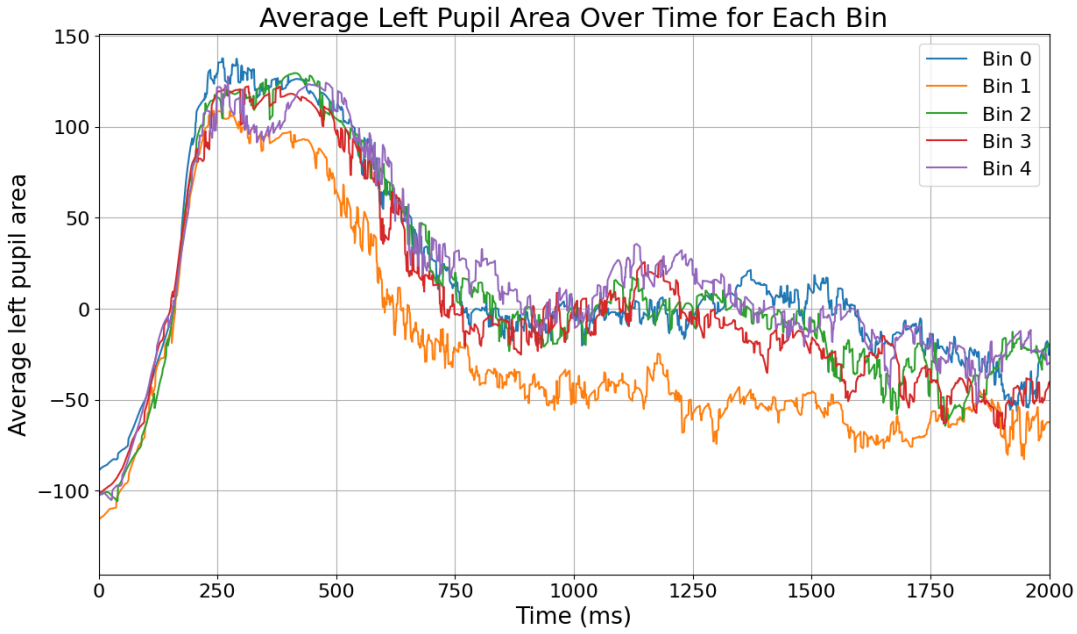

Figure 3.3: Average left pupil area change over time for each bin of surprisal ratings

trends in pupil dilation that were not apparent in the ERP data.

Among those external factors that could have influenced the pupil size are luminance and fatigue (Kiefer et al., 2016). Even if participants could take breaks after every chapter, this study involved prolonged reading, which could have led to fatigue influencing their pupil responses. Additionally, during mind wandering the pupil may not respond to stimuli or may show unusual patterns of dilation Smallwood et al. (2011). Given the length of the reading task in our study, it is possible that participants experienced periods of mind wandering that led to irregular pupil responses. This could have resulted in pupil dilation patterns that did not correlate with cognitive load induced by the stimuli.

4.3 Stimulus characteristics

A critical aspect of our findings revolves around the nature of the words used to calculate the surprisal ratings. Notably, some of the first words in the sentences shown as stimuli were stop words (e.g., "the", "and", "is"). Stop words are very frequent in language but carry little meaning by themselves (John Snow Labs, 2024). They are not as surprising as more meaningful words like nouns and verbs. Focusing on nouns or contextually rich words that carry more meaning and are more surprising could produce more interesting data about the cognitive processes involved in the brain's predictive mechanisms.

4.4 Limitations

Some limitations must be accounted for in this study. Firstly, the small sample size (N=2) significantly limits the generalizability of the findings. The lack of variability observed between the two subjects suggests that the pattern observed might not reflect the typical pattern seen in a larger and more diverse population. The small sample size also results is low statistical power, making it difficult to draw robust conclusions from the data.

We used GPT-2 to calculate the probabilities assuming that these surprisal rating would reflect the brain's prediction mechanisms. GPT-2 may not fully align with how the brain processes and predicts upcoming words. This assumption was made because we do not fully understand how the brain's prediction mechanism fully works. The debate mentioned in our introduction highlights that we do

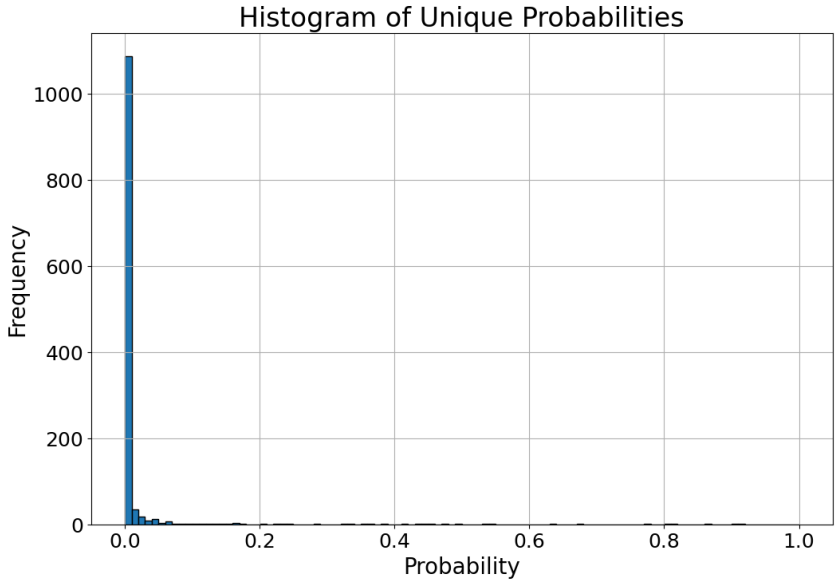

Figure 4.1: Probability distribution of the first word of each sentence calculated by GPT-2.

not know if predictive mechanisms are central to language comprehension or if they happen at all. If this assumption is incorrect, the comparison between the surprisal ratings calculated and the observed neural and physiological responses would be unreliable. However, Heilbron et al. (2022) did use GPT-2 and found evidence that brain responses are modulated by linguistic predictions. This indicates that GPT-2 can allign with neural activities to some extent, supporting its use in exploring how surprisal and predictability influence cognitive processes during language tasks.

Additionally, it's important to consider that GPT-2 was trained using modern language data whereas "Animal Farm" by George Orwell was written in 1945 using some outdated language. This could have resulted in some unusual probabilities that led to ratings that did not accurately reflect the surprisal of the participant.

Furthermore, the nature of the stimuli may have influenced the results. The stimuli were sentences from the novel "Animal Farm". Novels generally contain coherent and contextually consistent language. This consistency could have led the brain's prediction mechanisms to anticipate words that were not so far off from the actual word shown. This resulted in lower prediction errors than those that would arise from encountering completely unexpected words unrelated to the context. As seen in Figure 4.1, the probabilities calculated were very similar to each other, meaning that the surprisal ratings did not differ significantly. This is a major limitation, as it could be the reason why we don't see clear differences between the different levels of surprisal. Previous experiments used highly surprising sentences to trigger stronger cognitive and neural responses. Aurnhammer et al. (2021) conducted an experiment to investigate the effects of expectancy and lexical association on language processing, focusing on the N400 and P600 components of comprehension. They used sentences such as "Before he stacked the wood, the lumberjack sharpened the axe" to create an expected context and "Before he stacked the wood, the lumberjack ate the axe" to create an unexpected context. The second sentence is much more surprising than a typical sentence from a novel, as the verb does not fit the context at all. Sentences like these are more likely to trigger large surprisal or prediction errors in the brain. This difference in the level of surprisal might be a reason why we did not observe the N400 and P600 effects as prominently in our study.

Lastly, while our study emphasized natural reading conditions, the experimental setup does not replicate how people typically read at home. Using an EEG cap, chin rest, and a new mechanism to

read and advance sentences using the eye-tracker might have affected the results. This setup could have affected participants' reading behavior and cognitive load. Despite a training phase, using the mechanism for reading was likely unfamiliar, potentially influencing the observed neural and physiological responses.

The EEG records immediate brain activity and the participants had to fixate on the fixation dot (shown in Figure 2.1 for 0.5 seconds to switch to the next sentence. It is possible that the brain predicted the next word, and by the time the screen switched, the neural activity showed something else. The brain's prediction might have happened as soon as the participant finished reading the sentence and not when the screen switched. If assumption is true, the brain activity recorded from when the new sentence showed up could have been showing something else rather than the prediction hat happened right after the participant read the previous sentence. Even though the fixation time was very brief, participants could still have started to mind wander. This likelihood increases if the book's content is perceived as less engaging. The fact that participants had to stare at a dot and wait was not very natural and could have significantly affected the results.

4.5 Future research

Future research should address these limitations and expand on the findings to get a better understanding of the brain's predictive mechanisms in language comprehension.

Including more participants would increase the statistical power and allow for more robust conclusions. Diverse participant demographics could also provide insights into individual differences in language processing.

The use of GPT-2 for calculating word surprisal was based on the assumption that its predictions would align with the brain's predictive mechanisms. Future studies should explore other models and methods for quantifying surprisal and explore how that would affect the results. Ren et al. (2024) found a strong correlation between large language models and brain similarity, and their findings indicate that the similarity increases with the size of the models. With this information, it would be interesting to see how a larger model would affect the surprisal ratings and results of our experiment.

In this study we used a novel, so the sentences shown contained coherent and contextually consistent language which might not trigger significant prediction errors. Future research could create a story including sentences with higher levels of unexpectedness by using words that do not fit the previous context. This could trigger more pronounced cognitive responses and provide more insight into the brain's predictive mechanism. Additionally, Sarica & Luo (2021) state that a standard step of NLP tasks is the removal of stop words to focus on more contextually rich words. Focusing on more meaningful words, such as nouns or verbs, is important because these words carry significant meaning in context. They could trigger higher prediction errors in the brain compared to stop words, which carry little meaning by themselves. It would be interesting to start sentences using these rich contextual words.

This research emphasised natural reading so future research should aim to create an experimental setup to mimic how people read in everyday settings. Extending the training phase for participants would allow more time for them to become accustomed to the experimental setup, reducing the potential for unusual data due to unfamiliarity with the reading interface.

Furthermore, it would be interesting to analyse the ERPs at the point when participants fixate on the last word of the sentence rather than when the new sentence appears. This approach could reveal whether the brain's predictive mechanisms are actively engaged while reading the last word. This change could reveal whether the observed neural activity and pupil dilation were influenced by the waiting period, potentially recording activity unrelated to the predictive mechanisms.

Future research should continue to integrate NLP models with cognitive neuroscience methods. Iterative improvements in NLP models based on cognitive data could enhance their accuracy in replicating brain functions and deepen our understanding of human cognition. This deepening of our knowledge might provide further clarification on the role that the prediction mechanism plays and how significant it is in language processing.

5 Conclusions

This paper aimed to investigate how word surprisal affects cognitive processes during natural reading by using EEG and eye-tracking technologies combined with the predictive capabilities of the GPT-2 language model. This project was challenging as we do not fully understand how the brain's prediction mechanisms work and recreating natural reading conditions in a controlled experimental setup is complex. Accurately modelling the brain's prediction mechanisms using GPT-2 is particularly difficult due to the complex and dynamic nature of the human brain.

We hypothesised that higher surprisal ratings (less predictable words) would result in more pronounced N400 and P600 responses, and increased pupil dilation, reflecting greater cognitive load. However, we did not directly observe the expected N400 (semantic processing) and P600 (syntactic reanalysis) components in our EEG analysis, there was a potential shift in the N400 component. However, we did not find an effect on word surprisal level on this shifted component. This suggests there was no effect of the degree of surprisal on semantic processing. The pupil data showed more of a correlation between word surprisal and cognitive load, but these results were not conclusive due to some anomalies in the data. Despite the small sample size of only two participants this study produced interesting results, providing a foundation for future research.

This research built the groundwork for understanding how word surprisal affects natural reading, expanding beyond previous experiments that focused on specific words or sentences. By analysing the limitations of this experiment, we have identified areas for improvement that could make future studies more robust and successful. Reproducing this experiment with these improvements would be very interesting and could provide more conclusive insights into the brain's prediction mechanisms during natural reading.

References

Alamia, A., VanRullen, R., Pasqualotto, E., Mouraux, A., & Zenon, A. (2019). Pupil-linked arousal responds to unconscious surprisal. *Journal of Neuroscience*, *39*(27), 5369–5376.

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., & Crocker, M. W. (2021). Retrieval (n400) and integration (p600) in expectation-based comprehension. *Plos one*, *16*(9), e0257430.

Brouwer, H., & Crocker, M. W. (2017). On the proper treatment of the n400 and p600 in language comprehension. *Frontiers in psychology*, *8*, 1327.

Brown, C., & Hagoort, P. (1993). The processing nature of the n400: Evidence from masked priming. *Journal of cognitive neuroscience*, *5*(1), 34–44.

Delogu, F., Brouwer, H., & Crocker, M. W. (2021). When components collide: Spatiotemporal overlap of the n400 and p600 in language comprehension. *Brain Research*, *1766*, 147514.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, *8*(8), 1117–1121.

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological reviews*, *91*(4), 1357–1392.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain research*, *1626*, 118–135.

John Snow Labs. (2024). Text cleaning: Removing stopwords from text with spark nlp. *John Snow Labs*. Retrieved from `https://www.johnsnowlabs.com`

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, *82*(3), 3713–3744.

Kiefer, P., Giannopoulos, I., Duchowski, A., & Raubal, M. (2016). Measuring cognitive load for map tasks through pupil diameter. In *Geographic information science: 9th international conference, giscience 2016, montreal, qc, canada, september 27-30, 2016, proceedings 9* (pp. 323–337).

Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., & Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PloS one*, *13*(9), e0203629.

Kutas, M., & Federmeier, K. D. (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.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... others (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, *7*, e33468.

Ren, Y., Jin, R., Zhang, T., & Xiong, D. (2024). Do large language models mirror cognitive language processing? *arXiv preprint arXiv:2402.18023*.

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *Plos one*, *16*(8), e0254937.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(45), 1–12.

Smallwood, J., Brown, K. S., Tipper, C., Giesbrecht, B., Franklin, M. S., Mrazek, M. D., ... Schooler, J. W. (2011). Pupillometric evidence for the decoupling of attention from perceptual input during offline thought. *PloS one*, *6*(3), e18298.

Šoškić, A., Jovanović, V., Styles, S. J., Kappenman, E. S., & Ković, V. (2022). How to do better n400 studies: reproducibility, consistency and adherence to research standards in the existing literature. *Neuropsychology Review*, *32*(3), 577–600.

Steinhauer, K., Connolly, J. F., Stemmer, B., & Whitaker, H. A. (2008). Event-related potentials in the study of language. *Concise encyclopedia of brain and language*, 91–104.

Van der Wel, P., & Van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic bulletin & review*, *25*, 2005–2015.

Vo, T., & Gedeon, T. (2011). Reading your mind: Eeg during reading task. In *Neural information processing: 18th international conference, iconip 2011, shanghai, china, november 13-17, 2011, proceedings, part i 18* (pp. 396–403).

Yang, X., Chen, S., Chen, X., & Yang, Y. (2015). How distance affects semantic integration in discourse: Evidence from event-related potentials. *Plos one*, *10*(11), e0142967.