

Using EEG Signals to Examine Next-Word Predictability of Language Models in Reading

Mai-Boi Quach, B.Sc

Supervised by Asst. Prof. Graham Healy and Prof. Cathal
Gurrin



A thesis presented for the degree of Doctor of Philosophy

SCHOOL OF COMPUTING
DUBLIN CITY UNIVERSITY

December 2024

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Singed: **Boi Mai Quach**

Student ID No.: 20215661

Date: 2024-12-05

Acknowledgements

First of all, I would like to express my sincere appreciation to Graham Healy and Cathal Gurrin for being exceptional supervisors, patiently guiding me through the world of academia. Their continuous support, insightful guidance, and the scientific freedom they provided within a lively, honest, and stimulating research environment at Dublin City University were decisive for the success of my PhD research. Their mentorship has profoundly shaped me into a better researcher.

My gratitude extends to the members of my committee for taking the time to assess this thesis, and more importantly, for their contributions to my work and further research.

I am deeply grateful to my parents, my sister, and my friends, who provided invaluable emotional support and timely distractions whenever I needed them most. Their persistent help put my research endeavours into perspective and ensured my journey remained balanced and enjoyable.

Special appreciation goes to ML-Labs for offering several research seminars and industry talks. These events significantly enhanced my understanding and kept me at the forefront of cutting-edge techniques in Artificial Intelligence and Machine Learning. Thank you for the fun times during this PhD adventure.

Finally, the completion of this PhD program is not truly an ending, but rather the beginning of an exciting new adventure. Thanks to these enriching four years, I have grown into a better version of myself to embrace future challenges and opportunities. The invaluable experiences, mentorship, friendships, and lessons learned during this journey will continue to guide me throughout my career and life.

Contents

1	Introduction	3
1.1	Do Machines and Humans Behave in Similar Ways During Reading? .	3
1.2	Motivation	5
1.3	Contributions	9
1.4	Impact	10
1.5	Thesis Organisation	11
1.6	Author's Publications	13
2	Related Work and Background	15
2.1	Introduction	15
2.2	Neural Responses to Next-Word Prediction	16
2.2.1	ERPs in Neurolinguistics	16
2.2.2	Semantic ERP Effects in Reading	18
2.3	LMs for Next-Word Prediction Task	19
2.3.1	N-Gram Models	19
2.3.2	Transformer-Based Models	20
2.4	How Might LMs Be Related to Humans via Next-Word Predictability? .	23
2.4.1	Linguistic Prediction at the Computational Level	24
2.4.2	Encoding Brain Activity Using Surprisal	25
2.5	Conclusion	28
3	Experimental Design and Data Preparation for EEG-Based Reading Research: A Comprehensive Overview and Practical Example	29

3.1	Introduction	30
3.2	Experimental Design	31
3.2.1	RSVP-Based Paradigms	32
3.2.2	Advantages and Disadvantages of RSVP Paradigms	35
3.2.3	Important Features of EEG-Based Experimental Design	37
3.2.4	Synchronisation of Data Streams	38
3.3	Data Preprocessing	41
3.3.1	Common Artifacts	42
3.3.2	EEG Data Cleaning	46
3.4	A Practical Example of EEG-Based Reading Research	53
3.4.1	Experimental Design	53
3.4.2	Data Preparation	58
3.4.3	Comparison Procedures	60
3.4.4	Results	62
3.5	Discussion	65
4	Dublin EEG-Based Reading Experiment Corpus	67
4.1	Introduction	68
4.2	Method	71
4.2.1	Materials	71
4.2.2	Behavioural Word-Prediction Experiment	73
4.2.3	Data Acquisition	76
4.2.4	Data Preprocessing	76
4.3	Technical Validation	79
4.3.1	Behavioural Word-Prediction Dataset	81
4.3.2	Neural Responses in Reading	82
4.4	Discussion	86
4.4.1	Conclusion	86
4.4.2	Limitations	87

5 Decoding EEG Signals to Explore Next-Word Predictability in the Human Brain	89
5.1 Introduction	90
5.1.1 Next-Word Prediction Behaviour Observation	90
5.1.2 ERPs Effects in Lexical Classes	91
5.1.3 ERPs Effects in Grammatical Categories	91
5.1.4 Traditional ERPs Analysis Techniques	92
5.1.5 Classification-Based Decoding Techniques	93
5.2 Method	94
5.2.1 EEG Recording and Data Preparation	94
5.2.2 Univariate Analysis	96
5.2.3 SVM-Based Decoding Analysis	98
5.2.4 Analysis Procedures	101
5.2.5 Comparison of Performance Between Methods	103
5.3 Results	106
5.3.1 High vs. Low Cloze Probability: Analysis Over Word Classes .	106
5.3.2 High vs. Low Cloze Probability: Analysis Over Grammatical Categories	111
5.3.3 Results of Size Effects	116
5.4 Discussion	119
5.4.1 Key Findings and Interpretation	119
5.4.2 Theoretical Implications	121
5.4.3 Limitations	122
5.4.4 Conclusion	123
6 Encoding EEG Signals to Examine Human-Like Next-Word Prediction Behaviour in Language Models	124
6.1 Introduction	125
6.2 Method	127
6.2.1 Stimuli and Data Preparation	127

6.2.2	Top-1 Prediction Estimation	130
6.2.3	Surprisal Estimation	131
6.2.4	Brain Encoding Models	132
6.2.5	Regression Modelling	134
6.2.6	Significance Tests	135
6.3	Results	136
6.3.1	Performance and Human Alignment in Different Language Models	136
6.3.2	Neural Encoding Using Predictive Metrics from Human Cloze	140
6.3.3	Surprisal Estimate for All Encoding Models	141
6.3.4	Neural Correlates of Surprisal in GPT-2 Large Regression . .	145
6.4	Discussion	147
6.4.1	Key Findings and Interpretation	147
6.4.2	Limitations	148
7	Conclusion and Future Work	150
7.1	Ethical Considerations	150
7.1.1	Dual Use	150
7.1.2	Observational Bias	151
7.1.3	Protecting Privacy	152
7.2	Summary	152
7.2.1	Contributions	152
7.2.2	Research Question 1	153
7.2.3	Research Question 2	154
7.2.4	Research Question 3	155
7.3	Direction for Future Research	157
7.3.1	Unfolding Reading Processing in the Human Brain and LMs .	157
7.3.2	Translation of Thoughts to Text	159

Appendix	162
A Correlation	162
A.1 Pearson Correlation	162
A.2 Spearman Correlation	163
B Smoothing Techniques in N-Grams	163
B.1 Laplace Smoothing	163
B.2 Kneser–Ney Smoothing	164
C DERCo Dataset	166
C.1 DERCo Ethic Approval	166
C.2 Questionnaire for Five Articles	166
C.3 Transcripts of Five Articles	178
C.4 Online Consent Form for Amazon Mechanical Turk Experiment.	185
C.5 Human Next-Word Prediction Performance	186

List of Figures

1.1	Word cloud of keywords presented in this dissertation.	10
1.2	Reading threads of the thesis.	12
2.1	Deriving ERPs from EEG data.	17
2.2	The diagram of a self-attention layer in Transformer models illustrates how it learns to predict the next token in a sequence, as adapted from Li et al. [73]. A single-layer self-attention model performs a form of hard retrieval, precisely selecting high-priority tokens and then producing a convex combination of these as the output, from which the next token can be sampled.	22
3.1	Illustration of the word presentation for the first two words, “Once” and “upon”, in the common phrase “Once upon a time” in a typical RSVP experiment. The illustrated experiment begins with a blank screen and a fixation cross in the centre of the screen for 1000 ms. Each stimulus word is then presented for 200 ms, followed by a 300 ms blank screen.	33
3.2	Illustration of the RSVP-with-flankers paradigm with valid (left) and invalid (right) preview.	35
3.3	Illustration of the RSVP-with-flankers paradigm with parafoveal congruent (left) and incongruent (right) words.	36
3.4	Block diagram of the 1) shared triggers and 2) photodiode sensor techniques in synchronisation of data streams.	40

3.5	General preprocessing pipeline for EEG data. Epochs can be implemented before (dashed line) or after artifact removal.	42
3.6	Illustration of different physiological artifacts.	42
3.7	Block diagram of the blind source separation.	51
3.8	The procedure of word presentation for the first three words “Once upon a ...” in (a) RSVP experiment (b) RSVP-with-flanker experiment.	56
3.9	Our EEG event trigger board used in the DERCo project: The red-circled area connects to the EEG device to send triggers, while the blue-circled area connects to the photodiode, which detects light changes to mark stimulus onset.	56
3.10	Distribution of signal to noise ratios across channels (compared to Raw Epochs).	62
3.11	Decoding accuracies according to each pipeline.	64
4.1	The stimulus was transcribed in article 1 for the behavioral experiment. (a) Transcript (b) A ten-word sliding window example.	74
4.2	A ten-word sliding window was presented on the website for the article 1. (a) The next word is “the” (b) The next word is “next.”	74
4.3	The EEG epochs data preprocessing pipeline.	78
4.4	The structure of the dataset	80
4.5	Butterfly plots for EEG averages across all words in article 1 (a) across subjects (b) single subject (JPY86). Each time-series line graph corresponds to an electrode.	84
4.6	Scalp distributions of the response for low predictable words, high predictable words, and the difference between low and high predictable words (low minus high) in (a) RSVP experiment and (b) RSVP-with-flanker experiment. Amplitudes (μV) were averaged in 300-500 time window (for all electrodes).	85

4.7	ERP activity for midline electrodes in RSVP experiment (left) and RSVP with flankers experiment (right). The time range of significant differences ($p < 0.05$) between 300ms and 500ms post-stimulus are indicated.	86
5.1	Flowchart for SVM-based decoding	103
5.2	(a) ERPs plotted separately for content words (top row) and function words (bottom row) across the left hemisphere, midline, and right hemisphere electrode sites. ERPs are shown for high cloze (red) and low cloze (green) conditions. Gray shading indicates time points at which the values are significantly different from chance (p-values < 0.05) with a two-tailed test in the direction of the expected effect. (b) Raincloud plots displaying the absolute difference in N400 amplitude (measured from 300-500 ms) between content and function words across the left hemisphere, midline, and right hemisphere electrode regions. The plots combine boxplots, density distributions, and individual data points for both conditions. Note that for the content/cloze condition (green), ERPs to content words were averaged across cloze levels, while function/cloze condition (orange) represents function words across cloze levels. The lines connect the mean amplitude differences across these regions.	107

5.3 (a) Decoding accuracy for distinguishing between high and low cloze probability conditions plotted separately for content words (top row) and function words (bottom row) across the left hemisphere, midline, and right hemisphere electrode sites over time. Red shading indicates time points at which the value is significantly different from chance (p-values < 0.05) with a one-tailed test in the direction of the expected effect. (b) Raincloud plots showing decoding accuracy for content words (left) and function words (right) across the left hemisphere, midlines, and right hemisphere electrode regions during the N400 time window (300-500 ms). The plots include boxplots, density distributions, and individual data points. The black dashed line represents chance-level decoding (0.5), while the red lines indicate the mean decoding accuracy across regions. 109

5.4 (a) Grand average ERPs plotted separately for nouns (top row) and verbs (bottom row) across the left hemisphere, midline, and right hemisphere electrode sites. ERPs are shown for high cloze (red) and low cloze (green) conditions. Gray shading indicates time points at which the values are significantly different from chance (p-values < 0.05) with a two-tailed test in the direction of the expected effect. (b) Raincloud plots displaying the absolute difference in N400 amplitude (measured from 300-500 ms) for nouns and verbs across the left hemisphere, midline, and right hemisphere electrodes. The plots combine boxplots, density distributions, and individual data points for both conditions. Note that for the Noun/Cloze condition (green), ERPs to nouns were averaged across cloze levels, while Verb/Cloze condition (orange) represents verbs across cloze levels. The lines connect the mean amplitude differences across these regions. (c) Raincloud plots showing decoding accuracy for distinguishing between high and low cloze conditions during the N400 time window (300-500 ms) for nouns and verbs across the left hemisphere, midline, and right hemisphere electrode regions. The plots include boxplots, density distributions, and individual data points. The black dashed line represents chance-level decoding (0.5), while the red lines indicate the mean decoding accuracy across regions. 112

5.5	(a) Grand average ERPs plotted separately for determiners, pronouns, adpositions, and auxiliaries under high cloze (red) and low cloze (green) conditions across left hemisphere, midline, and right hemisphere electrode sites. Gray shading indicates time points at which the values are significantly different from chance (p-values < 0.05) with a two-tailed test in the direction of the expected effect. (b) Raincloud plots displaying the absolute difference in N400 amplitude (measured from 300-500 ms) for for each word class (determiner, pronoun, adposition, auxiliary) across the left hemisphere, midline, and right hemisphere electrodes. The plots combine boxplots, density distributions, and individual data points for both conditions. Note that for the Det/Cloze condition (green), ERPs to determiners were averaged across cloze levels, similar to Pron/Cloze (orange), Adp/Cloze (purple), and Aux/Cloze (pink). The lines connect the mean amplitude differences across these regions. (c) Raincloud plots showing decoding accuracy for distinguishing between high and low cloze conditions during the N400 time window (300-500 ms) for for each word class across the left hemisphere, midline, and right hemisphere electrodes. The plots include boxplots, density distributions, and individual data points. The black dashed line represents chance-level decoding (0.5), while the red lines indicate the mean decoding accuracy across regions.	115
5.6	Effect size (Cohen's d_z) drawn on the same scale for each analysis in 300-500 ms time-window comparisons. The error bars show ± 1 standard error.	117
6.1	Linear encoding model used to predict the neural responses to each word in the narrative before and after word-onset.	133

6.2	Performance comparison of various language model families (N-grams, GPT-2, GPT-Neo) and human benchmarks in a next-word prediction task, evaluated by per-model accuracy and joint prediction percentages “both correct” (green line) and “both incorrect” (red line).	136
6.3	Correlations between regression model results estimated from human cloze probabilities and neural responses for content and function words across time, shown for both lexical surprisal (left) and top-1 prediction (right) metrics. The encoding analysis was conducted for each electrode with significant encoding of predictive metrics and then averaged across electrodes. Asterisks indicate window time at which the value is significantly different ($p < 0.001$) with cluster-based permutation tests in the direction of the expected effect for content words (green line) and function words (purple line). The shaded regions represent the between-subject SEM (standard error of the mean) of the encoding models.	140
6.4	EEG encoding model comparison results using surprisal for content words (left) and function words (right) averaged across all channels over time (from -100 to 500 ms) for all participants. Significance levels indicate statistical differences between models, with $p < 0.05(*)$, $p < 0.01(**)$, and $p < 0.001(* **)$ based on two-tailed paired t-tests or Wilcoxon signed-rank tests. Error bars represent the SEM.	141
6.5	Surprisal-based encoding model comparison results for content words across all channel for individual participants. Significance levels indicate statistical differences between models, with $p < 0.05(*)$, $p < 0.01(**)$, and $p < 0.001(* **)$ based on two-tailed paired t-tests or Wilcoxon signed-rank tests. Error bars represent the SEM.	143

6.6	Surprisal-based encoding model comparison results for function words across all channel for individual participants. Significance levels indicate statistical differences between models, with $p < 0.05(*)$, $p < 0.01(**)$, and $p < 0.001(* **)$ based on two-tailed paired t-tests or Wilcoxon signed-rank tests. Error bars represent the SEM.	144
6.7	Correlations between regression model results estimated from GPT-2 Large model and neural responses for content and function words across time, shown for both lexical surprisal (left) and top-1 prediction (right) metrics. The encoding analysis was conducted for each electrode with significant encoding of predictive metrics and then averaged across electrodes. Asterisks indicate window time at which the value is significantly different ($p < 0.001$) with cluster-based permutation tests in the direction of the expected effect for content words (green line) and function words (purple line). The shaded regions represent the between-subject SEM (standard error of the mean) of the encoding models.	145
6.8	Full EEG Topographies of the effects of lexical surprisal. The topographies show the Pearson r of encoding prediction correlations (upper row) and the respective TFCE-corrected significance $-\log_{10}(p)$ (lower row).	146
1	Online Consent Form for Behavioural Word-Prediction Experiment. .	185

List of Tables

2.1	Hyper-parameters of GPT-2, GPT-Neo families. n_{layers} , n_{head} , d_{model} , and n_{params} respectively refer to number of layers, number of attention heads per layer, embedding size, and number of parameters. . . .	23
3.1	Comparison between photodiode and shared trigger approaches . . .	41
3.2	Baseline number of trials for each preprocessing pipeline (AutoReject, FASTER, and Riemannian Potato) across high and low conditions for each subject. Highlighted cells indicate the baseline numbers used for comparative analysis.	60
3.3	Mean (and standard deviation) of performance for the described cleaning algorithms on different task.	65
4.1	Descriptive statistics of reading materials (M = mean, SD = standard deviation).	72
4.2	Human next-word prediction performance. % shows the percentage of correctly predicted words for each story and part of speech. “No. words ($p \geq 0.5$)” shows the number of words with a cloze probability above .5.	82

5.1	Results of the univariate analysis procedure. N refers to the number of participants. Baseline no. of trials refers to the number of trials used for analyses in each participant. LH, M, RH represent the Left Hemisphere Electrode, Midline Electrode, and Right Hemisphere electrode sites, respectively. Cluster p-value denotes the p-value threshold used for cluster-based permutation tests.	99
5.2	Information about the SVM-based decoding procedure for each case. N refers to the number of participants. No. of folds indicates the number of folds used in Stratified K-fold cross-validation. L iteration represents the number of decoding iterations for each participant. Baseline no. of trials refers to the number of trials used for decoding in each participant. Note that a p-value threshold of 0.05 was used for the cluster-based permutation tests.	104
6.1	Correlation comparisons of various language model families (N-grams, GPT-2, GPT-Neo) against human benchmarks in a next-word prediction task. Metrics include Pearson and Spearman correlation coefficients r for top-1 prediction and surprisal. Statistically significant correlations with $p^* = 0.001$ are indicated in two-sided test.	138
1	Human next-word prediction performance. % shows the percentage of correctly predicted words for each story and part of speech.	186

List of Abbreviations

AI	Artificial Intelligence
AR	AutoReject
BSS	Blind Source Separation
CNNs	Convolutional Neural Networks
ECG/EKG	Electrocardiography
ECoG	Electrocorticography
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
ERPs	Event-related Potentials
FASTER	Fully Automated Statistical Thresholding for EEG Artifact Rejection
FDR	False Discovery Rate
fMRI	Functional Magnetic Resonance Imaging
GPT	Generative Pre-trained Transformer
HIT	Human Intelligence Task
ICA	Independent Component Analysis
ICs	Independent Components
KS	Kolmogorov-Smirnov
LDA	Linear Discriminant Analysis
LLMs	Large language models
LMs	Language Models
LSTM	Long Short-Term Memory
MEG	Magnetoencephalography
ML	Machine Learning
MVPA	Multivariate Pattern Analysis
NLP	Natural Language Processing
OSF	Open Science Framework
RBF	Radial Basis Function
RF	Random Forest
RNN	Recurrent Neural Network
RSVP	Rapid Serial Visual Presentation
SNR	Signal-to-noise Ratio
SVM	Support Vector Machines
TFCE	Threshold-Free Cluster Enhancement

Glossary

Next-Word Predictability	The ability to predict upcoming words.
Language Model	A model that can predict the next word in a given sequence of words.
Human-Like Language Model	A language model that can generate, understand, and interact with human language in a way that closely mimics natural human communication.
Encoding Model	Encoding uses stimuli to predict activity.
Decoding Model	Decoding uses activity to predict information about stimuli.
EEG Epochs	The fixed-length segments of continuous EEG data, usually time-locked to a specific event like a stimulus or a response.
N400 amplitude	The mean voltage, measured in μV , 300–500 ms post onset of stimulus.
N400 latency	The peak delay, measured in milliseconds, in the N400 time window.
Surprisal	The measure of the unexpectedness of a target word.
Top-1 Prediction	The highest probability assigned by the model to be the next word.
Microsaccades	The small, rapid, involuntary eye movements that occur during fixation.

Cloze procedure	The experiment involves asking participants to predict and complete unfinished sentences or passages based on the accumulated preceding context.
Cloze value	The proportion of people who fill a gap in a sentence with that specific word.
Top-down processing	When readers use their prior knowledge, experiences, and expectations to make sense of the text.
Bottom-up processing	Understanding a text by first decoding individual letters, words, and sentences, then building up to grasp the overall meaning.
Content words	Words that carry the most meaning of a text.
Function words	Words that provide grammatical structure but carry little meaning on their own.

Using EEG Signals To Examine Next-Word Predictability of Language Models in Reading

Boi Mai Quach

Abstract

Language Models (LMs) are primarily developed to address different natural language processing (NLP) tasks, and are not intended to reflect human reading comprehension processing. However, since these models are trained with the same written materials that humans also process in reading, they should share many of the same abilities in the reading comprehension processes as a human, particularly in next-word predictability. Comparing the accuracy of different types of LMs and humans in the next-word prediction task demonstrates their predictive capabilities but falls short of confirming whether LMs and the human brain process reading in a similar manner. To address this, the thesis used electroencephalography (EEG) signals to examine next-word predictability of different LMs during reading.

The main contributions of this thesis are as follows. First, we present a comprehensive resource for guiding EEG-based reading experiments, introduce a tailored preprocessing pipeline, and provide DERCo (Dublin EEG-based Reading Experiment Corpus), an openly accessible dataset combining EEG and next-word prediction data. Second, we use DERCo to analyse how the brain responds to various word categories (e.g., content words, function words, and grammatical classes), shedding light on the interaction between top-down and bottom-up processing during reading. We demonstrate an improved decoding methodology that strengthens our analytical capabilities. Additionally, we highlight the effectiveness of a decoding approach that enhances analytical power compared to the traditional event-related potentials (ERPs) in EEG data analysis. Lastly, leveraging surprisal, an information-theoretic metric, alongside accuracy, we build brain encoding models for different LMs and human prediction production to capture neural responses at the word level. Our

evaluation reveals that while more advanced language models exhibit closer alignment with human prediction patterns, they fail to fully reflect the human-like reading processes observed in brain signals.

Chapter 1

Introduction

In the opening chapter of this dissertation, we provide an introduction to the field of cognitive plausibility and computational language models' ability in reading processing and explain the rationale behind pursuing this area of research. Furthermore, we state the research questions to be addressed in the scope of this thesis and then highlight the contributions and impact of the presented work.

1.1 Do Machines and Humans Behave in Similar Ways During Reading?

Humans invented reading and have passed down this complex skill across generations through language. Gough and Tunmer [1], and Hoover and Gough [2], pointed out one of the first theories of reading, also known as the Simple View of Reading. This theory indicates that both the ability to decode words and comprehend language are required to master reading comprehension.

When reading a word without prior context, the brain engages in what is known as a “bottom-up” process, beginning with letters and later accessing higher-level linguistic structures (lexical word groups, grammatical categories). As readers focus on these linguistic units, they begin to predict next words that these units form. Once the words are recognised, they are translated into inner speech, from which the

reader extracts meaning. In this process, reading comprehension is often considered an automatic outcome of accurate word recognition [3, 4].

Top-down processing is the opposite approach. This standpoint holds that word comprehension occurs in a text, which can guide us to the correct meaning before the word is fully revealed. For example, we can infer that the statement “I drink my coffee with cream and ...” is likely to end with “sugar.” Humans rapidly use top-down cues to disambiguate and predict upcoming words [5–7]. Simultaneously, the bottom-up process in this context also indicates that the upcoming word is more likely to be a noun, given the syntactic structure. It is noteworthy that word identification (related to bottom-up processing) and comprehension components (related to top-down processing) work together and can substantially affect reading comprehension [8, 9].

These two processes develop day by day during an individual’s life via learning and experiences. This suggests similarities to the process of building a language model, which is trained on billions of sentences drawn from diverse online sources, including Wikipedia, books, news, articles, and more. Importantly, next-word prediction is also the primary task used in training these LMs [10–12].

Traditional computational LMs were developed to determine the probability of a given word in a sentence [13]. They predict the most likely next token in a sequence by utilising the preceding context from the training data. The language modelling objective can be optimised using neural language models [14, 15]. By significantly increasing the size of the training corpus and using additional training objective beyond contextual word prediction [16, 17], pre-trained neural language models have rapidly become the best-performing standard for many natural language processing tasks [18]. The recent growth of large language models (LLMs), such as Generative Pre-trained Transformer (GPT) models, has been showing some promise that LMs can mimic the human brain in reading because they attempt to learn language, capture abstract concepts, and afford a remarkable diversity of uses [19].

Although the mechanism to learn reading in the human brain and LMs may

sound similar, LMs do not completely behave like people [19, 20]. For example, given the sequence of words, “Once upon a . . .”, the human brain can instinctively associate it with a fairy tale, imagine a king or queen, predict the next word as “time,” and evaluate whether or not the upcoming word is suitable with the previous context. However, in this case, LMs are trained specifically to predict the next word, such as “time,” but they are not designed to understand long-term narrative structures or truly perceive the scenario [21].

Modern language models, particular LLMs are trained on billions of sentences which far exceeds the average number of sentences a single individual is likely to acquire over their lifetime. Humans, however, understand and generate language with prodigious speed and accuracy. In addition, they can handle ambiguity and artifacts with ease, and are highly adaptable to new domains and inputs.

Hence, there are still many open research questions to be considered at the intersection of interpretability of LMs and cognitive science. Within the scope of this thesis, we aim to leverage decoding and encoding models to investigate the next-word predictability of various LMs and humans. Our goal is to evaluate which LMs most closely reflect human cognitive processes in reading by comparing their performance in predicting EEG signals. In the following section, we present the three main research questions to be explored in this thesis.

1.2 Motivation

Although the field of Artificial Intelligence (AI) has made impressive strides in recent years, LMs still struggle to illustrate an adequate neurally-mechanistic picture of human language processing. However, several existing studies have suggested that the brain is systematically organised in a hierarchy similar to language models [22–26]. Next-word predictability is one of the main factors of the human language processing mechanism, which importantly supports a language model to be cognitively plausible [27]. Meanwhile, previous studies on the cognitive neuroscience of language have tried to explore the connection between the changes in our brain and

manipulations of linguistic representations [28–30]. Still, however, a gold standard theory of how the brain processes language has not been completely agreed upon by researchers [31–33].

Furthermore, the availability of publicly accessible datasets using neural activity data for semantic-level information in reading is limited, which poses a challenge to further discoveries in the field. Designing and conducting reading experiments wherein neural responses are captured is painstaking due to the complexity and time-consuming nature of deploying EEG experiments successfully. These reasons drove us to build and publish the DERCo (Dublin EEG-based Reading Experiment Corpus), which is a language resource combining electroencephalography (EEG) and next-word prediction data obtained from participants reading narrative texts.

The main goal of this work is to leverage brain activity signals from human language processing during reading to unfold next-word predictability. More specifically, we apply emerging brain decoding and encoding approaches to add evidence about the nature of language processing in human brain and LMs during reading comprehension. Because of the similarity between human reading behaviour and language models, we rely on it to evaluate how well humans and different type of language model predict the upcoming words, and then indicate the potential to further align LMs with human cognitive processes. In order to chronologically address these aspects, we begin with the following hypothesis.

Hypothesis *Since language models are trained with the same written materials that humans also process in reading, they should share many of the same abilities in the reading comprehension processes as a human, particularly in next-word prediction ability.*

To effectively investigate this hypothesis, the following three research questions have been defined as a guiding framework of concrete, feasible, and testable milestones for the projects completed in this dissertation.

Research Question 1 (RQ1). **How do we design a standard framework for EEG-based reading experiment?**

In order to simulate reading scenarios, for those researchers who do not have experience in designing experiments in this field, they are frequently overwhelmed because the variety of experimental options are numerous and disorienting. They can struggle to select an appropriate experimental paradigm, determine a suitable sample size, or choose the correct preprocessing techniques for their signals. Therefore, we contribute guidelines for designing EEG-based reading experiments, summarising and analysing several data preparation techniques.

Prompted by challenges that many data troves in neuroscience are not readily available and are often developed under strict experimental conditions such as compliance with the European Union’s General Data Protection Regulation (GDPR), we build an EEG-based reading dataset specifically tailored to investigate the relationship between next-word predictability in language models and neural responses in the human brain. We end with (1) an openly available dataset, (2) a standardised EEG preprocessing pipeline for EEG-based reading datasets, and (3) a combination of annotation data with the EEG-based reading data for next-word predictions.

Beyond addressing this research question, we ultimately hope that this will enable other researchers to run more EEG-based reading experiments. The DERCo dataset is a useful resource for researchers and practitioners to evaluate and reuse within a wide range of neuroscience and AI fields.

Research Question 2 (RQ2). **Are neural responses to upcoming words attributable to quantitative differences in lexical categories (content versus function words) and grammatical categories (e.g., nouns, verbs, adjectives, pronouns, etc.)?**

This research question has previously been examined using traditional ERP approaches [34–36], but no consensus has been reached. Due to the rise of Machine Learning (ML) techniques, we currently use ML-based classifier to decode EEG signals, providing further support for past findings.

In this study, we utilise an EEG-based decoding technique alongside the traditional ERP method to test two competing hypotheses. The first states that whether

differences in word class account for the neural response variations observed at different predictability levels. The second hypothesis examines whether the recognition of grammatical categories, initially generated by bottom-up processes, influences top-down processing.

To understand the contribution of top-down and bottom-up processes during reading comprehension, we used the DERC_o dataset, and modelled neural responses as a function of word class and lexical categories. Furthermore, by using the EEG-based classification approach, we aim to determine whether the decoding-based method enhances effect sizes and statistical power in detecting brain activity differences between conditions during reading, compared to traditional ERP analysis methods.

Research Question 3 (RQ3). Which type of language model most closely aligns with human next-word predictability during reading comprehension?

The training objective of LMs is similar to a psycholinguistic benchmark known as the cloze task [37], which measures a word’s predictability within a given context. However, LMs lack the rich set of experiences that people possess, and humans can demonstrate a high degree of creativity. While LMs are trained to predict the next word, they are not trained to anticipate long-term scenarios. Therefore, although an LM may achieve high accuracy in next-word prediction, it remains unclear whether this high accuracy indicates that the LM behaves similarly to a human in reading processing.

Leveraging information-theoretic surprisal [13], along with neural responses during the reading process, we aim to address the research question by investigating next-word predictability more comprehensively. Surprisal is high when the current word is unexpected given the preceding context. In other words, surprisal reflects a backward-looking cognitive process.

To assess human parity in these models’ training objectives, we first compare the top-1 prediction measures of three language model families, namely n-grams, GPT-

2 [11], and GPT-Neo [38] to those of human participants in the DERCo dataset [39]. Next, we estimate surprisal for the distribution of next-word probabilities using these models. Finally, we build two regressors (also known as brain encoding models), each corresponding to a metric modelling the ERPs elicited by words for different language models and human predictions. The quality of these predictions is then evaluated using the Pearson correlation between the predicted and observed ERPs.

In this chapter, we aim to investigate which LMs exhibit a closer correspondence to human productions through their next-word prediction abilities, particularly in content words. However, this correspondence is not reflected by prediction performance but rather by the way these models correlates with neural response during reading, through the lens of information theory. Moreover, our results also indicated that larger models are not always better at modelling ERPs or capturing neural patterns during reading comprehension.

1.3 Contributions

First, we provide a valuable resource to instruct researchers in conducting EEG-based reading experiments, with detailed explanations and analyses. In addition, we propose an effective EEG preprocessing pipeline for word-level EEG recordings in a narrative context. Furthermore, we leverage our procedure to collect the DERCo dataset as a practical application. Consequently, the DERCo dataset is our outcome of this practical case which is a language resource combining electroencephalography (EEG) and next-word prediction data obtained from participants reading narrative texts. Additionally, DERCo is currently available to the research community and is specifically designed to address our research questions.

Next, we leverage the DERCo dataset to investigate the differential brain responses elicited by various lexical and grammatical categories under human next-word prediction behaviour in reading. As a result, we provide evidence for the interaction between top-down and bottom-up processes in human reading compre-

The word cloud in Figure 1.1 summaries the key terms in this thesis. According to all important key words, we can discern the main objective of this line of research. Applying EEG signals to investigate the reading behaviours and understand the reading processing in different generations of LMs is my Ph.D. study mission. While state-of-the-art LMs exhibit impressive capabilities, questions remain about whether they can truly replicate human reading behaviours or they are just predictive machines. The ultimate goal is to push the boundaries of LMs by bridging the gap between LMs and human through deepening our understanding of how the human brain processes reading and the correlations between their next-word predictability.

The impact of this work is twofold. First, the foundational research conducted benefits multiple disciplines such as AI, neuroscience, and linguistics. It will improve our understanding of human reading behaviours under the information theory and cognitive aspects. This will provide insights into how we can improve computational language models to achieve similar human capabilities. Second, we believe that developing interpretable human-like language models for reading processing will bring great advances to the field of NLP. The insights gained into the cognitive domain of language can be leveraged to build human-like language models that generalise better and are more easily interpretable and cognitively plausible. In a broader context, this research lays the groundwork for future breakthroughs in the field of neuroscience and for making language models more efficient and intuitive.

1.5 Thesis Organisation

This dissertation is organised as follows. Chapter 2 provides a comprehensive overview of our research, structured as a narrative that covers essential content-related backgrounds. This storytelling approach helps readers grasp both the study’s scope and the methodological techniques used throughout the research. In Chapter 3, we present a comprehensive guide to the design and preparation processes involved in EEG-based reading experiments, combining both theoretical insights and

practical cases. Chapter 4 introduces our data collection efforts for the DERCo dataset. We then leverage this dataset to examine the reading comprehension process in the human brain and in LMs. In Chapter 5, we describe a large-scale study on decoding EEG signals to investigate how different types of words and grammar categories activate distinct brain responses during reading comprehension, under human next-word prediction behaviour. Chapter 6 focuses on brain encoding models, which are used to determine which language models most closely mirror human reading predictability when analysed through the lens of information theory. Finally, Chapter 7 discusses the insights gained from previous chapter, their limitations, and the potential for further research directions.

The reading threads in Figure 1.2 are split by topics of interest according to reader’s orientation. For readers who are concerned about in conducting their own EEG-based reading experiments and processing their datasets, follow the chapters indicated by the dotted arrows. If the readers are interested in leveraging neural responses to examine cognitive hypotheses in language processing, then follow the solid arrows. If researchers looking to use EEG signals to evaluate and interpret reading predictability in language models and humans, please follow the chapters according to the dashed arrows. Otherwise, readers may choose to follow the thesis in linear order to gain insights from the author’s perspective.

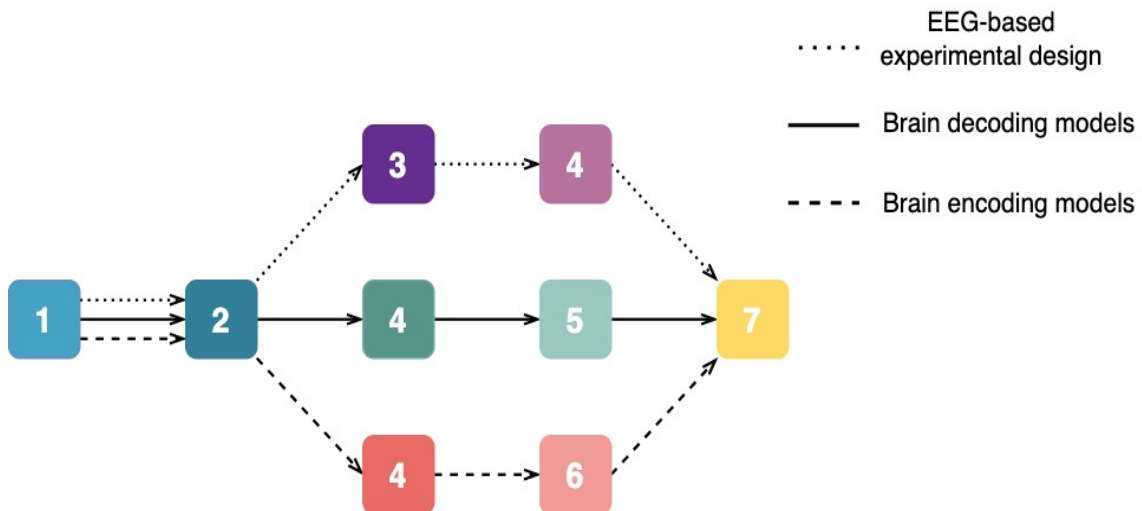


Figure 1.2: Reading threads of the thesis.

1.6 Author's Publications

This thesis is based largely on the following publications presented in chronological order:

- **Quach, Boi Mai**, Gurrin Cathal, and Graham Healy. “Sentiment analysis: using detrended fluctuation analysis of EEG signals in natural reading.” In Proceedings of the 29th Irish Conference on Artificial Intelligence and Cognitive. (AICS 2021)
- **Quach Mai, Boi** , Binh T. Nguyen, Gurrin, Cathal, and Healy, Graham. “Experimental Design and Data Preparation for EEG-based Reading Research: A Comprehensive Overview and Practical Example.” IEEE Access (2024). (Under review).
- **Quach Mai, Boi** , Gurrin, Cathal and Healy, Graham. “DERCo: A Dataset for Human Behaviour in Reading Comprehension Using EEG.” Nature Sci. data (2024).
- **Quach Mai, Boi** , Gurrin, Cathal and Healy, Graham. “Decoding EEG Signals to Explore Next-Word Predictability in the Human Brain.” (Accept to CogSci 2025)
- **Quach Mai, Boi** , Gurrin, Cathal and Healy, Graham. “Encoding EEG Signals to Examine Human-Like Next-Word Prediction Behaviour in Language Models”. The 63rd Annual Meeting of the Association for Computational Linguistics. (Under review).

Further published papers, which are outside the scope of this thesis:

- Gómez, Carlos, Niamh Belton, **Boi Quach**, Jack Nicholls, and Devanshu Anand. “A simplistic machine learning approach to contact tracing.” arXiv preprint arXiv:2012.05940 (2020).

- Healy, Graham, Tu-Khiem Le, **Boi Mai Quach**, and Minh-Triet Tran. “Overview of the NTCIR-16 RCIR Task.” In Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies. NTCIR, 2022.

Chapter 2

Related Work and Background

2.1 Introduction

Readers understand written texts while subconsciously predicting upcoming words based on the context provided by preceding words. Thus, next-word prediction is an integral part of human reading processing. LMs are also trained to accurately predict the next token given previous context. Thus, next-word predictability serves as a cognitive concept for comparing how well different LMs capture human-like reading comprehension abilities. However, can we claim that these LMs can “think” like humans during reading? When attempting to train larger models with more knowledge, can we transform LMs into human-like models of reading processing? When it comes to thought, we need to examine brain activity. This is because when people engage in language comprehension, their brains display particular patterns of electrical activity [40]. Therefore, rather than examining next-word prediction performance between various LMs, we should establish a link between next-word predictability and neural responses during reading. These assumptions sidestep a deeper problem addressed we expose in this dissertation.

The structure of this chapter is as follows: Section 2.2 delves into the neural responses associated with next-word prediction, with a focus on ERPs in neurolinguistics. It introduces key concepts such as the role of ERPs in capturing neural activity and explores semantic ERP effects observed during reading comprehension.

Section 2.3 discusses LMs in the context of next-word prediction. This section reviews statistical LMs such as n-gram models, highlighting their historical importance and limitations, and contrasts them with transformer-based LMs, which have revolutionised language modelling through superior contextual understanding. Finally, Section 2.4 introduces the relationship between LMs and human predictive processing, emphasizing their alignment in linguistic prediction and exploring how brain activity can be encoded and interpreted using surprisal-based metrics. This chapter thus bridges cognitive neuroscience and AI providing a foundation for understanding predictive processes in both humans and machines.

2.2 Neural Responses to Next-Word Prediction

2.2.1 ERPs in Neurolinguistics

EEG signals are recorded from electrodes placed on a participant’s head and connected using a conductive gel. The modulations in the recorded electrical potentials are commonly referred to as brain waves or EEG waves [41]. Continuous EEG recording provides a relatively coarse measure of brain processes and hence requires careful consideration on how it is effectively utilised in experimental studies. Several techniques have emerged over the years to mitigate noise in EEG in order to enable the measurement of EEG phenomena linked with cognitive processes. These phenomena are often elicited in controlled experimental settings, where components of the EEG signal that reflect informative responses to stimuli can be measured and studied. One such type of EEG phenomenon, ERPs was first discovered by Pauline and Davis in 1935-1936 and later published by Davis [42].

ERPs in a continuous EEG recording can be identified by time-locking the EEG signal to specific events, such as the presentation time of stimuli, and averaging the time-locked EEG signals. They are often used to analyse brain responses to various stimulus (e.g., visual, auditory, or somatosensory). Although ERP components are a part of EEG signals, their amplitudes are relatively weak compared to the on-

going background activity, which results in a low signal-to-noise ratio (SNR) [43]. Several ERPs can be used to study language comprehension. Still, it is necessary to clarify and understand the characteristics of each distinct ERP component and how language processing could trigger changes in these components during reading [44].

In neurolinguistics, many researchers have long studied this principle to uncover the mechanism behind the cognitive processes while reading. Berkum et al. [45] conducted experiments to determine the extent to which readers make predictive inferences from an unfolding piece of text. In their research, readers were asked to complete a truncated story where many of them came up with the same words. The evidence from ERPs suggested that next-word predictability is a pervasive aspect of ordinary language comprehension, affecting several levels of the comprehension system involved. ERPs have shown that the language system is sensitive to mismatches between its expectations and the actual linguistic input, with ERP signals differing significantly when the incoming material is unexpected. [40, 46]. Figure 2.1 illustrates how ERPs are elicited from EEG data when studying brain activity related to a specific event, in this case, the word “reading”.

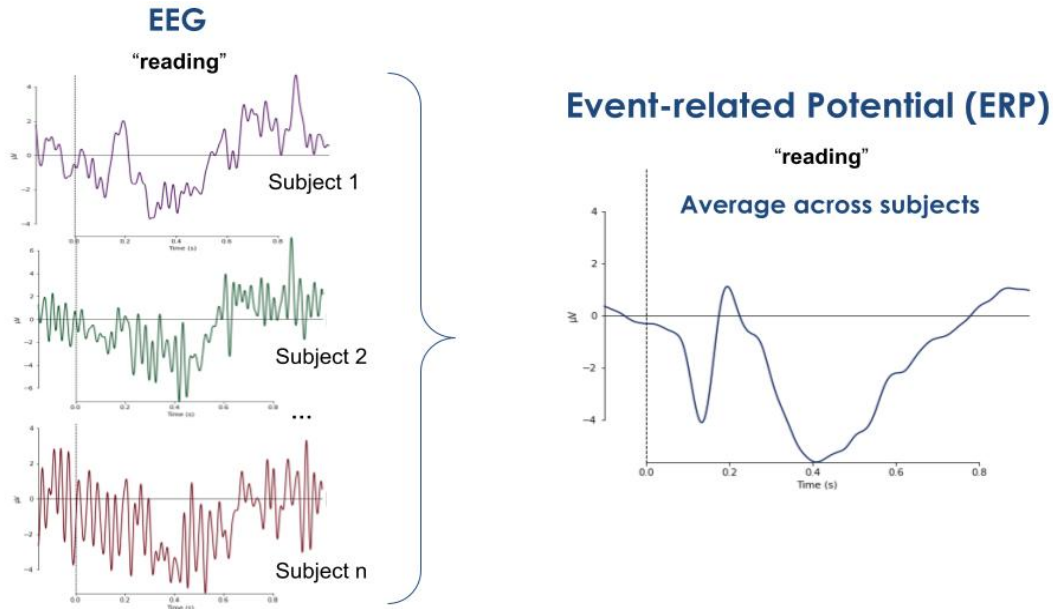


Figure 2.1: Deriving ERPs from EEG data. Left: Raw EEG signals from multiple subjects for the word “reading” (a word stimulus in our experiment). Right: ERP of the word “reading” obtained after averaging EEG signals across all subjects.

2.2.2 Semantic ERP Effects in Reading

ERPs have been extensively utilised in many language studies, with particular focus on the N400 component [47–53], a negative-going deflection typically occurring 300–500 ms after word onset. The N400 response has been well studied, and it has been established to be associated with semantic complexity [52], acting as a probe for semantic processing. As shown in Figure 2.1, a negative-going deflection peaking around 400 ms post-stimulus onset, within the 300–500 ms time window, represents the N400 ERP component.

Two commonly-used measurements of N400s are amplitude and latency. For the N400 amplitude, Petten and Kutas [48] investigated the correlation between word characteristics (frequency and position) in a sentence and the N400 amplitude. They concluded that low-frequency words were correlated with larger N400s than the high-frequency words when they came up early in a sentence. Additionally, N400 amplitudes are reduced by factors that might be expected to ease processing, such as repetition [54, 55], and context aids in the interpretation [52, 56]. Moreover, all words produce an N400 response, which reflects the process of accessing the word or concept. When the prior context allows us to predict semantic features or words, the N400 amplitude is attenuated and the word is easier to retrieve [57]. This phenomenon is well-explained by top-down prediction theories, which effectively account for the differences in N400 responses between function and content words. Moreover, different theories about the role of the N400 in both processes have shed light on how meaning processing unfolds [52].

Besides studying N400 amplitude effects, researchers have used measures of N400 latency as lexical processing could shorten or lengthen the response latency in reading. The latency is the duration time between stimulus representation and the time point of the N400 peak. However, it is inconsistent when analysing N400 effects in reading comprehension. Thus, the N400 latency is frequently researched with the N400 amplitude in most semantic studies. Compared to the N400 amplitude, most variables correlated to language comprehension that affect the N400 latency are not

linguistics factors but mainly related to the subjects and experimental conditions such as age-related variables [58, 59], language proficiency level (e.g., bilinguals [60, 61] or trilinguals [62]), neurological and psychiatric diseases [52] (e.g., Alzheimer’s disease [63–65] and schizophrenia [66, 67]), and stimulus’s presentation rate in experimental design [49]. In this research, we examine the N400 amplitude rather than its latency. To minimize the effects of latency variability, our EEG data were collected from participants who were 18 years or older, native English speakers, and without any neurological disorders.

2.3 LMs for Next-Word Prediction Task

Next-word prediction, a type of language modelling, is one of NLP’s fundamental tasks. LMs have advanced remarkably in recent years. Modelling approaches currently range from conventional n-gram models (e.g., bigram, trigram, etc.) to more recent advanced transformer neural network models [10–12, 68]. When humans read a text while subconsciously predicting the next word, they incorporate preceding information to predict upcoming words. Regarding the same principle, typical language models from these two types of LM architectures will be discussed in this section.

2.3.1 N-Gram Models

One of the straightforward methods used to estimate the probability in upcoming word prediction tasks is the n-gram model. In principle, these probabilities are calculated from relative frequency counts. The n-gram models first take a very large corpus and count the number of times that the given word is seen in the corpus. Mathematically, to compute the probability of a word w given some history h or $P(w|h)$, we need to answer the question “Out of the times we saw the history h , how many times was it followed by the word w ?” For example, take the sentence beginning “The four leaf clover was the color...”, and we want to know the probability that the

next word is “green”. Using a bigram model, we would compute $P(\text{green}|\text{color})$ and $P(\text{other single words}|\text{color})$ to determine the more probable guess between these two words. Continuing this example, one might imagine that the model’s guess would be even more accurate if we computed $P(\text{green}|\text{The four leaf clover was the color})$, making a 7-gram model. However, such a model would take enormous computing power and a much greater amount of time than the bigram model to compute. Since good estimates can be made based on smaller models, it is more practical to use bi- or trigram models. The idea that a future event (in this case, the next word) can be predicted using a relatively short history (for instance, one or two words) is called a Markov assumption. A primary limitation is that n-gram models are usually based on local contexts of only a few words and thus ignore background information and long-range dependencies. Advances in NLP have led to development of various architectures, such as Transformer-based language models, in order to address these disadvantages.

2.3.2 Transformer-Based Models

The transformer architecture was introduced in 2017 by researchers at Google Brain and uses self-attention mechanisms to focus selectively on words in the text that are most relevant in a context [69]. Since then, several transformer-based models have been developed and widely adopted for various NLP tasks. One of the most prominent of these is GPT [15], a transformer-based autoregressive language model that uses masked self-attention to predict the next token in a sequence, based on the preceding left-to-right context.

GPT-family models are trained on vast corpora crawled from the internet to predict upcoming words accurately given the preceding context. The self-attention mechanism, a primary innovation of the transformer architecture, calculates weights for every word in a sentence as they relate to every other word in the context, enabling the LM to predict words which are likely to be used in sequence. The self-attention mechanism allows each word to attend to every other word in the

sequence parallelly, weighing their importance for the current token. In this way, it can be said that LMs can “learn” the rules of grammar and capture the previous context by statistically modelling word usage patterns in language. Moreover, the flexibility of the transformer architecture allows for the training of context-specific representations across a wide range of domains [70–72].

The mechanics of how self-attention in transformer models learns to predict the next token in a sequence have been further examined by Li et al. [73]. Their analysis reveals that the self-attention mechanism acquires a two-step process for next-token prediction. First, the self-attention layer performs a “hard retrieval” operation, precisely selecting the high-priority input tokens most relevant to the last input token. It then applies a “soft composition” step, creating a convex combination of these retrieved high-priority tokens from which the final next-token prediction can be sampled. In this way, the self-attention layer learns to both identify the most salient context and then synthesise that context into a coherent next-token prediction, representing a sophisticated language modelling capability within a single self-attention module. The overview of their result is illustrated in Figure 2.2.

GPT2

GPT-2 [11] was developed by researchers at OpenAI using a larger dataset and with more parameters to produce an even stronger language model. GPT-2 comes in various versions, distinguished by the number of parameters (model size). All versions had up to 1.5 billion parameters (XL model), which was ten times more than GPT-1 (117M parameters). It is trained on several unlabelled text (e.g. Wikipedia, books, movie scripts) corpora. The model will learn to estimate the probability of any given word sequence, even unseen words. The use of GPT-2, therefore, offers opportunity to assess the extent to which prediction is feasible in natural reading, and it also provides a means of quantifying precisely the probability distribution of any word in any context, including those words with very low probability.

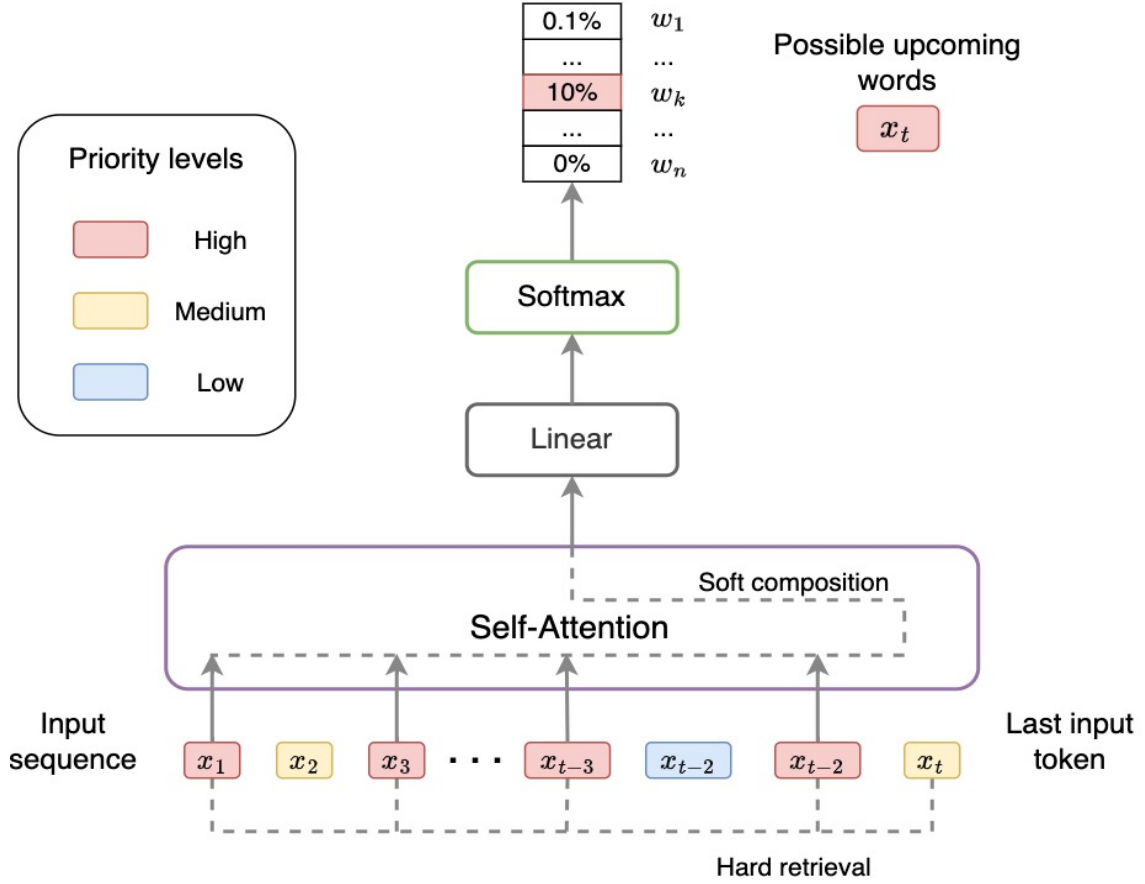


Figure 2.2: The diagram of a self-attention layer in Transformer models illustrates how it learns to predict the next token in a sequence, as adapted from Li et al. [73]. A single-layer self-attention model performs a form of hard retrieval, precisely selecting high-priority tokens and then producing a convex combination of these as the output, from which the next token can be sampled.

GPT-Neo

GPT-Neo [38], developed by researchers at EleutherAI, serves as an open-source alternative to GPT-3, replicating its transformer-based architecture while offering greater accessibility, and the capabilities of large-scale language models. Unlike proprietary models, GPT-Neo allows users to train and deploy models on local hardware, offering greater flexibility and reducing dependency on cloud-based systems. While both GPT-Neo and GPT-2 leverage the Transformer architecture, GPT-Neo adopts several innovations from GPT-3, such as pre-layer normalization, post-layer normalization, and rotary positional encodings, enabling better handling of large-scale models and longer sequences. GPT-2 remains simpler but highly effective,

especially for medium-scale applications. GPT-Neo’s flexibility and open-source nature give users more control over the architecture and training process, making it a more versatile option for researchers and developers.

In this thesis, we examine the next-word predictability of GPT-2 and GPT-Neo families. The hyper-parameters of model variants from the two GPT families are outlined in Table 2.1.

Model Name	n_{layers}	n_{head}	d_{model}	n_{params}
GPT-2 Small	12	12	768	~124M
GPT-2 Medium	24	16	1024	~355M
GPT-2 Large	36	20	1280	~774M
GPT-Neo 125M	12	12	768	~125M
GPT-Neo 1.3B	24	16	2048	~1.3B
GPT-Neo 2.7B	32	16	2560	~2.7B

Table 2.1: Hyper-parameters of GPT-2, GPT-Neo families. n_{layers} , n_{head} , d_{model} , and n_{params} respectively refer to number of layers, number of attention heads per layer, embedding size, and number of parameters.

2.4 How Might LMs Be Related to Humans via Next-Word Predictability?

In reading, both LMs and the human use the same inputs (i.e., text data), which suggests that the statistical properties of linguistic inputs may end up being represented in similar ways in both systems. Humans understand language in reading based on the interaction between top-down and bottom-up processes (as mentioned in Section 1.1). On the other hand, LMs learn language from real-world textual data (from various sources) with minimal or no explicit prior knowledge of linguistic structure. LMs do not break down sentences into parts of speech or follow explicit syntactic transformations. Instead, they learn to encode sequences of words into numerical vectors, called contextual embeddings, which are used to predict the next word. Once trained, the next-word prediction mechanism allows them to create new, contextually appropriate, and grammatically correct text [74, 75].

Although human and LMs approached language understanding in different ways, they share core computational principles rooted in next-word predictability [76]. Both systems rely on the ability to anticipate next words in a sequence, a crucial behaviour for efficient language understanding. A key link between these systems lies in the surprisal measure (as detailed in Section 2.4.1). This measure captures more than just prediction accuracy; it quantifies the cognitive effort involved in processing language, aligning closely with neural responses observed in human brains during reading comprehension. Notably, prior research has demonstrated that surprisal, derived from LMs, is correlated with human neural responses, supporting the hypothesis that similar computational principles govern both systems [76–79].

2.4.1 Linguistic Prediction at the Computational Level

Word predictability effects fit into a broader picture of human cognition, in which individuals continuously integrate new input with context to make predictions about upcoming events and test those predictions against their perceptual input from the utterances they hear and read [80]. But what cognitive processes underlie these predictability effects? One view is that predictability effects reflect the cognitive costs associated with probabilistic inference over sentence interpretations [79]. This perspective, grounded in information theory [13], frames prediction as an intrinsic function of a generative, probabilistic mental processor. This processor performs incremental probabilistic inference over a vast (potentially infinite) space of possible sentence analyses as the input unfolds. Under this framework, linguistic units convey quantifiable information, with measures such as surprisal (the unexpectedness of a word given its prior context). In general, surprisal serve as an useful metric for quantifying word-by-word predictability during incremental sentence processing [81]. With this theoretical foundation in mind, we now review this information-theoretic measure of language processing.

Surprisal

Surprisal is a measure of the unexpectedness of a target word. Hale [82] and Levy [83] argued that the less expected a word is in a given context, the higher its surprisal. Considering an example “Peter won the championship. Afterward, he was in seventh ...”. If readers know the idiom, they can guess that the missing word is “heaven.” That said, no new information is conveyed in this case. The surprisal is computed by determining the probability of a word given its context, applying a logarithmic transformation, and converting the result to its negative value. If the upcoming word is 100% predictable, the surprisal will be equal to 0. On the other hand, if we have a very unpredictable word in a sentence with a low probability of occurrence, the negative log would give a high surprisal value. The natural logarithm is used to provide nothing more than a scaling. The surprisal quantifies the amount of information conveyed by a word, for instance, this word will convey 10 bits of information, or this word will contain only 10 bits of information if we choose a log base of two. In cognition, it is reflected in the amount of cognitive processing effort required for comprehension.

2.4.2 Encoding Brain Activity Using Surprisal

Neural Responses Prediction under Cloze Estimates

Unlike language models, determining the exact probability of the next word generated in the human mind remains an unattainable goal due to the complexity and opacity of neural computations. Nevertheless, predictability in psycholinguistics is commonly studied using the cloze procedure [37], a traditional approach which involves asking participants to predict and complete unfinished sentences or passages based on the accumulated preceding context. This approach has been widely regarded as the gold standard for estimating human lexical probabilities, and cloze probability has emerged as the primary metric for contextual word predictability [47, 84, 85].

The cloze procedure offers several advantages. First, it indirectly reflects human subjective probabilities, capturing how individuals perceive the likelihood of specific linguistic outcomes. Second, cloze-based estimates have proven superior to corpus-derived probabilities in predicting human reading patterns [86]. Empirical studies consistently demonstrate that words with higher cloze probabilities elicit smaller N400 responses compared to words with lower cloze probabilities [47, 52, 57, 85]. Furthermore, research has identified a strong linear correlation between word cloze probability and lexical processing difficulty [85, 87].

Research has consistently demonstrated that surprisal serves as a reliable predictor of neural responses during reading, particularly in relation to the N400 component. Michaelov and Bergen [88] found that when measuring the predictability of words using surprisal, this metric effectively predicts variations in N400 amplitude, which serves as a neural indicator of processing difficulty during language comprehension. Frank et al. [89] further supported these findings by analysing EEG data from participants reading identical sentences and examined four distinct ERP components. Their results specifically highlighted that surprisal estimates significantly predict the amplitude of the N400 only, with more surprising words eliciting larger negative N400 responses in EEG recordings. Lindborg et al. [90] provided additional evidence by indicating that semantic surprisal effects are specifically confined to the N400 time window, and its effect topography aligns closely with conventional ERP analyses of expected versus unexpected words.

Neural Responses Prediction under Language Model’s Probability

Effects of surprisal, which can be generated on a word-by-word basis, can be derived from various language models, including n-gram models, simple Recurrent Networks [91], Recurrent Neural Network (RNN) Encoder-Decoder framework [92], and Long Short-Term Memory (LSTM) networks [93]. Studies by Smith and Levy [87], Goodkind and Bicknell [94] have thoroughly evaluated these models with respect to behavioural indicators of processing difficulty. More recently, Transformer-

based language models have gained prominence, dominating numerous NLP tasks, which highlights the growing interest in evaluating their capabilities, particularly in terms of next-word predictability [95–98].

Surprisal modelling from LMs has been commonly applied to predict neural responses during language comprehension. Surprisal, estimated by simple, efficient, and widely applied trigram models (also known as second-order Markov models), has been successfully used to show its positive correlation with the N400 effect observed in reading studies [99–101]. For more advanced LMs, Heilbron et al. [102] demonstrated that GPT-2’s word-by-word un(predictability) estimates align with ERPs in EEG recordings when participants listened to audiobooks. Their findings revealed a central negative response at 400 ms, similar to the well-known N400 component. As further evidence, when compared with non-transformer-based LMs, Michaelov et al. [103] found that GPT-3 surprisal provided the best fit for N400 amplitude, suggesting that effects of expectancy, plausibility, and contextual semantic similarity could all be accounted for by variations in word predictability.

Within transformer-based LMs, Hao et al. [96] showed GPT-2 as the most effective at predicting psycholinguistic patterns, outperforming both TransformerXL [104] and XLNet [105], in terms of psycholinguistic predictive power. These results indicated that GPT-2 may better capture lexical surprisal that aligns with human predictability, potentially making it a strong candidate for psycholinguistic modelling in comparison to more complex architectures. However, when examining the GPT family, Shain et al. [79] challenged claims that more advanced LMs should exhibit stronger logarithmic relationships between contextual predictability and processing difficulty. They found that surprisal estimates from GPT-3 were not more “super-logarithmic” than those from smaller models such as GPT-2, even though GPT-3 is a larger and more powerful model.

2.5 Conclusion

In this chapter, we have established the theoretical and empirical foundations necessary to understand the relationship between human language processing and LMs. Thanks to the review of neural responses, particularly ERPs, we have gained insight into the predictive mechanisms employed by the human brain during language comprehension, revealing how contextual information shapes anticipatory processes. Similarly, the exploration of language models—from traditional n-gram models to advanced transformer-based architectures—illustrated how computational systems emulate these predictive capabilities through probabilistic modelling of sequential linguistic data.

Furthermore, the discussion highlighted next-word prediction as a critical link between cognitive neuroscience and AI. By analysing how LMs align with human predictive processing and how neural responses can be analysed in terms of surprise, we identified significant overlaps and divergences between these domains. This synthesis not only enhances our understanding of linguistic prediction but also opens avenues for evaluating whether LMs can approximate human-like cognition or “thinking” in the context of reading and language comprehension.

Finally, this background and literature review chapter establish a comprehensive foundation for the dissertation’s core objective: examining how language models can mimic the predictive neural processes of human cognition. This foundational knowledge will guide subsequent empirical and theoretical investigations, connecting computational modelling with cognitive neuroscience.

In order to investigate next-word predictability in LMs and human productions, a dataset specifically designed for this purpose is necessary. Therefore, the next chapter will fully describe how to collect EEG-based reading experiment data (see Chapter 3) and introduce the DERCo dataset (as outlined in Chapter 4).

Chapter 3

Experimental Design and Data

Preparation for EEG-Based Reading

Research: A Comprehensive

Overview and Practical Example

The purpose of this dissertation is to utilise EEG signals to examine human-like next-word predictability in different LMs during reading comprehension. To address this, the first and most important step of this thesis is to design an EEG-based reading experiment and propose a preprocessing pipeline particularly suited to our dataset. However, EEG-integrated reading experiments require a common framework for researchers as current approaches are numerous and often disorienting. Therefore, this chapter aims to address the first research question *“How do we design a standard framework for EEG-based reading experiments?”* Additionally, this chapter does not only serve for the whole research but also provides comprehensive guidelines for designing EEG-based reading experiments, providing valuable materials for researchers in this field of neuroscience. Moreover, we will leverage our DERCo dataset collection project as a practical case study to validate the effectiveness of our proposed EEG preprocessing pipeline in this chapter. Finally, we

ultimately hope that this work will encourage other researchers to run EEG-based reading experiments and achieve significant outcomes.

3.1 Introduction

Existing studies in the cognitive neuroscience of language have sought to explore the connection between the changes in our brain and manipulations of linguistic representations [28–30]. Three measurement techniques are frequently employed in brain connectivity studies: functional Magnetic Resonance Imaging (fMRI), Magnetoencephalography (MEG), and EEG. Compared to fMRI, EEG provides measurement of brain activity at higher temporal resolutions, in the milliseconds range, but with a poorer spatial resolution. In addition, EEG is often preferred over MEG and fMRI due to its portability and lower cost. As a result, EEG is widely used in cognitive research, particularly in a broad range of reading experiments.

Publicly available datasets focused on EEG-based reading tasks are limited [106]. Thus, researchers often have to design and collect data specific to their purposes. Nonetheless, designing and conducting reading experiments to measure neural responses can be a painstaking process. In particular, beginners often struggle with how to start their research, for example, in selecting the appropriate type of EEG-based reading experiment. Moreover, most literature review papers focus on data preprocessing and analytic methods but do not adequately cover experimental designs.

Even after successfully collecting data, researchers frequently face many challenges during data processing, as brain signals are often contaminated and hard to analyse. Many researchers apply existing pipelines to their data, which can lead to misleading interpretations. This often goes unrecognized due to insufficient foundational knowledge of EEG signal processing. For neuroscientists, although they are typically familiar with using EEG data, there remains uncertainty about whether their approaches are suitable. Moreover, there is a trade-off to consider when choosing different approaches for preprocessing EEG data. Traditional methods can be

time-consuming, whereas state-of-the-art techniques often introduce additional complexity to the preprocessing pipeline.

The reasons outlined above open an opportunity and motivate the pursuit of this work. The main contributions of this chapter are as follows:

- We summarise common reading experimental designs that use EEG, as well as analyse the advantages and disadvantages of these setups. From this, we provide a list of key points to support researchers in choosing a proper type of experiment for their studies.
- We outline in-depth techniques for EEG signal preprocessing, both in general and with a focus on reading experiments.
- After covering various techniques and experiment design types, we propose a practical EEG-based reading experiment design and a corresponding EEG preprocessing pipeline. The final results will be evaluated in Chapter 4, where we introduce the DERCo dataset.

The rest of this chapter is structured as follows. First, we describe and analyse reading experiments that capture brain activity. This is followed by a critical review of the methodological approaches used for preprocessing EEG signals. Next, we present a practical implementation of an EEG-based reading data collection experiment. Finally, we conclude the key contributions of this chapter and discuss future directions of reading research.

3.2 Experimental Design

In reading, informative brain activity that reflects perceptual and cognitive processes occurs when a subject interacts with a given stimulus [107]. Therefore, experimental setups are primarily focused on how the stimulus is displayed. To simulate natural reading scenarios, many studies have used a co-registration approach of eye movement and EEG signals, taking advantages of eye tracking to determine where a target

word in a sentence is presented when it falls into the participant’s fovea. Common paradigms for this design include gaze-contingent paradigms [108] and free-viewing paradigms [109–111]. Nevertheless, this approach can lead to an overlap between time-locked neural responses, as these responses last longer than the individual fixations and saccades typically seen in reading [112]. The experimental design must also address the complex calibration between the eye-tracking (ET) and EEG systems. Moreover, this method often requires a high-end eye tracker, as reading experiments demand higher measurement accuracy [113] compared to other visual experiments such as those involving image or video tracking.

Limited research questions related to brain activity in reading can be explored without using an eye-tracker, where the Rapid Serial Visual Presentation (RSVP) paradigm is one of the most common approaches. A typical RSVP experiment presents one word at a time on a screen in sequence, using a fixed presentation time for each word, where a subject is required to keep their gaze consistently fixated during reading. This is fundamentally different from how natural reading occurs in that previous and upcoming words can be perceived in parafoveal vision [114, 115]. In this chapter, we will focus on RSVP-based paradigms, as this method can be implemented in many labs, regardless of whether or not they have access to eye-tracking devices.

3.2.1 RSVP-Based Paradigms

The traditional RSVP paradigm requires participants to keep their eyes at the centre of the screen while words are sequentially presented in the foveal region at a constant rate [116] (as illustrated in Figure 3.1). Using RSVP can speed up reading times by increasing the frequency of displayed words via the Presentation Speed parameter. This is because RSVP removes the need for eye movements and each incurring a saccadic duration due to the required time to jump from word to word [117]. This presents the question of whether a faster Presentation Speed reduces the reading comprehension of a reader. To investigate this hypothesis, Kosch et al.

[118] conducted experiments with three Presentation Speeds using EEG to determine individual information processing levels in terms of changes in RSVP display parameters. Their conclusions agreed with previous findings regarding the presentation rate. At 200 words per minute, participants read slower than their regular speed, which resulted in an increase in reading comprehension scores that were reflected in the EEG correlations of higher working memory load and engagement during reading. At the same time, there were no significant improvements between the Presentation Speed at 350 WPM and 500 WPM and normal reading for text comprehension and cortical activity. Thus, words are usually presented at a stimulus onset asynchrony (SOA) in the range of around 400 ms up to 1000 ms depending on the purpose of the research [119] in the RSVP paradigm.

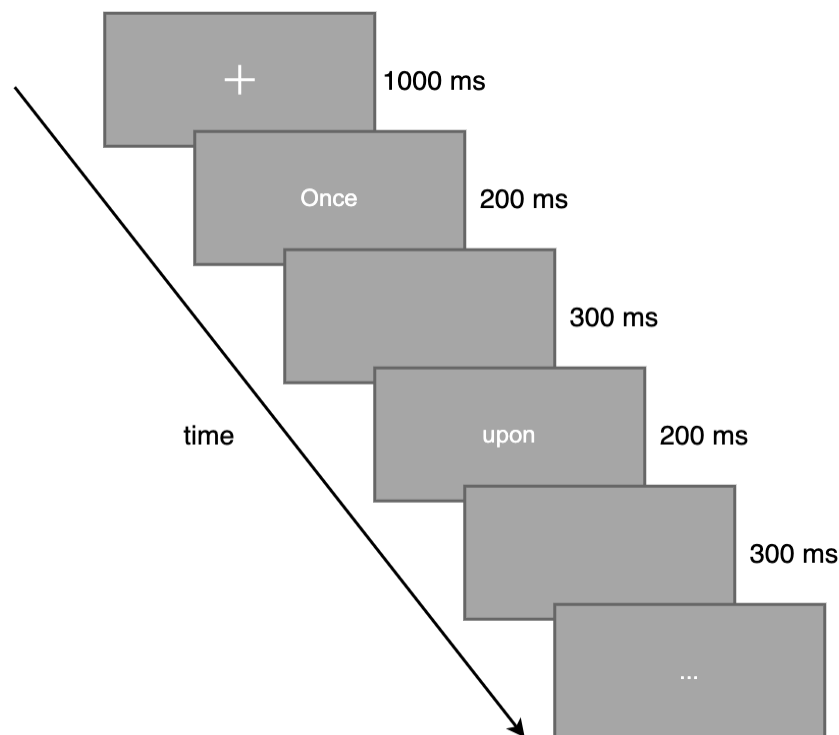


Figure 3.1: Illustration of the word presentation for the first two words, “Once” and “upon”, in the common phrase “Once upon a time” in a typical RSVP experiment. The illustrated experiment begins with a blank screen and a fixation cross in the centre of the screen for 1000 ms. Each stimulus word is then presented for 200 ms, followed by a 300 ms blank screen.

According to Rayner et al. [115], there is a considerable decline in information processing due to the reduction in visual acuity from the foveal region to the

parafovea and then the periphery. Although traditional RSVP experiments ignore many properties of natural reading, especially in parafoveal processing, the RSVP-with-flankers experiment can overcome these issues. In this paradigm, a sentence is displayed in word triplets with consecutive words appearing on the centre of the screen. The three words ($n - 1$), (n), and ($n + 1$) are, respectively, the preceding word, the target word, and the next word. The word at fixation is always visible while the words on neighbouring sides are masked by letter strings, commonly referred to as flankers. In most studies using this technique [120–123], the flanker may be valid previews (either sentence-congruent, sentence-incongruent words) or invalid previews (pseudowords) depending on the objective of the research.

Figure 3.2 illustrates an RSVP-with-flankers paradigm using valid and invalid previews. In the invalid condition, the letters of upcoming pseudowords were either replaced with a string of “x” characters or partially revealed by showing the first one, two, or three letters (right panel). In contrast, our study aimed to simulate natural reading conditions; therefore, we used a valid preview (left panel) instead of an invalid one, which typically consists of a masked, random-letter non-word. Similarly, Figure 3.3 depicts the use of congruent and incongruent word flankers in the RSVP-with-flankers experiment. However, since natural reading usually involves congruent words, we retained the original words in the transcript and did not modify the next-word flankers to be incongruent.

Many of the studies using RSVP-with-flankers also incorporate ERP analysis. Zhang et al. [122] conducted the RSVP-with-flankers paradigm to examine the semantic processing of parafoveal words during the reading of sentence lists. The result concluded that the time-locked N400 was more negative for the incongruent flankers compared to the congruent ones. That also suggests one can extract the semantic integration of parafoveal words based on the N400 component during the reading process. In terms of the sentence level, Stites et al. [124] reported the level of N400 effects for an incongruent target word would reduce or vanish after that word left the preview position toward the foveal vision. Payne et al. [125] then

replicated the work of Stites et al. [124] to explore aspects of higher-order language comprehension during natural reading, which refers to the more complex, advanced cognitive processes involved in understanding meaning beyond basic word decoding. They found that semantic access was initiated in the parafoveal vision; however, the task-dependent, integrative processing could not be evoked with its normal window time for the target word unless it occurred in the fovea. Furthermore, Li et al. [123] indicated that the effect of N400 on the foveal word only appeared when that word fell into the preview position and was contextually congruent with the sentence. On the contrary, there was no N400 evoked from the incongruent words.

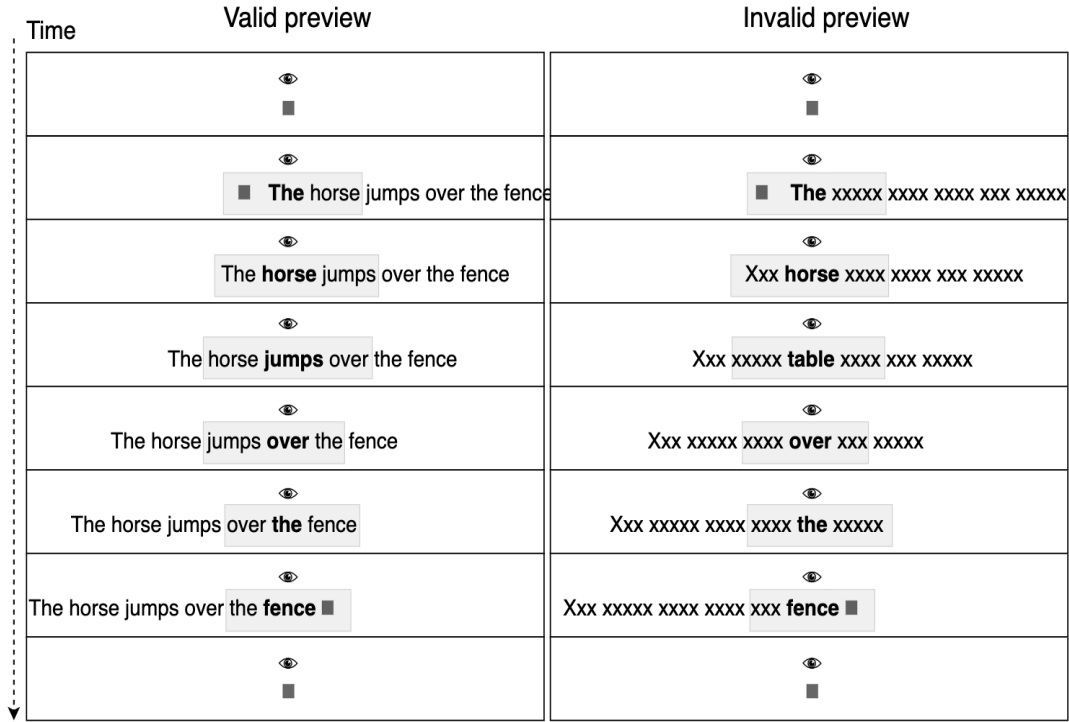


Figure 3.2: Illustration of the RSVP-with-flankers paradigm with valid (left) and invalid (right) preview.

3.2.2 Advantages and Disadvantages of RSVP Paradigms

Participants' eyes move from fixation to fixation during normal reading. Reading RSVP studies are not naturalistic because they eliminate three important properties: parafoveal processing, saccade execution, and attentional load. Thus, the RSVP-with-flankers paradigm was introduced to modify the classic RSVP paradigm, and

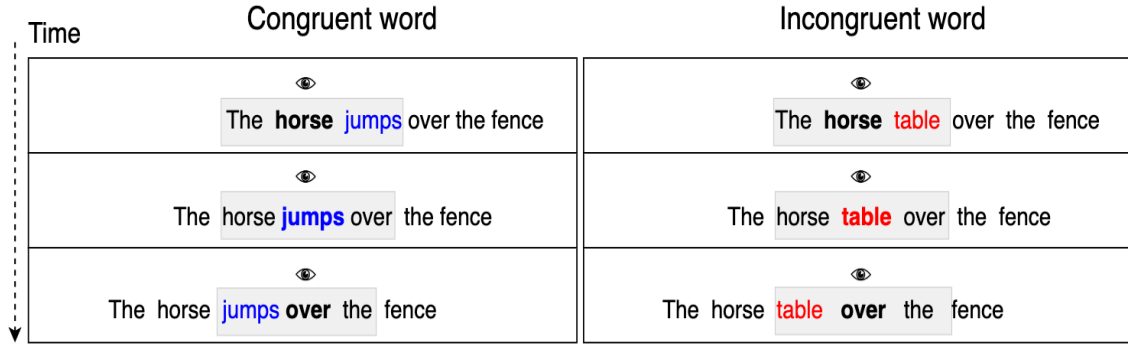


Figure 3.3: Illustration of the RSVP-with-flankers paradigm with parafoveal congruent (left) and incongruent (right) words.

brings a range of benefits. First, this approach simulates a closer approximation of natural reading scenarios than the conventional RSVP paradigm since parafoveal previews of the upcoming words are shown. Indeed, Li et al. [123] examined an N400 congruency effect in an RSVP-with-flankers experiment. The finding indicated that participants could quickly integrate the information about soon-to-be-seen words using congruent parafoveal words. This insight is essential since visual processing already happens in the parafovea and periphery during regular reading, alongside foveal vision. Where there are constraints on the presence of eye movements, the RSVP-with-flankers paradigm is a useful approach to eliminate ocular artifacts, which are of notable concern in ERP studies. This paradigm allows control over presentation speed and space to create an equal quality for the reader’s perception of all previews as much as possible. These controls could explain the more consistent semantic Parafoveal-on-foveal-like effects observed in the RSVP-with-flankers paradigm than the gaze-contingent boundary paradigm [121, 126].

Critically, the RSVP-with-flankers paradigm enables a continuous measure of brain electrical activity evoked in several stages of word processing to the parafoveal and foveal information. Thus, it sidesteps some of the processing complexities encountered when dealing with the temporal overlap between the FRPs elicited by consecutive fixations when using co-registration paradigms. The first reason could be that the RSVP-with-flankers paradigm applies a fixed time and rate of visual inputs. Another reason is that a significant overlap between FRPs will occur due to

the variation in fixation duration of those from subsequent and preceding saccade targets [111] that cannot be present in the RSVP-with-flankers implementation.

Despite its benefits, the RSVP-with-flankers paradigm also suffers from several drawbacks. The first one is in the limitation of experiment settings where participants are firmly instructed to maintain fixation on the centre of the screen. As a result of this, participants cannot shift their gaze, e.g., to a parafoveal region. Additionally, word lists in the RSVP-with-flankers paradigm navigate leftwards. A word begins on the right side of the screen and then moves step-by-step to the left. However, in normal reading, the eye-movement reading habit is opposite to this operation where a leftward bias in readers' attention was found since people had predicted the presentation's direction and tried to follow the stimulus. In the RSVP-with-flankers implementations, forcing participants to hold fixation is contrary to natural reading [127]. This is because people tend to skip the words that are easily predictable [128] or of a high frequency [129], making a longer saccade or regression in a sentence. Another disadvantage is the reduction of lexical-parafoveal processing. The reason could be that the word lists presentation limits the semantic context, such as displaying an array of five words instead of a complete sentence [130].

3.2.3 Important Features of EEG-Based Experimental Design

Poor experimental design can result in inadequate data and inconclusive or misleading findings that fail to address the hypothesis under investigation. Therefore, it is necessary to carefully conduct and consider each stage in an experiment's design, from the experimental questions and participant recruitment to the interpretation of the final results. A scientific study starts with research questions/hypotheses and ends with a conclusion or an evaluation.

EEG-based reading experiments often consist of a number of blocks, and each block includes several trials per condition under examination. According to Light

[131], a typical EEG experiment generally has 16 to 20 participants and typically a large number of trials per condition. The number of subjects is decided based on the signal-to-noise characteristics of the data, the type of methodology applied, how many trials will be collected from each participant, and the expected effect size. For instance, an ERP experiment should typically have a minimum of 50 trials on a per-participant-condition combination to achieve a sufficient SNR, but this is based on empirical research, not a strict principle [132]. Also, according to [132], the number of EEG electrodes used will depend on the selected approach and the inferences that experimenters intend to make.

Furthermore, experimental design can directly affect the final outcome. A well-designed EEG experiment has five characteristics [133]. First, an experiment should be easily reproduced in the future; thus, it needs to be as simple as possible. Second, when testing a specific hypothesis, the experiment must take factor effects and related risks into account. Third, experimenters need to minimise the running cost of experiments that will achieve significant results. Fourth, data analysis and evaluation should be planned in advance. Finally, the conclusion should relate to the aims of the work and have broad validity. Regarding developing an EEG laboratory, and instructions for subject preparation, readers can refer to Light et al. [131] and Ledwidge et al. [134].

3.2.4 Synchronisation of Data Streams

The RSVP-based paradigm can help achieve high precision time synchronisation between EEG devices and stimulus display devices due to the fact that the timing of stimulus presentation is in some way controlled, however, system delays still occur between devices. More importantly, in reading experiments, we often analyse word-level synchronised EEG segments, which typically last around 200-250 ms. Thus, any delay during the experiment will directly affect the synchronisation, leading to inaccuracies in the analysis. Therefore, before running the experiment, we should design a synchronisation between the EEG device and the stimulus display device to avoid

any delays during the experiment. Using an eye tracker in gaze-contingent boundary paradigm or co-registration of eye tracking and EEG paradigms is considered an effective approach to synchronising data streams. Nevertheless, for paradigms that do not use an eye tracker e.g., the RSVP-based experiments, we can still achieve high synchronisation accuracy. The most common solutions to synchronising data streams involve using shared triggers and/or a photodiode sensor.

Shared Triggers

In the Figure 3.4, a shared trigger system is implemented using a splitter cable attached to the parallel port of the stimulation computer. The parallel port is designed to send multiple bits of data simultaneously, which allows for fast data transmission. In the experimental setups, each pin on the port can represent a unique binary signal (on or off). These signals can be used to communicate specific triggers (event markers) to an EEG amplifier when the stimulus words presents.

Sharing triggers like this may seem outdated, but many EEG systems and other laboratory equipment are already equipped with parallel port interfaces, making it easy to integrate without needing additional adapters. Moreover, the parallel port provides almost instantaneous communication, with negligible delays between the signal being sent and received. This ensures precise timing in experiments where every millisecond counts. Furthermore, since the parallel port sends bits of data at once, it allows for multiple triggers or event markers to be sent simultaneously. This is useful when distinguishing between different types of stimuli in real time.

One drawback of this method is the requirement for the necessary connections (i.e., cables and ports) if they are not available in the system. Additionally, parallel port connections can be more susceptible to electrical noise than more modern digital solutions. This is particularly important in EEG setups where noise must be minimised. Another disadvantage relates to latency variances due to the operating system and software.

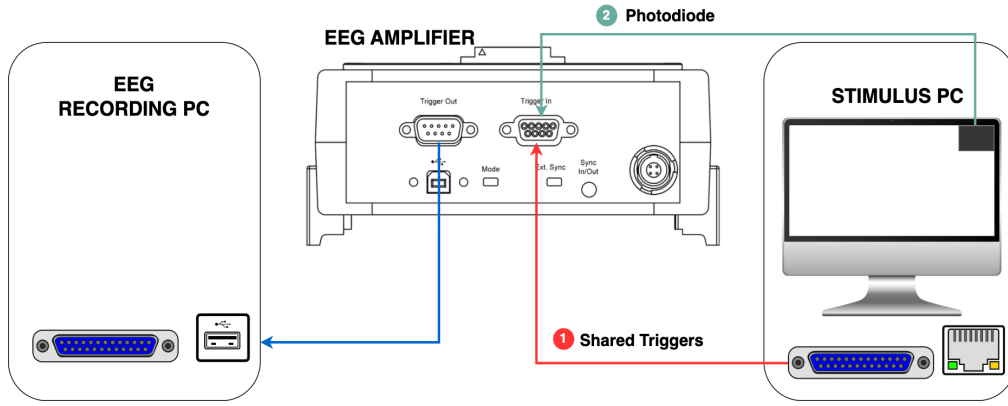


Figure 3.4: Block diagram of the 1) shared triggers and 2) photodiode sensor techniques in synchronisation of data streams.

Photodiode Sensor

A photodiode measures light and translates it to a voltage signal through a TTL (transistor-transistor logic) pulse to the EEG amplifier. It can be connected to the corner of the stimulus presentation monitor [135] as shown in Figure 3.4. Alternating word stimuli events can use a small box at that part of the screen, triggering the photodiode. This creates a marker in the EEG amplifier data corresponding to the stimulus change, allowing for the time synchronisation of all other data recorded by the platform. Different intensity of light (when the screen turns brighter or darker) can be used in the box to indicate different trials.

This method has been applied in existing EEG-based studies and has demonstrated many advantages [136–138]. Firstly, the photodiode measures the exact moment the stimulus is displayed, completely eliminating any software-related delays or variability due to hardware limitations. Secondly, since it detects the actual visual event rather than relying on computer-generated triggers, it offers higher precision in synchronising visual stimuli with EEG recordings. Lastly, a photo sensor bypasses potential display lag introduced by different monitor refresh rates or operating system delays. For example, if the monitor’s refresh rate is 60 Hz, visual stimuli can be delayed by up to 16.67 ms, which could introduce unwanted timing errors, while the photodiode captures the stimulus as soon as it is actually displayed.

The main disadvantage of this approach is the requirement for additional hard-

ware, including the photodiode sensor and extra cabling. Table 3.1 shows comparisons between these two solutions in terms of many aspects.

Aspect	Photodiode	Shared Trigger
Timing Precision	Highly accurate, as it detects the true screen stimulus.	Moderate precision, due to software and system-dependent timing.
Latency	No delay as sensor captures exact visual stimulus onset.	Small potential delays due to hardware and system issues.
Cost	Higher cost due to additional sensor hardware.	Low-cost, uses standard cables and ports.
Potential Errors	It can be influenced by ambient light.	Timing errors due to monitor refresh rates, system lag.
Usage	Requires additional hardware and setup (sensor placement).	Easier to set up, uses common hardware like parallel ports.

Table 3.1: Comparison between photodiode and shared trigger approaches

3.3 Data Preprocessing

Raw EEG data typically requires a number of preprocessing steps and transformations before meaningful measures can be extracted, e.g., ERP component amplitudes. However, when extracting such measures across various experimental conditions, researchers must deal with several problems due to the lack of standards for data preprocessing protocols.

Moreover, the collected EEG data can be different due to the dissimilar quality in EEG devices and electrode placement standards [139]. Given the importance of standardisation, this section will outline common steps that researchers frequently conduct to obtain clean data for reading experiment analysis. Firstly, we broadly discuss different types of artifacts that are present in EEG signals and present strategies to mitigate them with a range of commonly-used techniques. Following this, we introduce epoching and grand averaging, which involves extracting the time-locked segments of EEG corresponding to particular events of interest. These are also important steps in a general framework for EEG processing (as illustrated in Figure 3.5). Note that the “epochs” step can be performed either before or after artifact

removal, depending on the approach used for artifact removal.



Figure 3.5: General preprocessing pipeline for EEG data. Epochs can be implemented before (dashed line) or after artifact removal.

3.3.1 Common Artifacts

A comprehensive understanding of the different types of artifacts present in EEG recordings is essential for researchers in selecting appropriate techniques and strategies for artifact removal, as well as for mitigating and recognising potential confounds that such methods may introduce. Generally, EEG can be temporally and spatially contaminated by artifacts arising from two sources - physiological and non-physiological. The former includes internal sources of physiological artefacts such as ocular, muscular, and cardiac artifacts (depicted in Figure 3.6). Non-physiological artifacts can result from events like physical movements of the subject, or come from external environmental sources, nearby electrical devices, movement of electrodes, and cable connections [140]. In this section, we will discuss common artifacts often present in EEG and outline several techniques that can be used to mitigate their impact on the meaningful analysis of EEG.

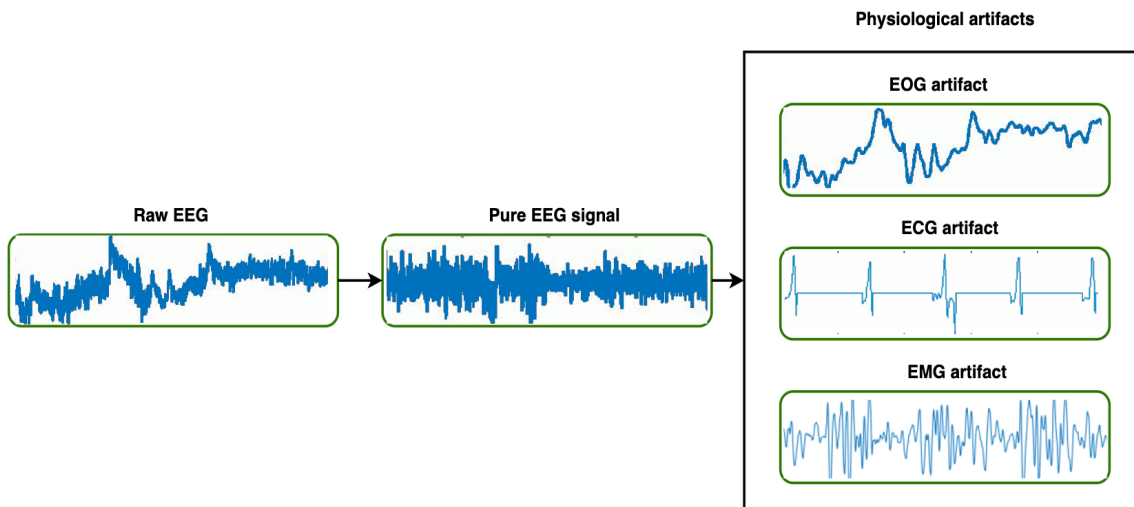


Figure 3.6: Illustration of different physiological artifacts.

Physiological artifacts

Ocular Artifacts

Eye movements and blinks are significant sources of artifacts in EEG signals, namely ocular artifacts. Eye-movement artifacts are often present in scalp-measured EEG, as there is a charge between the retina and cornea of the eyes (cornea-retinal dipole). When the eyes undergo movement, this dipole imparts spatially varying and additive electrical signals into the EEG. Further compounding the issue, the voltage can be orders of magnitude larger than the neurally generated EEG signals. In addition, eye blinks also introduce contaminants in EEG due to contact of the cornea with the eyelid during blinking [141]. Notably, the electrical scalp characteristics of this latter source of noise are spatially stable.

The signals created by eye movements and neural cognition are simultaneously detectable by EEG on the scalp. Electrooculography (EOG) can be used to measure such ocular signals with high isolation from EEG. However, EOG and EEG recordings can be contaminated by each other. A classic approach to dealing with EOG artifacts requires participants to avoid eye blinks and eye movements during experiments. However, this solution may suppress and inadvertently change a person's ongoing cognitive processes, thereby affecting the interpretation of the results [142]. Moreover, employing such a strategy is sometimes infeasible, especially in experiments performed on children, disabled people, uncooperative subjects, and more [143]. It is sometimes preferred to use artifact correction to eliminate ocular artifacts from EEG signals instead of directly instructing participants to abstain from such behaviours.

Muscle Artifacts

Muscle (electromyogram) activity near the head, such as the neck and facial muscles, during natural reading, can cause excessive noise in EEG signal recordings. These signals are known as muscular artifacts and can be found in different sets of electrodes of the scalp. These noisy signals are typically characterised by high-frequency

activity (greater than 20Hz) [144]. As a result, these artifacts can overlap with the EEG signals at beta (13-30Hz), and gamma (30-50Hz) bands in the frequency domain [144, 145].

The number of studies on EMG (Electromyography) artifacts is minimal compared with ocular artifacts. One explanation is that they are challenging to stereotype. There are fewer strategies to remove such muscular artifacts from EEG based on their different frequencies because they can stem from different muscle groups. According to a study by Goncharova et al. [144], for instance, the frequencies found in the muscle artifacts were approximately 20-30 Hz and 40-80 Hz in the frontal and temporal muscles, respectively. Furthermore, they also indicated higher component frequencies around 100 Hz for the posterior head muscles. In reading, the extraocular muscles controlling saccadic eye movements can produce activity that peaks around 65 Hz [146]. Nevertheless, there was no consistent pattern for these signals in the same frequency bands across participants, due to their gender and different types of contractions [147].

Unlike EOG artifacts, trials containing EMG artifacts are frequently removed using a screening process (trial rejection) instead of using various mathematical methods to remove their contribution to the EEG signal. In brief, the removal of muscular artifacts is particularly challenging. Whitham et al. [148, 149] used a neuromuscular blockade approach to allow the EEG signals to be collected without muscle artifacts. Along with their high-frequency properties, muscular artifact signals are short in duration. Several techniques have been proposed as suitable for dealing with muscular contamination [150]. Fatourechi et al. [145] reported that about 80% of 250 papers surveyed ignored the existence of muscle artifacts, and only 3.2% of them applied a screening process to identify EMG artifacts.

Cardiac Artifacts

Cardiovascular artifacts originate from the electrical activity of the heart and include those related to heartbeat, blood flow, pulse, and electrocardiogram records

(ECGs or EKGs). These potentials have a frequency of around 1.2 Hz [151]. The heart potentials propagate via the human body and can be difficult to disentangle from EEG signals. This leads to spike potentials, directly affecting the ERP components' amplitude. Similar to EOG signals, ECGs possess a typical pattern, also known as P-QRS-T shape [152]. Based on this characteristic, several artifact correction techniques have been applied, such as Independent Component Analysis (ICA) and continuous wavelet transformation [153, 154]. The success of the denoising approaches in detecting and removing cardiac artifacts from EEG signals can be found in several studies [155–157].

Non-physiological artifacts

Non-physiological artifacts primarily stem from technical issues encountered during experimental procedures. The most common one is power line interferences that manifest in the frequency spectrum as a series of narrow peaks at multiples of the power line frequency (50 Hz or 60 Hz), depending on the geographical region. Many approaches have been proposed to attenuate power line artifacts, including low-pass and notch filters [158]. Often, the frequency spectrum of the EEG data will be cut off below the power line frequency using a low-pass filter. The primary motivation is to enhance slowly varying features relative to fast fluctuations judged less relevant in EEG, with line artifact reduction being an added benefit. The downsides of this are (a) loss of high-frequency information, (b) limited attenuation at line frequency unless the cutoff is steep, and (c) waveform distortion, particularly if the cutoff is steep and the filter impulse response consequently long [159].

Alternatively, a notch filter, or set of notch filters, can be applied to mitigate the signal at the line frequency and its harmonics. Attenuation at other frequencies is minimal if the notches are narrow. The main disadvantage, however, is that a notch filter's very long impulse response can lead to artifacts triggered by the onset, offset, or transitions within the data [160]. Other techniques can be used such as frequency domain filters [161, 162], regression based on a reference signal [163], ICA

[164], and other spatial filtering techniques [165].

Instrument artifacts as a type of extrinsic artifact are caused by electrode misplacement and cable movements. These artifacts can be reduced by using appropriate electrode connection procedures and planning. Impedance fluctuations, as another source of artifacts, derive from the properties of the electrode-skin interface. These can be avoided by carefully attaching electrodes to the scalp and using gel to increase the conductivity between electrodes and the scalp’s surface.

3.3.2 EEG Data Cleaning

Epochs

A continuous EEG signal can be divided into EEG segments known as epochs, which are often extracted with respect to specific events for which the exact timing is known. For example, epochs can be extracted for fixation events using event markers or triggers (refer to Section 3.2.4 for examples).

Determining a suitable onset and offset (i.e., duration) for epochs concerning specific eye movement events is heavily dependent on the research question(s) under investigation in a study. For instance, if the researchers want to examine the sequential or overlapping frames of continuous EEG signals during a resting-state activity [166], it is reasonable to extract longer epochs that extend beyond the stimulus offset. However, epoch intervals are typically short in the reading studies, e.g., -100 ms to +500ms for a word stimulus.

EEG measurements in the epoch time window typically undergo baseline subtraction, where the mean value between a set of time points preceding the event to which the epoch was extracted is subtracted from all the values in the epoch. Baseline correction is commonly employed after epoching continuous EEG signals. This is done so as to minimise voltage offsets and drifts because of skin hydration, skin potentials, and static electricity charges [158].

The step of dividing continuous EEG data into epochs can be implemented either before or after data cleaning, depending on the chosen artifact removal techniques

in preprocessing pipeline. In some workflows, epochs are extracted before artifact removal, while in others, data is cleaned first and then segmented into epochs.

Detection and Removal of Bad EEG Epochs

When working with EEG data, it is common to encounter bad trials, or epochs, due to noise or artifacts. While ICA has become a widely used method for EEG preprocessing, particularly for artifact removal, it is less effective at handling sporadic, non-repeating, and non-similar bursts of noise, which can further degrade ICA performance. These bursts, mainly caused by non-physiological artifacts, including changes in the conductivity between the electrode and the scalp due to head or body movements, electrode impedance imbalance, and faulty wiring, can be removed manually [167] or automatically [168–170]. Ignoring the presence of such bad data can negatively impact downstream analysis. For instance, when multiple trials time-locked to a stimulus are averaged to estimate an evoked response, a single bad trial can corrupt the average, as the mean of a dataset is not robust against strong outliers. Therefore, bad epochs detection and removal should be conducted in the early stage of EEG pre-processing to minimise its influence.

To address this issue, several algorithms have been developed to automate the detection and removal of bad epochs, such as FASTER [168], AutoReject [170], and Riemannian Potato [169]. In the following work, we propose an EEG preprocessing pipeline where epoching is applied early in the process. After preliminary rejection of bad epochs, ICA is applied in combination with ICLabels [171] to identify and remove artifacts. Finally, the remaining clean data is grand-averaged across various conditions for analyses. In Section 3.4.4, we evaluate the performance of these algorithms when integrated with ICA and ICLabel, aiming to determine the most effective approach for our pipeline. This combination facilitates efficient data cleaning while reducing the impact of noisy trials on the final analysis.

FASTER

FASTER (Fully Automated Statistical Thresholding for EEG artifact Rejection) identifies and drops bad epochs using a global outlier detection method [168]. In this process, several statistical measures, e.g. amplitude, deviation and variance, are calculated for each epoch across the entire dataset. These measures are then transformed into z-scores, and any epoch with an absolute z-score exceeding a predefined threshold (normally 3) is considered “bad” and is subsequently removed from the dataset. This global outlier detection ensures that noisy or artifact-laden epochs, which could negatively affect subsequent analyses, are excluded. A similar rejection process can be applied for detecting and removing bad channels.

AutoReject

AutoReject [170] (AR) is a cleaning algorithm designed to automatically determine the optimal threshold for rejecting or repairing EEG epochs, without requiring the researcher to manually adjust parameters. This ease of use makes AR highly applicable compared to other cleaning methods that often require the user to set rejection thresholds. AutoReject operates in two main phases: Global AutoReject and Local AutoReject.

In the Global AutoReject phase, the algorithm uses a widely-known statistical method of cross-validation in combination with Bayesian optimization to identify the best global rejection threshold. The goal is to minimise the difference between the mean of the clean signals and the median of all signals, effectively aligning the overall median closer to the mean of the good signals.

Following this, Local AutoReject is conducted to repair or exclude individual trials. Electrodes within each trial are evaluated based on their peak-to-peak amplitude. If too many electrodes in a trial are deemed bad, the entire trial is rejected. However, if only a subset of sensors are problematic, they are repaired through spatial interpolation from the neighbouring good sensors. This two-step process ensures efficient and effective cleaning of the data, whilst preserving as much useful information as possible.

Riemannian Potato

Riemannian Potato [169] is a clustering method used to detect and reject bad epochs in multichannel MEG/EEG data. It works by first calculating the covariance matrices for all trials, which serve as the foundation for identifying outliers. A central covariance matrix is then defined, which represents the “center” of the clean data distribution. A distance threshold is established around this central matrix, forming a volume in Riemannian space. This volume is analogous to a sphere in Euclidean space, but in Riemannian space, the shape resembles a potato, hence the name. Any trial with a covariance matrix that lies outside of this “potato,” meaning its Riemannian distance exceeds the predefined threshold, is considered an outlier and rejected. This approach provides an efficient method for detecting and removing noisy or artifact-laden trials, enhancing the overall quality of EEG data [172–174].

However, bad epochs removal also has certain limitations, such as degraded performance in EEG analysis due to loss of information, unreasonable group-level statistics because of changes in the EEG’s dimensions (e.g., stimulus, channel, frequency, etc.), and increasing the potential risk of errors for large-scale EEG data when using EEG tools. In brief, instead of removing bad epochs, reconstructing signals is a potential alternative that helps overcome shortcomings. There are several useful interpolation approaches for signal reconstruction from bad channels, including spherical spline interpolation [175, 176], higher-order polynomials [177], nearest-neighbor [176] and radial basis function [178], deep learning-based methods [179].

ICA Combined With ICLabels

ICA is a widely used method for EEG preprocessing and has become a de facto standard in the field. Multi-channel EEG signals consist of a mixture of neural activity and various artifact sources. ICA is employed to separate the underlying sources from the recorded EEG channels. It enables the decomposition of EEG signals into independent components (ICs), allowing the removal of components

associated with artifacts, thereby yielding cleaner EEG signals [180]. ICA-based algorithms have been effectively used to filter out various types of artifacts in EEG data, including ocular artifacts [181, 182], EMG artifacts [183, 184], ECG artifacts [153, 185], and power line noise [186, 187].

The ICA approach can be applied to EEG data given a number of assumptions, including independent sources mixing, sources that are non-Gaussian or only one source being Gaussian, and that are a greater number of channels than the number of cortical sources of activity [141]. ICA is the primary component-based technique used in Blind Source Separation (BSS), enabling the identification and removal of artifacts. When BSS methods are applied, the axes onto which the data are projected are altered. The word “separation” comes from separating an EEG record into independent components according to the projection onto different spaces. Looking at signals from many perspectives allows researchers to obtain important information rather than a particular projection. To illustrate the algorithm, assuming that X is an electric potential signal generated by an EEG electrode $X = [x(1), x(2), \dots, x(N)]$ which is mixed by a stationary linear matrix of constants A , and S is the original artifacts source signal. The relationship between such values is shown as follows,

$$X = AS \tag{3.1}$$

To obtain the observed signals, the BSS algorithm is an extended version:

$$Y = WX \tag{3.2}$$

where Y is the assessment of sources, and the W is the reverse mixing of X .

Subsequently, the components (that contain artifactual signals) are removed, and the EEG data is reconstructed. Figure 3.7 shows the general schematic of artifact removal using BSS algorithms.

Based on the BSS’s formula 3.1, the ICA attempts to estimate a demixing matrix W . Several ICA-based techniques were developed to give a good approximation

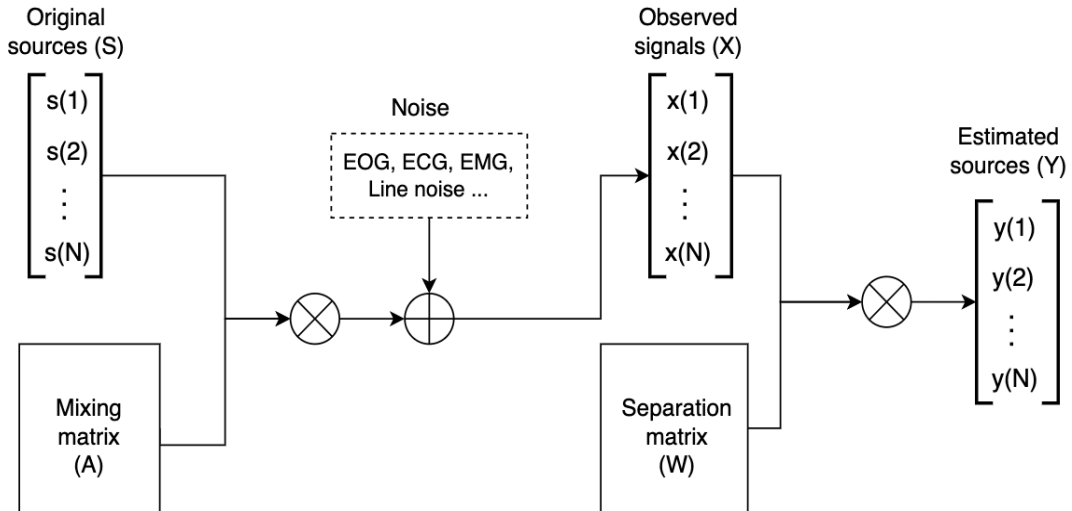


Figure 3.7: Block diagram of the blind source separation.

for W since it is difficult to determine accurately in practice. Two common estimators are used to measure these variables: maximisation of non-Gaussianity and minimization of mutual information. In the former method, the FastICA (deflation and symmetric) is representative. This approach applies an iterative technique to find the independent components by estimating the matrix W via maximising non-Gaussianity [188]. Another learning technique inspired by mutual information is InfoMax ICA. This estimation calculates independent components by maximising the entropy of a neural network with non-linear outputs [180].

One problem that prevents ICA from becoming an automated data cleaning tool is that it requires manual annotation of which ICs are related to brain activity and which are artifacts. This process requires expertise and is labour-intensive and prone to human error, especially in large datasets. To address this challenge, the ICLabel project [171] was developed to automate the classification of ICs from EEG and MEG data. ICLabel utilises a machine learning model trained on a large dataset of over 200,000 ICs from more than 6,000 EEG recordings, helping to categorise ICs into different types, such as brain-related activity, muscle artifacts, eye movements, heartbeats, and other sources of noise. The ICLabel model provides an efficient and accurate way to classify ICs, reducing the need for manual inspection and speeding up data processing in electrophysiological studies.

By integrating ICLabel into the ICA workflow, the combination offers a streamlined, automated approach to EEG data cleaning. ICA decomposes EEG signals into independent components, while ICLabel automatically classifies these components into categories, including brain signals, eye movements, muscle activity, heartbeats, power line noise, channel noise, and others. Together, they form a robust pipeline that reduces human intervention, increases accuracy, and accelerates the preprocessing of large or complex EEG datasets. This makes the method suitable for both large-scale studies and real-time applications, offering significant improvements over manual component inspection and artifact removal.

Grand Average

Grand averaging of epochs is a common technique used in EEG analysis to obtain a representative signal by averaging the data across multiple trials and subjects, increasing the SNR for time-locked neural activity and reducing the influence of random. Grand-average ERP plots are calculated by summing all single-trial EEG waveforms at each time point and dividing by the number of epochs. It is often done by separating epochs according to different conditions [158], where grand-average ERPs can be compared between conditions. The grand average is particularly useful in ERP studies, as it highlights consistent brain responses across trials and participants. Moreover, epochs can be averaged on more than one condition or across nearby electrode sites involving a specific brain area of interest (e.g., occipital grand-average ERP). Please refer to Subsection 2.2.1 and Figure 2.1 for an example of the grand average step.

After exploring the components of a typical EEG preprocessing pipeline for an EEG-based reading experiment, the next step is to apply these techniques and select the most appropriate methods for each stage of the pipeline. To achieve this, we will use our DERCo dataset collection project as a practical case study for Section 3.4.

3.4 A Practical Example of EEG-Based Reading Research

Reading is a complex cognitive process that involves various linguistic, perceptual, and predictive mechanisms. One of the critical components of reading comprehension is the ability to anticipate or predict upcoming words in a sentence. This next-word prediction ability is fundamental not only to efficient reading but also to general language processing. While much research has been conducted to understand how individuals comprehend language, the neural activities underlying predictive processing during reading remain a topic of growing interest.

ERP components such as the N400, which is sensitive to semantic incongruities, and the P300, related to expectancy and attention, have been widely used to investigate language processing during reading. These components provide valuable information about how the brain anticipates and processes expected or unexpected words in a sentence.

In this section, we will examine a practical study conducted by Quach et. al [39], which investigates the neural mechanisms of next-word prediction during reading using EEG signal. The study serves as an illustrative example for designing and implementing EEG experiments in the context of reading processing. By going through their research, we will explore the critical steps involved in setting up an EEG experiment, including participant selection, stimuli and experimental design, EEG recording, and data preparation.

3.4.1 Experimental Design

Overview

An EEG-based reading experiment was conducted to further validate behavioural next-word predictability obtained from an online study with a large cohort of 500 participants on Amazon Mechanical Turk (discussed in Subsection 4.2.2). In the online experiment, participants were asked to predict upcoming words and complete

a story using the cloze procedure (as described in Subsection 2.4.2). While the cloze procedure was suitable for the online experiment, it was not appropriate for use in the EEG experiment. For accurate time-locking in EEG analyses, it was crucial to present stimuli at a steady rate, minimising the impact of eye movements on the EEG signals i.e. using RSVP [189]. Cloze probabilities were computed for each word to quantify predictability, providing a metric used to analyse semantic context effects in the EEG recordings. In the following parts, we will outline the step-by-step implementation of a typical EEG-based reading experiment.

Participants

EEG data was recorded from 22 participants who were native English speakers, originating from the UK, Ireland, and the USA. All subjects had normal or corrected-to-normal vision and no neurological disorders, by self-reporting. The participants were employees or students at Dublin City University (18-51 years of age, mean age 23.86; 7 women, 15 men). Of these, two participants (males) were excluded from data analysis due to contaminated EEG from excessive eye movements. Every participant received a voucher gratuity of €25 upon completion of the experiment, which lasted approximately two hours. All participants provided informed consent for their participation, the re-use of the data prior to the start of the data collection, and were aware that they could withdraw participation at any point during the experiment. Data collection was conducted with approval from Dublin City University Research Ethics Committee (Reference: DCUREC/2023/085) (see Appendix C.1).

Stimuli & Experimental Design

Two types of RSVP paradigm were used for collecting the DERCo dataset for each participant: classic RSVP and RSVP-with-flanker, with a stimulus presentation rate of 200 ms. This rate is due to the fact that when a person reads a text, a word is fixated on for around 200 to 250 ms before a saccade occurs and the next word is fixated on [112, 190]. Both experimental conditions were similar, except that the

RSVP-with-flanker used triple-word stimuli, whereas the classic RSVP presented only one word at a time.

More importantly, participants in the EEG-based reading experiment were not explicitly made aware that their word predictability was being investigated. There were five short articles, each corresponding to one session. Although we allocated 25 minutes per story in the cloze experiment because participants needed time to type their predictions, each session lasted approximately 4 minutes (the shortest story) to 7 minutes (the longest story) in the EEG experiment.

In the experiment, the title was displayed at the beginning of each session. The session then began with a blank screen and a fixation cross in the centre of the screen for 1000 ms. After that, words were presented as white letters in lowercase 60-px Arial font on a gray background on a 1920 x 1080 Full HD monitor. Each stimulus word was presented for 200 ms, followed by a 300 ms blank screen. In the RSVP-with-flanker, the preceding word and the target word were revealed, while the succeeding word was hidden by a string of Xs, corresponding in length to that of the masked word. PsychoPy [191], an open-source package for creating experiments in behavioral science, was used for stimulus presentation. The RSVP experiment implementation is illustrated in Figure 3.8.

In the experiment, a photodiode was used to generate accurate EEG event markers by detecting changes in a small rectangular area in the corner of the monitor, which alternated between black and white. When a word stimulus was presented, this rectangle turned black, and when the stimulus disappeared, it changed back to white. The photodiode, connected to this corner of the display, captured these changes in light intensity, ensuring precise detection of the stimulus onset and offset.

Figure 3.9 shows the photodiode board we specifically designed to detect stimulus onsets and offsets in the EEG-based reading experiment. The blue-circled area connects to the photodiode, which detects the black-white transitions, while the red-circled section connects to the EEG device, sending event markers based on the detected changes. When the photodiode senses a transition (stimulus onset or

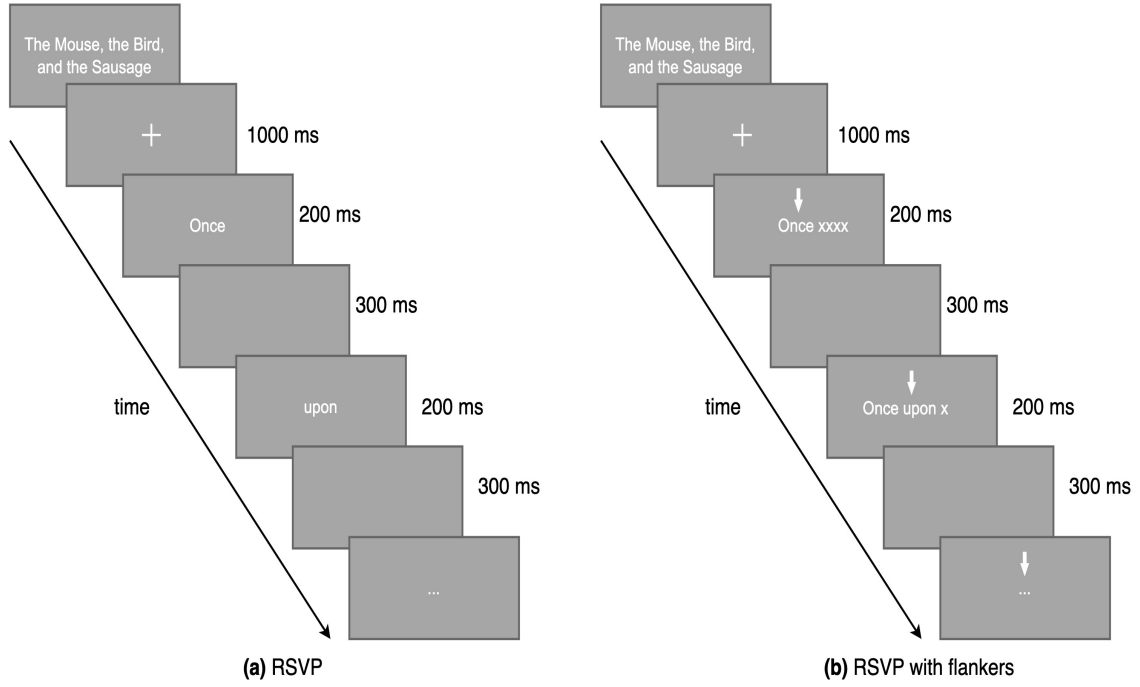


Figure 3.8: The procedure of word presentation for the first three words “Once upon a ...” in (a) RSVP experiment (b) RSVP-with-flanker experiment.

offset), the board sends a trigger to the EEG system, ensuring tight synchronisation between the visual stimulus and the recorded brain activity. This setup is crucial for studying time-locked neural responses to word stimuli during the experiment at a millisecond level.

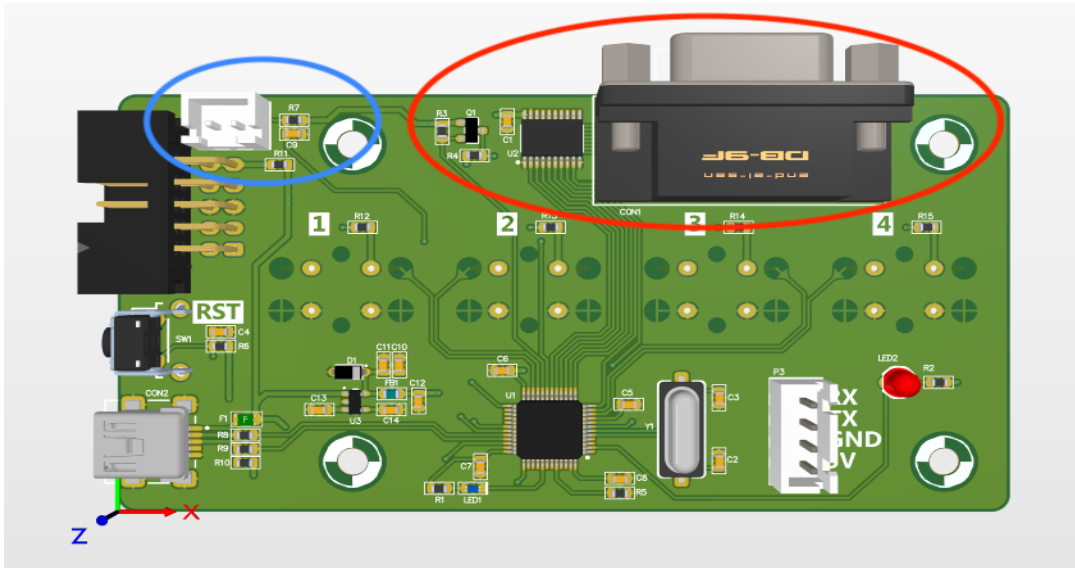


Figure 3.9: Our EEG event trigger board used in the DERCo project: The red-circled area connects to the EEG device to send triggers, while the blue-circled area connects to the photodiode, which detects light changes to mark stimulus onset.

Procedure

An EEG data collection took place at the School of Computing in Dublin City University. After participants had been given an overview of the experiment procedure, they were provided with a Plain Language Statement outlining the study's objectives, the attention-related metrics to be recorded, the usage of the collected experimental data, the extent of their involvement in the research study, and details about the associated risks, benefits, and inclusion criteria. Subsequently, they were required to sign a written informed consent form to participate in the study.

Participants were seated in a comfortable chair, approximately 100 cm in front of a 1920 x 1080 Full HD monitor in a semi-darkened sound-attenuated room. They were instructed to read (via RSVP) the five short stories in silence, where the first three stories were presented using RSVP (i.e. word-by-word display), and the last two stories were presented using RSVP-with-flankers (i.e. three consecutive words displayed at a time with the upcoming word represented with X's). Participants completed a test block before conducting the real experiment. Participants were asked to remain relaxed, and to minimise their movements during RSVP presentation, a common instruction to minimize movement artefacts in EEG recordings.

The stimuli were displayed following the description in the Stimuli & Experimental Design. After the final word disappeared, the participants answered comprehension questionnaires related to what they had just read. The primary purpose of these basic questions on the story content was to ensure the subject maintained their attention during the task i.e. the participants needed to pay attention in order to be able to answer these questions afterwards.

Sessions were self-paced i.e. participants could take a 5-10 minute rest between sessions until they were ready to proceed. After completing five sessions, the participants were required to answer certain questions about their experience during the experiment. The questionnaire for the five articles are included in Appendix C.2.

3.4.2 Data Preparation

EEG Data Acquisition

The EEG data was recorded using an ActiCHamp EEG system [192] with a 32-channel active electrode cap, with electrode positions following the international 10-20 system [193]. The signals were collected at a sampling rate of 1,000 Hz with a 500 Hz IIR (Infinite Impulse Response) low-pass filter, referenced to an internal reference (virtual ground) of the amplifier. The impedance of each electrode was carefully checked before recording to ensure good conductivity, and were checked again before each session to maintain all electrode impedances below $5\text{k}\Omega$. EEG signals were continuously recorded during each session. A photodiode connected to an external circuit was affixed to the top right hand corner of the screen and was used to generate markers (corresponding with word presentations) on the EEG recording. This allowed for correctly aligned epochs corresponding to word presentations to be extracted after the experiment. Participants were not able to see the visual marker used to trigger the photodiode i.e. it was physically occluded. Please refer Section 3.4.1 for more information, and Figure 3.9 about the trigger board.

Data for Comparison

As mentioned in Section 3.3.2, various algorithms can be applied to automate the detection and removal of bad epochs. In this section, we will compare the performance of each algorithm by examining their effectiveness in removing bad epochs.

ICA allows for the decomposition of EEG epochs into ICs. The corrected EEG signals are obtained by discarding the ICs containing artifacts. The Picard algorithm [194] was used for maximum likelihood independent component analysis. To automatically identify artifact-containing ICs, we applied the ICLabel classifier. The remaining ICs were then reconstructed to create clean EEG epochs.

There is no established evidence as to which bad epoch detection algorithm works best in conjunction with ICA and ICLabel for building a preprocessing pipeline in an EEG-based reading experiment. Therefore, this section will compare three different

preprocessing pipelines: AutoReject + ICA + ICLabel, FASTER + ICA + ICLabel, and Riemannian Potato + ICA + ICLabel.

The raw EEG data was filtered with bandpass cutoffs from 0.1 Hz to 45 Hz using a Finite Impulse Response (FIR) filter with a Hamming window. Subsequently, using the stimulus onset triggers, we segmented the raw EEG data into multiple epochs.

The DERCo dataset aims to examine the next-word predictability based on each word-level analysis. We used two non-overlapping groups of *low predictable* words (cloze $p \leq 0.06$, $n = 989$), and *high predictable* words (cloze $p > 0.52$, $n = 962$), extracted from the Mechanical Turk dataset (as outlined in Subsection 4.3.2). After filtering, the number of trials for high and low cloze probability levels varied between participants.

To solve this, we standardised the number of trials used for analysis across subjects by determining the smallest number of trials at each cloze probability level across participants and setting these as a baseline. We then randomly selected an equivalent number of trials from other participants, ensuring that the distributions of these selected trials (based on cloze values) closely matched the baseline distributions. This random selection process was guided by the Kolmogorov-Smirnov (KS) test [195], which assessed the similarity between distributions. The process iterated up to 5000 times, each time randomly selecting a subset and computing the KS statistic, with the subset yielding the lowest KS statistic considered the best match. The smallest number of trials for each pipeline is outlined in Table 3.2.

For example, for the “AutoReject + ICA + ICLabel” pipeline, the smallest number of trials for high and low cloze probability was observed in subject “SAB93” with 777 and 775 trials, respectively. Thus, we extracted the cloze values from each trial of this participant to construct distribution baselines for the high and low cloze conditions and then collected the same number of trials with similar distribution baselines from other participants.

Subject ID	AutoReject		FASTER		Riemannian Potato	
	<i>High</i>	<i>Low</i>	<i>High</i>	<i>Low</i>	<i>High</i>	<i>Low</i>
TRA37	836	844	902	929	895	932
JPY86	901	921	919	929	932	946
ACB71	888	908	907	928	892	915
LRK01	929	968	927	953	944	977
SAB93	777	775	951	973	940	958
OLW10	885	924	902	945	914	962
WHR08	902	944	891	931	888	932
NXB64	937	967	919	946	938	961
LYY64	927	959	919	940	935	960
SIT48	941	961	886	910	912	949
MRB47	873	921	888	923	828	881
WAL74	933	956	919	944	943	965
HMK96	863	879	919	927	941	962
RRO98	938	963	914	936	919	948
QFT39	931	947	915	934	940	968
QPI83	934	958	912	942	941	964
WJX11	884	940	906	947	932	960
MNY88	927	950	910	935	907	929
UJM92	858	931	878	933	912	962
DGR11	928	948	904	939	927	956

Table 3.2: Baseline number of trials for each preprocessing pipeline (AutoReject, FASTER, and Riemannian Potato) across high and low conditions for each subject. Highlighted cells indicate the baseline numbers used for comparative analysis.

3.4.3 Comparison Procedures

We can attempt to evaluate the effectiveness of our cleaning procedures. This is a challenging task, as defining the quality of a signal, particularly in EEG, is not straightforward. Different “qualities” of the same signal are required depending on different tasks. Another issue is that it is impossible to completely distinguish between the “true signal” (i.e., brain activity) and noise, as they are always intertwined during recording. While all cleaning techniques aim to mitigate noise, they may also affect the true signal. In our case, we propose two metrics for comparing different pipeline approaches: mean-based SNR and classification-based decoding accuracy.

Signal-to-Noise Ratio

One method for estimating the SNR is to assume that the true underlying signal is the one obtained by averaging multiple recordings of the same event. This approach is closely aligned with our understanding of ERPs, making it a good proxy for signal quality when assessing brain activity reliably evoked by a specific stimulus. However, it is not suitable to examine signal quality when studying more variable brain states, such as oscillations, which may cancel out during averaging.

For each electrode, we divide the mean signal amplitude over time by the standard deviation of electrode time course. The standard deviation of the error is more precisely estimated once we remove the signal. This is also equivalent to the variance of the average signal in relation to the total variance. Let us examine the situation where the response to a word stimulus w (one of W stimuli in total) is r_w . In this case, the SNR is:

$$\text{SNR} = \frac{P_W}{P_N} = \frac{\mathbb{E}[r_s^2]}{\sigma_N^2}$$

where $\mathbb{E}[\cdot]$ denotes the expectation over stimuli, i.e., $\frac{1}{W} \sum_w r_w^2$ if stimuli are equiprobable, and σ_N^2 is the noise variance, i.e., it measures the amount of variability that still occurs when the stimulus is held to a fixed value, which can be measured in practice by taking the variance of the distribution of responses across trials where that word is repeated.

Classification-Based Decoding

Since we aim to investigate the general-purpose performance of our cleaning algorithms, we will not rely only on the previously mentioned SNR but also consider decoding performance in different settings as a measure of signal improvement. We investigated N400-like components within the 300 - 500 ms time window using decoding techniques. In Chapter 5, we concluded that highly predictable words generate more distinct N400 effects across lexical classes compared to less predictable words. Therefore, we will apply the same decoding technique from Chapter 5 to

EEG data preprocessed using three different pipelines. If our cleaning algorithm has improved the signal, this should also enhance the decoding accuracy between the two conditions.

3.4.4 Results

Results of Mean-based SNR

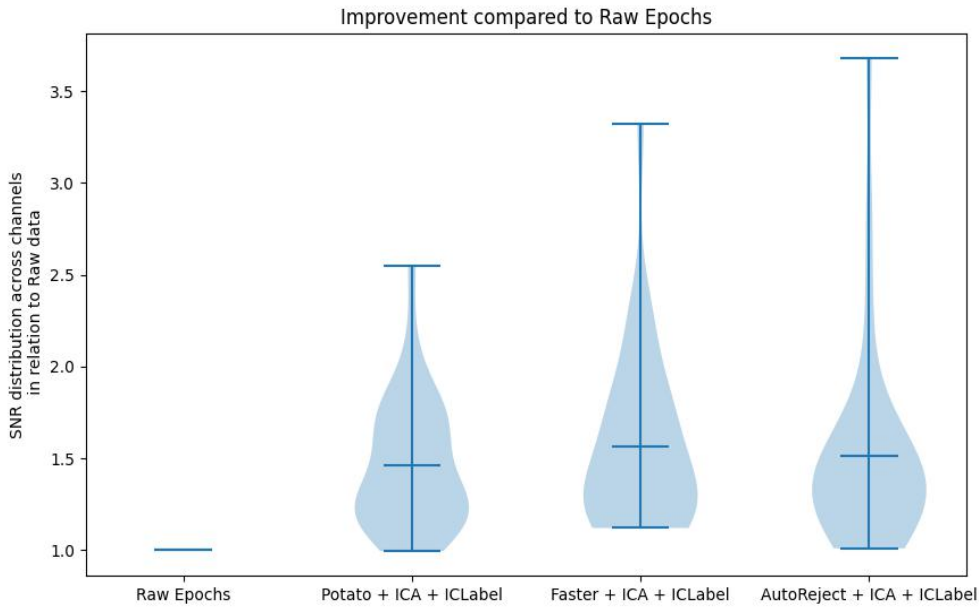


Figure 3.10: Distribution of signal to noise ratios across channels (compared to Raw Epochs).

In this analysis, we compared three different cleaning pipelines including: Potato + ICA + ICLabels, FASTER + ICA + ICLabels, and AutoReject + ICA + ICLabels, in terms of their ability to improve the SNR in EEG data. The SNR distribution across channels for each pipeline was measured relative to the raw epochs, which served as the baseline, as visualised in Figure 3.10. As expected, all algorithms improved the SNR to varying degrees, as they are specifically designed for noise reduction and signal enhancement.

Regarding the Potato + ICA + ICLabel pipeline, the result shows a moderate improvement in SNR compared to the raw epochs. Nevertheless, the variability

across channels is relatively high, as indicated by the wider spread of the distribution. This suggests that while the method improves the signal, the degree of improvement varies significantly across different channels.

For the FASTER + ICA + ICLabel pipeline, the results indicate a more consistent improvement in SNR across channels. The distribution has lower variability compared to the Potato pipeline, suggesting more consistent performance across the dataset. The overall increase in SNR compared to the raw data is notable, indicating that this cleaning pipeline is more effective at mitigating noise and preserving signal integrity.

The combination of the AutoReject algorithm with ICA and ICLabel demonstrates the most effective performance in improving SNR across channels. This pipeline achieves the highest SNR improvements relative to the raw epochs, indicating its superior ability to remove noise while preserving the true signal.

While the FASTER pipeline yields a slightly higher mean SNR than that of AutoReject, it exhibits greater variability across channels, as evidenced by the wider distribution in the violin plot. This inconsistency suggests that FASTER’s performance is more sensitive to certain types of data or noise within the dataset, with some channels achieving better SNR and others performing worse.

In contrast, AutoReject offers a more robust and stable cleaning solution across channels. This consistency is often more beneficial in practical applications than higher mean performance that fluctuates significantly between channels. The AutoReject pipeline’s reliable performance across all channels makes it a more dependable choice for noise reduction in EEG data processing.

The primary objective of this section was to identify a suitable approach for EEG-based reading experiments. While maximising SNR is important, it is crucial to evaluate the decoding accuracy achieved with each pipeline in the actual EEG reading task. These evaluation metrics ensure that our chosen method not only cleans the data effectively but also contributes to a better understanding of the cognitive processes involved in reading.

Results of Classification-Based Decoding

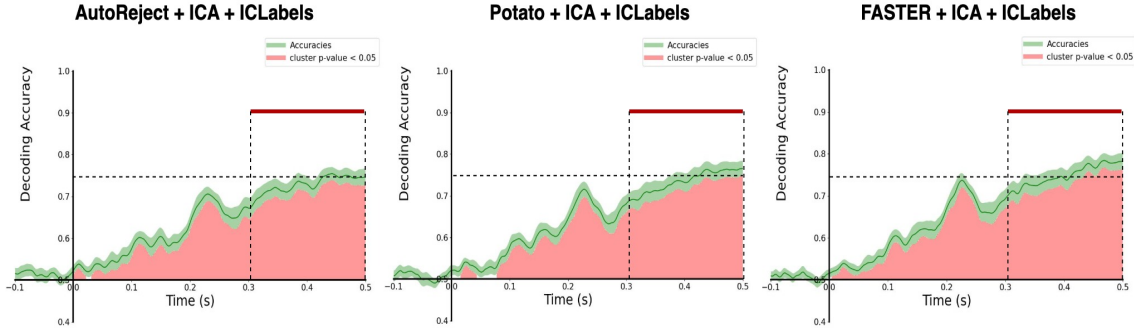


Figure 3.11: Decoding accuracies according to each pipeline.

Figure 3.11 illustrates the decoding accuracies for two conditions (high and low cloze probabilities) across different cleaning pipelines. The main focus is on the time window of 300-500 ms, where N400-like components are expected to be observed. In general, across three pipelines, the decoding technique successfully reveals significant differences between high and low cloze probability words in the N400 time window (300-500 ms) compared to other time courses based on decoding accuracy. Although the performances did not differ significantly among pipelines in the N400 time window, we can still distinguish and compare them to identify the most effective pipeline.

Table 3.3 summarises the details of the mean-based SNR and decoding accuracies, which serve as our evaluation metrics (as mentioned in Subsection 3.4.3). Nevertheless, if a pipeline obtains the best performance but is time-consuming to process, we cannot consider it the most effective one. Thus, processing time was another key factor in our analysis. While AutoReject + ICA + ICLabel achieves the highest SNR (0.113 ± 0.073), it is significantly time-consuming (≈ 7212 s), making it less practical. FASTER + ICA + ICLabel offers the highest decoding accuracy (0.747 ± 0.025) with a shorter processing time (≈ 1420 s), thus, it is the most efficient option overall. Potato + ICA + ICLabel provides a reasonable balance between accuracy and runtime, but FASTER remains the optimal choice for both performance and efficiency.

Method	Runtime	Decoding accuracy	(Mean-based) SNR
Raw Data	-	0.660 ($\pm$ 0.041)	0.079 ($\pm$ 0.052)
AutoReject + ICA + ICLabel	$\approx$ 7212 s	0.728 ($\pm$ 0.021)	0.113 ($\pm$ 0.073)
Potato + ICA + ICLabel	$\approx$ 1493 s	0.734 ($\pm$ 0.026)	0.106 ($\pm$ 0.065)
FASTER + ICA + ICLabel	$\approx$ 1420 s	0.747 ($\pm$ 0.025)	0.111 ($\pm$ 0.065)

Table 3.3: Mean (and standard deviation) of performance for the described cleaning algorithms on different task.

3.5 Discussion

In this chapter, we addressed the research question of how to design a standard framework for EEG-based reading experiments by proposing a comprehensive guideline that integrates experimental design and data preprocessing used specifically to EEG-based reading experiments. The first part outlined details on how to design such an experiment using RSVP-based paradigms, providing the advantages and disadvantages of each approach. The novelty of this section lies in its detailed explanation of each paradigm and its instructions on setting up a high-quality experiment. Most of the published literature to date on EEG-based reading experiments either ignores the experimental approaches [196–198], or only provides a brief introduction of them [199, 200]. Thus, this step is extremely important to decide the success of a study since it not only affects the quality of collected data but also reflects the researcher’s objectives.

The use case in Section 3.4 described a practical EEG-based reading research, starting from brainstorming the idea, progressing through the design and execution of the experiment. An innovative aspect of this section is to compare different cleaning pipelines to determine the most suitable one for reading experiments. EEG processing pipelines have not been standardised and are still a developing area of research. Their design depends on how researchers understand their experiments and collected signals, requiring careful implementation of cleaning steps to extract the necessary features. For this research, we focused primarily on reading datasets,

thus, the cleaning pipelines must meet evaluation metrics such as SNR and decoding performance for reading tasks.

The comparison across pipelines reveals a clear hierarchy in decoding performance. While the raw epochs show fluctuating and unreliable accuracy, all three pipelines (Potato, FASTER, and AutoReject) demonstrate improvements in decoding accuracy, particularly within the 300-500 ms time window, which reflects the N400-like component. Among the three, AutoReject + ICA + ICLabel consistently outperforms the others by achieving the highest mean-based SNR and the most stable decoding accuracy. Nonetheless, this pipeline demands substantially more processing time. FASTER + ICA + ICLabel also shows strong improvement in the mean-based SNR, with slightly more variability than AutoReject in classification-based decoding, but it has the shortest processing time compared to the others. Although this comparison gives us useful insights into cleaning performances for a reading-based dataset, it would be beneficial to investigate more use cases. Additionally, including more cleaning algorithms would be valuable, provided their implementation is feasible and efficacy.

This chapter not only addressed the first research question but also established a foundation for building a robust EEG-based reading dataset for the next-word prediction task, paving the way for success in addressing subsequent research questions of the thesis. Furthermore, the field of EEG-based reading research is far-reaching; thus, this work might not cover all of the existing literature for each and every reading experiment. However, after reading this chapter, readers can have a clear roadmap for their research process. In addition, the authors believe this research will encourage further exploration and development of other methodologies, as well as address various challenges encountered in investigating cognitive processes during reading.

Chapter 4

Dublin EEG-Based Reading Experiment Corpus

In Chapter 3, we described the process of collecting EEG data from an EEG-based reading experiment, which is part of the DERCo dataset collection project. This chapter will provide a more detailed examination of the DERCo dataset, emphasizing how EEG data correlate with reading behaviour data collected from human participants during reading comprehension. The DERCo dataset comprises behavioural data collected from 500 participants recruited through the Amazon Mechanical Turk online crowd-sourcing platform, along with EEG recordings from 22 healthy adult native English speakers. The online experiment was designed to examine context-based word prediction by a large sample of participants, while the EEG-based experiment was developed to extend the validation of behavioural next-word predictability. Online participants were instructed to predict upcoming words and complete entire stories. Cloze probabilities were then calculated for each word so that this predictability measure could be used to support various analyses pertaining to semantic context effects in the EEG recordings. EEG-based analyses revealed significant differences between high and low predictable words, demonstrating one important type of potential analysis that necessitates close integration of these two datasets. This work does not directly aim to solve any research questions in the dissertation; how-

ever, the dataset is a result of the first research question: *“How do we design a standard framework for EEG-based reading experiment?”*, mainly used to address next two research questions: (1) *“Are neural responses to upcoming words attributable to quantitative differences in lexical categories (content versus function words) and grammatical categories (e.g., nouns, verbs, adjectives, pronouns, etc.)?”* discussed in Chapter 5, and (2) *“Which type of language model has as close of a next-word predictability as a human?”* discussed in Chapter 6. Moreover, this material also provides a valuable resource for neurolinguistics researchers, due to the word-level EEG recordings in context.

4.1 Introduction

The availability of publicly accessible EEG datasets for semantic-level information in reading is limited [106]. Researchers often design and collect data with limited consideration of reuse of the data for other researchers to conduct investigations or test hypotheses, making these datasets difficult to reuse in other studies. For example, Dufau et al. [201] released an EEG dataset with a thousand words to examine the time course of orthographic, lexical, and semantic influences on word-level information. Thus, it could not be reused for studies related to decision-making or reading behaviours. In contrast, Davis et al. [202] conducted a reading-based experiment with complex decision-making (CDM) tasks to investigate how brain activity in CDM was distributed across frequencies and brain regions. Participants in this experiment were instructed to make decisions related to specific relational contexts: selecting a bus, choosing a friend, deciding on a job, selecting a medication, and opting to participate in research. Consequently, this dataset is not suitable for testing general semantic relation hypotheses in the context of reading comprehension.

Moreover, designing and conducting reading experiments wherein neural responses are captured is painstaking due to the complexity and time-consuming nature of successfully deploying EEG experiments. Although this chapter does not directly address any of the research questions in the thesis, the DERCo dataset

plays a crucial role in supporting the brain decoding and encoding models discussed in Chapter 5 and Chapter 6. The higher the quality of the dataset we collect, the better the results we can obtain. Therefore, we decided to build an EEG-based reading dataset, which significantly contributed to our understanding of cognitive mechanisms underlying reading processing.

Studying brain activity during reading is a complex challenge for neuroscientists [203–207], as it involves disentangling complex cognitive processes from continuously measured brain signals that are often contaminated with various sources of non-neural activity. Over time, EEG has become an important measurement technique not only for neuroscientific research but also for those in the field of AI, in particular for NLP. Working with EEG and AI in combination has enabled researchers to explore the similarities and differences between brain activity and AI models in language processing [76, 78, 208], mapping neural signals to natural language [207, 209], decoding linguistic information from the brain [204, 210], or encoding language into neural activity representations [211].

Two notable EEG-based reading datasets made available for other researchers are the Kilo-word ERP dataset and ZuCo dataset. Dufau et. al [201] published the Kilo-word ERP dataset in 2015, consisting of EEG data from 75 subjects performing a lexical decision task on approximately 1,000 English words. This dataset has been reused in neurolinguistic studies [212] and for teaching ERP analysis [213]. However, the paradigm used to capture this data presents word stimuli independent of each other without context, i.e. word stimuli presentations are not part of a sentence. The ZuCo 1.0 dataset [214] was first released in 2018, updated to ZuCo 2.0 [215] in 2019, and has been applied in numerous studies [207, 210, 216, 217]. This dataset comprises EEG measures co-registered to eye-tracking events during natural reading. However, the dataset’s focus is on NLP tasks such as entity recognition, relation extraction, and sentiment analysis using machine learning algorithms. Hence, the ZuCo dataset does not completely focus on single-trial responses at the word level.

To balance the disadvantages and advantages of these two datasets, our dataset

focuses on single-trial neural responses collected in a narrative context. Single-trial responses in our experiment are synonymous with word-level EEG responses. This approach is particularly important as it allows us to perform analyses such as comparing high versus low predictable words in context. While responses may be averaged within a category (e.g. high probability), word-level (single-trial) responses are still ultimately required to perform analysis on language-related ERP components [218].

Moreover, our dataset includes data from two types of experiments, including classic RSVP and RSVP-with-flanker paradigms (please refer to Section 3.2.1 for more details). Horizontal EOG was included as a part of the dataset, motivating researchers to examine horizontal eye movements should they wish to further investigate this. From our investigation of the data, there was minimal eye movements, but users of the data may decide to impose stricter thresholds regarding minor eye movements (e.g. microsaccades). Also, these paradigms use a fixed presentation rate for each word, effectively driving the speed at which participants read [118, 119, 121].

An important feature of this dataset is human annotation data via a behavioural word-prediction experiment conducted using Amazon Mechanical Turk [219] for 500 participants. The word “behaviour” was intended to indicate that participants were required to predict upcoming words [220]. Behavioural experiments are designed to test a belief or prediction and to discover new insights [221]. They are frequently conducted on crowdsourcing platforms for tasks related to psychology, neuroscience, and cognition [222]. In our research, this experiment captures the predictions for next words from 500 participants, allowing us to measure the predictability of words in the context. Combining this dataset with the EEG-based reading data for 22 participants, we demonstrate ERP differences using high-and-low cloze probabilities for next-word predictions. This work aims to provide evidence that the word-level predictability metrics derived from our behavioural word-prediction experiment are sufficiently robust and relevant to be effectively incorporated into the EEG-based

reading experiment data.

We anticipate that this dataset will facilitate AI-related research, including the development of brain encoding and decoding models, research in cognitive science and linguistics, and advancing the capabilities of natural language processing systems. For instance, it will enable researchers to study the nature of brain signals underlying semantic-level information in reading contexts. Additionally, by standardising EEG data formats and providing thorough metadata annotation, this dataset will support data reuse with different modalities such as fMRI, MEG, eye tracking data, etc. The technical validation of this dataset, described further below, demonstrates its reliability and ensures it meets the necessary standards for examining human next-word prediction ability through neural responses during reading comprehension.

4.2 Method

In our experiments, we used the same materials for both the EEG-based reading (as described in Subsection 3.4) and the behavioural word-prediction experiment. The script of the five articles are reported in Appendix C.3. The study was approved by the Dublin City University Research Ethics Committee (Reference: DCUREC/2023/085) for both experiments. Participants' responses were de-identified i.e. each individual data was represented by a unique random ID with no recording of name. The ethic approval was mentioned in Subsection 3.4.1 and Appendix C.1. They also provided permission for their data to be shared as a part of informed consent process. The two experiments must use the same materials because their datasets should be used in conjunction.

4.2.1 Materials

The primary purpose of this dataset was to enable researchers to investigate reading behaviours and comprehension, necessitating the selection of materials that avoided

using advanced/unfamiliar words and did not require specific domain knowledge, such as that needed for understanding a scientific article. Many previous studies of brain activity during reading experiments have utilised naturalistic stories for the collection of neural data, including the Alice datasets [223], Le Petit Prince (The Little Prince) corpus [224], the Harry Potter datasets [225, 226], and the Adventures of Sherlock Holmes datasets [78, 227].

Based on these considerations, five short stories from Grimms’ Fairy Tales were selected, namely: “The Mouse, the Bird, and the Sausage” (Article 1), “Straw, Coal, and Bean” (Article 2), “Poverty and Humility Lead to Heaven” (Article 3), “The Death of the Little Hen” (Article 4), and “The Wolf and the Fox” (Article 5). These fairy tales include a large, rich set of verbal descriptions without the use of advanced vocabulary, making them comprehensible to individuals with varying degrees of knowledge.

Additionally, these stories are not commonly known, thereby mitigating the impact of familiarity. Table 4.1 presents the overall descriptive statistics of the materials split by story. It shows the variability in article length, sentence structure, and word usage, with *distinct word counts* ranging from 224 to 301, and *sentence counts* from 23 to 36. The table also includes mean and standard deviation values for *words per sentence* and *word length*, indicating the average and variability within these metrics across the articles. When creating the descriptive table, we removed currency symbols, numerals, and punctuation.

		Story 1	Story 2	Story 3	Story 4	Story 5
No. of words		564	466	493	660	776
No. of distinct words		243	235	226	224	301
No. of sentences		33	32	23	36	34
Word per sentences	M	17.09	14.56	21.43	18.33	22.82
	SD	8.43	7.16	7.89	13.61	11.41
Word length	M	4.17	4.09	3.89	3.83	3.97
	SD	1.94	2.01	1.92	1.65	1.91

Table 4.1: Descriptive statistics of reading materials (M = mean, SD = standard deviation).

4.2.2 Behavioural Word-Prediction Experiment

Participants

A total of 500 participants completed an upcoming word prediction task (behavioural word-prediction experiment) on Mechanical Turk [219], a crowdsourcing marketplace on Amazon. Each HIT (Human Intelligence Task) corresponded to one of the five stories, meaning that we had 100 subjects per story. Demographic information was not collected, and participants remained completely anonymous. The HITs were only available for participants in the USA, Canada, Australia, England, and Ireland, aged 18 years or older. Potential participants saw the HIT on the Amazon Mechanical Turk platform and then decided whether to complete the tasks for compensation. Before beginning the HIT, participants were required to sign a digital consent form with three conditions: 1. You voluntarily agree to participate, 2. You are 18 years of age or older, 3. You are a native English speaker (as shown in Appendix C.4). Volunteers participating in our experiment were compensated with a fee of \$1.2 to \$1.5 per story, depending on the story’s length. Each HIT (per story) lasted 25 minutes on average.

Stimuli & Experimental Design

The experiment was presented to participants as an HTML webpage with the task flow controlled by jsPsych [228], which is a JavaScript framework for creating behavioral experiments running in a web browser. The words were displayed in 21-px Arial font with a black color at the centre of the webpage, while the page had a white background color.

The online experiment was conducted based on the cloze procedure as mentioned in Section 2.4.2. For each story, the title was initially displayed, and then the first two words in the content were revealed, where participants were then asked to predict (by typing) the next word. After providing their prediction, the actual upcoming word was disclosed, where participants were then asked to predict the next word in the sequence. Once ten words were displayed on the screen, the left-most word was

then removed, and the upcoming word was presented. This procedure was repeated until participants provided their predictions for all upcoming missing words. The ten-word sliding window is illustrated in Figure 4.1, and Figure 4.2 shows the display used on the website.

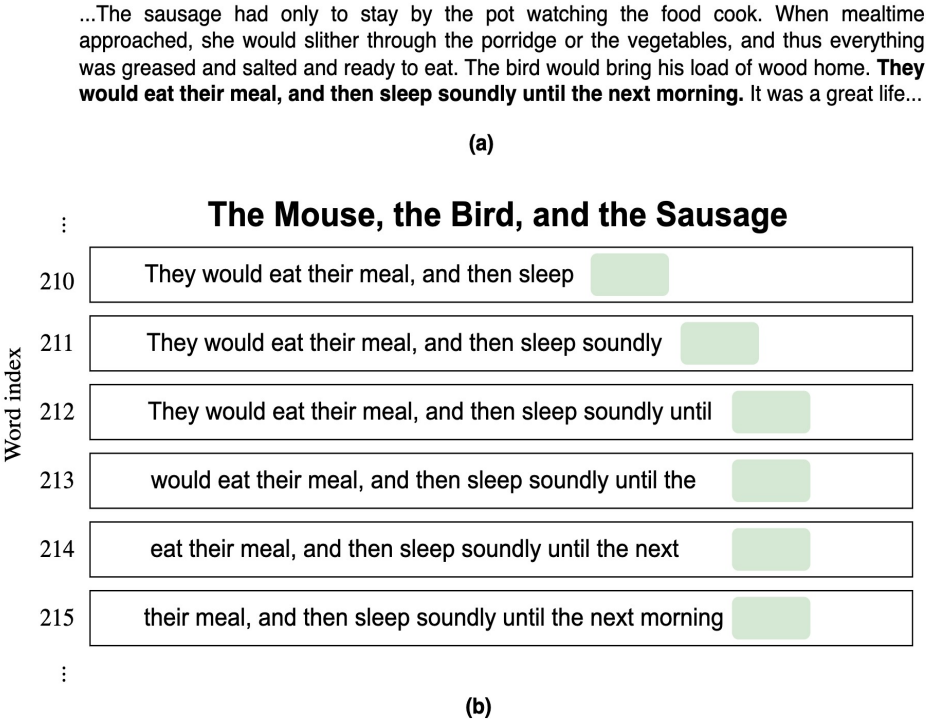


Figure 4.1: The stimulus was transcribed in article 1 for the behavioral experiment. (a) Transcript (b) A ten-word sliding window example.

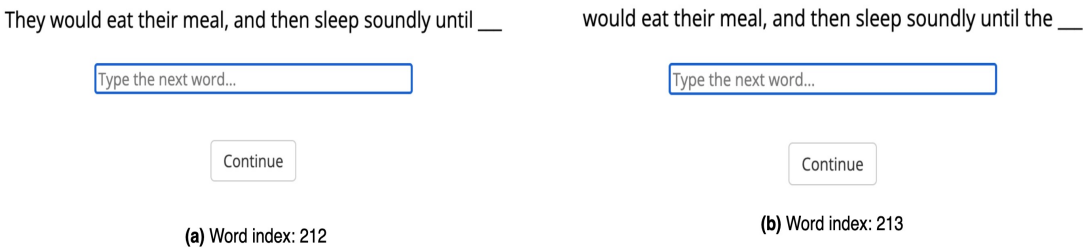


Figure 4.2: A ten-word sliding window was presented on the website for the article 1. (a) The next word is “the” (b) The next word is “next.”

Since our main focus is on next-word predictability, we designed our experiment based on the cloze procedure due to its application in brain activity analyses for reading and audio experiments. Goldstein et al.[76] conducted a cloze experiment using a sliding-window behavioural paradigm similar to ours. In their experiment, participants initially saw two words from a story, predicted the next word, and then

continued to predict subsequent words as they were revealed one by one. After displaying ten words, the earliest word was removed to make space for the next word in the sequence. Varda et al. [229] implemented a similar experimental design. Additionally, they conducted a cloze experiment based on the original EEG-based reading dataset [99, 230]. The materials used included a total of 205 English sentences, ranging from 5 to 15 words, with an average of 10 words. This average of ten words was also the number we chose for our experiment. Although participants had access to more context in the cloze experiment compared to the EEG-based experiment in Section 4.2.2, both studies achieved significant results in analysing predictability in human brain activity based on cloze probability.

Procedure

Participants on Mechanical Turk [219] found the experiment by browsing for HITs. After accepting the HIT, the participants were provided with task instructions and a web link to navigate to the experiment’s website. An informed consent form was displayed after the welcome screen. Participants then proceeded with a test block to familiarise themselves with the experiment. Before moving on to the next step, we emphasised three important points to ensure data quality and prevent technical interruptions during the experiment: 1. Please do not skip words, 2. Please do not use online resources and, 3. Please do not reload the page. After completing the entire story (i.e. providing a complete set of upcoming word predictions), they were then asked to answer some short questions about the text they had just read to test their comprehension (included in Supplementary File 2). Finally, a survey code was provided at the end of the experiment. The participants could then use this code to complete the HIT task on the Mechanical Turk platform. A task was considered successful if the participant adhered to the rules and used the exact survey code provided by the experimenters. After experiment completion, participants were not allowed to take part in further experiments in this study.

EEG-based Reading Experiment

All details regarding the participants, stimuli, experimental design, and procedure of this EEG-based reading experiment have already mentioned in Section 3.4.1.

4.2.3 Data Acquisition

Amazon Mechanical Turk Dataset

We used the Amazon Web Services to save collected data from the HITs. Amazon S3 [231] was used for storage to save the raw data after participants completed the task. Each survey code corresponded to a folder for each participant, and the raw data was saved in a CSV file.

EEG Dataset

All information about the acquisition of the EEG dataset has been provided in Section 3.4.2.

4.2.4 Data Preprocessing

Amazon Mechanical Turk Dataset

For each response in the prediction data, we remove punctuation and convert them to lowercase. Additionally, we also conducted checks to ensure the data quality. Namely, any spelling errors were corrected; for instance, ‘mornin’ is a typo for ‘morning’, ‘brige’ is a typo for ‘bridge’. Another scenario involves participants typing too quickly, pressing adjacent keys on their keyboards, resulting in grammatically incorrect words such as ‘sround’ instead of ‘around’ because the character ‘s’ is next to ‘a’. Importantly, these errors were not due to participants predicting incorrectly but would ultimately impact subsequent analyses. If a (misspelled) word only differed by one letter to its correct version (i.e. including a letter, removing a letter, or changing a single letter), the misspelling was rectified in the preprocessing stage.

Each response corresponded to a word at a unique location in the article. Keeping the words as indices ostensibly results in duplicates for analysis. Importantly, however, even though the words are the same, people may have different predictions or thoughts based on the context of the sentence, the topic, or their knowledge [6, 232]. Thus, we combined the position and story index for each word to create new indices.

In order to organise the necessary information for use, we designed three schemas corresponding to questions and predictions. All 100 results for each story were also saved in CSV files. We have published the dataset (both raw and preprocessed data) on the Open Science Framework (OSF). Please refer to the OSF project [39] for more information about data organisation, an introduction, size, authors, and schema of the DERCo dataset.

EEG Dataset

As discussed in Subsection 3.4, our results demonstrated that although the Autoreject + ICA + ICLabel pipeline outperformed other approaches in terms of mean-based SNR improvement, it was also the most time-consuming preprocessing pipeline. In contrast, the FASTER + ICA + ICLabel pipeline achieved the highest decoding accuracy in the classification task and consumed the least computational time compared to other pipelines. Consequently, we combined these two pipelines to preprocess the EEG data in this study, aiming to balance accuracy and efficiency.

The raw epochs were then input into an EEG epochs data preprocessing pipeline (as described in Figure 4.3) to create preprocessed epochs. The description of the preprocessing is outlined below.

First of all, the raw EEG epochs were imported into the pipeline using the MNE package version 1.4 [233]. Common average referencing (CAR) was applied to all data. The next stage of the pipeline aimed to mitigate artifacts in the EEG recordings. The initial step involves outlier removal using FASTER [168]. Epochs were

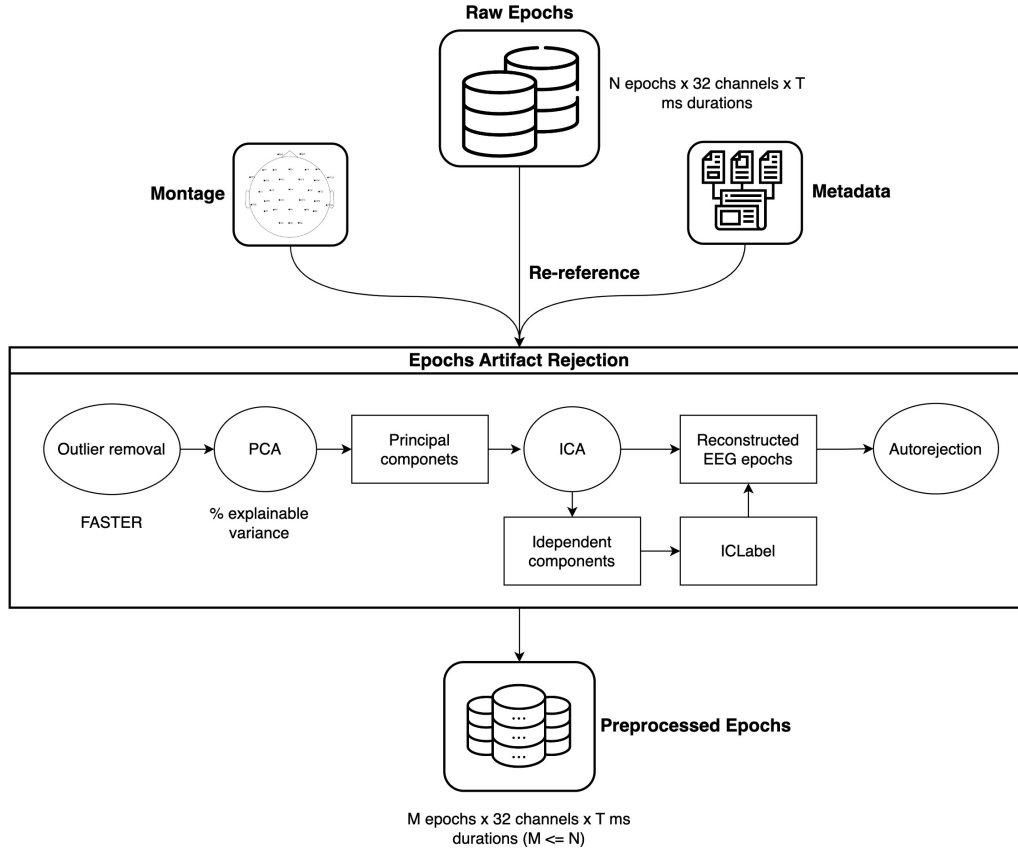


Figure 4.3: The EEG epochs data preprocessing pipeline.

rejected based on a global threshold on the z-score, exceeding 3 standard deviations for epoch amplitude, deviation, and variance range. The outputs were then preprocessed by applying ICA [234] to further improve the SNR. The number of principal components generated from the pre-whitening principal component analysis (PCA) step was selected based on the percentage of explainable variance. In this study, 99% explainable variance was used, which was accounted for by the principal components retained after dimensionality reduction.

ICA allows the decomposition of EEG signals into ICs, and then the corrected EEG signals can be obtained by discarding ICs containing artifacts [180]. The Picard algorithm [194] was used for maximum likelihood independent component analysis. To automatically determine which ICs contain artifacts, we applied the ICLabel classifier [171]. The remaining ICs were then reconstructed to create clean EEG epochs.

Following the ICA process, Autoreject [170] was applied for automatically iden-

tifying and removing artifacts. Autoreject uses unsupervised learning to estimate the rejection threshold for the epochs. In order to reduce computation time, which increased with the number of epochs and channels, we fitted Autoreject on a representative subset of epochs randomly picked from 25% of the total epochs. The number of cleaned epochs ultimately was less than or equaled the number of raw epochs. Importantly, this means for some word indices in stories there are no corresponding EEG epochs after trial rejection for some participants.

Data Records

The dataset is available on the OSF website [39]. There are two main folders corresponding to two experiments: the Behavioural Word-Prediction Experiment and the EEG-based Reading Experiment. The behavioural word-prediction experiment folder consists of three subfolders: “experienced_question”, “prediction”, and “question”. Each subfolder contains 5 CSV files corresponding to the participants’ responses across 5 articles. Supplementary File 2 contains all the questions asked in the experiment.

In the EEG-based reading experiment folder, there is an “answers” folder containing the answers of all participants collected after they completed reading each article. The main folder “EEG_data” has two subfolders: “raw” and “preprocessed”, which have the same structure. They contain 22 subfolders, each corresponding to a participant. Each participant’s folder has five subfolders for five articles, named “article_0”, “article_1”, etc. In each article folder, there is an epoch file in *fif* format. The directory tree for the repository of the DERCo project is shown in Figure 4.4.

4.3 Technical Validation

In this section, we examine next-word predictability — a key aspect in reading behaviour [235] — to validate that the word-level predictability measures calculated in the behavioural word-prediction experiment dataset can be meaningfully inte-

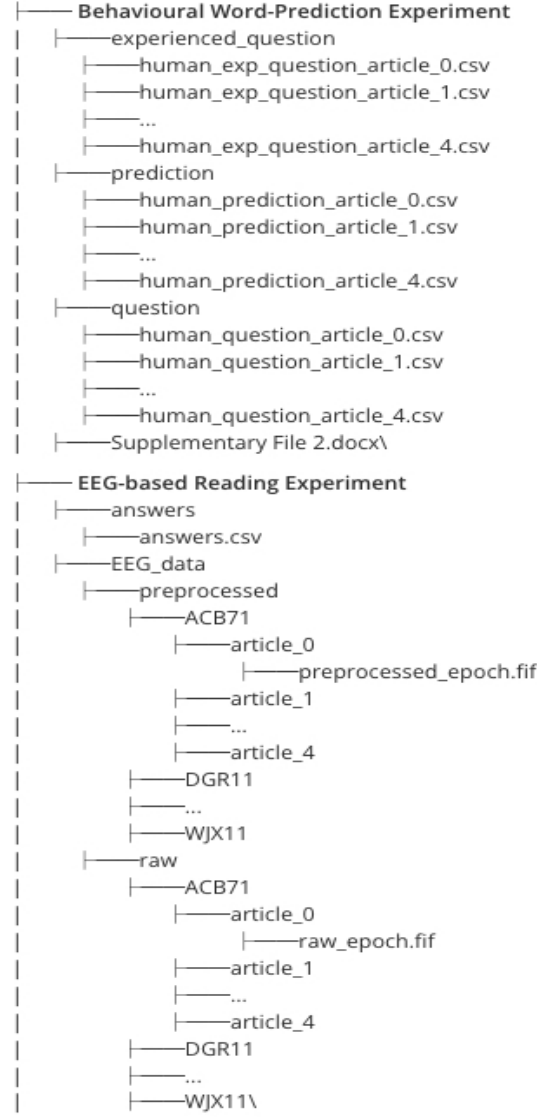


Figure 4.4: The structure of the dataset

grated with the EEG-based reading experiment dataset. To achieve this, we created the DERCo dataset using data collected from two experiments: the “Behavioral Word-Prediction Experiment” and the “EEG-based Reading Experiment”. After applying the techniques described in the Data Preprocessing section, we obtained neural responses based on word-level and semantic-level information. However, we could not collect the exact predictions of the next words for each participant using the EEG-based reading experiment. Therefore, the behavioral word-prediction experiment was carried out with a large sample size to measure the predictability of each word (at a group level) in the texts. By combining the data collected from both experiments, we have EEG responses and predictability values for each word

in the articles. The data in the behavioral word-prediction dataset was validated to demonstrate participants’ proficiency in predicting subsequent words and to assess participants’ reading comprehension of the articles. Given the significant results at the semantic level, we then analyse next-word predictability based on EEG signals at the word level using the predictability measures calculated from the behavioural word-prediction experiment dataset. This validation exercise is crucial to demonstrate that these two datasets are meaningfully aligned, ensuring they can be used together for effective analysis and investigation.

4.3.1 Behavioural Word-Prediction Dataset

Using the *general questions* in each story (see Appendix C.2), we estimated a participant’s prior knowledge of the recently read narratives and identified whether the articles used excessively advanced or unusual vocabulary. The result indicated that a significant majority of subjects, ranging from 98% to 100% of the total, had not read the stories before. Despite the unfamiliar stories, most participants agreed that no advanced or uncommon words were used. Only the final and longest story had agreement from 89% of participants, whereas the other articles achieved percentages between 95% and 98%.

As an estimate of participants’ overall reading comprehension (as a proxy measure to ensure participants correctly read the texts), a proportion was calculated based on the number of correct answers to multiple-choice questions for each subject. As a result, a substantial number of participants, between 98 and 100 participants, showed a performance greater than or equal to 60%, indicating their ability to comprehend and recall the stories’ content to provide accurate answers after reading.

The ability to anticipate upcoming words is commonly linked with the “cloze probability” or “cloze value” [47, 236, 237]. It is calculated as the proportion of participants that picked the same target word for the next word prediction. Cloze values range from 0 to 1, where when 0 indicates that no participant predicted the same next word, and 1 indicating all participants predicted the same next word.

Notably, high predictability was observed for all words from all parts of speech in a sentence, and not just the last words in a sentence. More than 1000 words out of about 3,000 words in all articles had a cloze probability higher than 0.5. Of these, 62.74% of words were functions words, 20.32% were nouns, 6.24% were verbs, 5.35% were adjectives and verbs, and 5.35% were classified as other. The parts of speech were identified using the Python library IPA [238].

Table 4.2 provides further information about prediction accuracy for each story and word categories, including content words (nouns, verbs, adjectives, and adverbs), function words, and others. The accuracy refers to the percentage of times participants correctly predicted the upcoming words in the story. We found that the number of sentences decreases in the order of articles 4, 5, 1, 2, and 3, corresponding to a reduction in the percentages of content words in the same sequence (38.77%, 35.48%, 31.67%, 24.46%, and 19.59%). As a result, articles with more sentences show higher accuracy in predicting content words (nouns, verbs, adjectives, and adverbs). For all words, we included the descriptive statistics in Appendix C.5. These findings suggest that humans are proficient in anticipating next words in narrative texts when asked to do so.

Story	Accuracy	No. words ($p \geq 0.5$)	% nouns	% verbs	% adjs & advs	% function words	% others
1	45.02%	180	23.89%	5.00%	2.78%	60.56%	7.78%
2	40.95%	139	15.83%	2.16%	6.47%	72.66%	2.88%
3	41.34%	148	14.19%	2.70%	2.70%	75.00%	5.41%
4	55.32%	294	21.77%	7.14%	9.86%	56.80%	4.42%
5	41.60%	248	22.18%	10.48%	2.82%	58.47%	6.05%
Total	45.17%	1009	20.32%	6.24%	5.35%	62.74%	5.35%

Table 4.2: Human next-word prediction performance. % shows the percentage of correctly predicted words for each story and part of speech. “No. words ($p \geq 0.5$)” shows the number of words with a cloze probability above .5.

4.3.2 Neural Responses in Reading

The experimental design differed between the behavioural word-prediction experiment and the EEG-based reading experiment due to their distinct requirements,

as mentioned in the procedure sections for both experiments. For the online experiment, the cloze procedure was appropriate, allowing participants to type their predictions for upcoming words. However, this method was unsuitable for the EEG-based reading experiment, which required precise time-locking of stimulus presentation to enable ERP analysis [158]. Despite the differences in the stimulus presentation paradigms between both experiments, we show significant results, demonstrating that word-level predictability measures calculated in the behavioural word-prediction experiment dataset can be meaningfully integrated with the EEG-based reading experiment dataset, as outlined below.

EEG data was segmented into epochs, starting 50 ms before and lasting 550 ms after the word-onset. Each grand-averaged word ERP was calculated by averaging the neural responses across participants for each channel generated by a given word. Figure 4.5 shows the EEG averages across all words for 32 electrodes in the first article between all subjects and a selected individual (user-id = JPY86). Consistent with the findings of the previous study by Dimigen et al. [199], we can see a positive peak at around 100 ms post-word onset, also known as P100. A second positive peak (P200) can be identified at around 200 ms after word-onset.

For a technical validation, that incorporated both the behavioural word-prediction dataset (Mechanical Turk) with the EEG dataset, we investigated N400-like components within the 300 - 500 ms time window. The ERP was examined across midline electrodes Fz (Midline Frontal), Cz (Midline Central), Pz (Midline Parietal), and Oz (Midline Occipital). Using midline electrodes in EEG studies provides balanced recordings of brain activity. They capture significant neural activity from central brain regions involved in cognitive processes, reduce hemispheric lateralisation effects, and ensure consistency and comparability due to their standardized use across studies [239].

We used two non-overlapping groups of *low predictable* words (cloze $p \leq 0.06$, $n = 989$), and *high predictable* words (cloze $p > 0.52$, $n = 962$), calculated from the Mechanical Turk dataset. Figure 4.6 provides a more detailed view of the topogra-

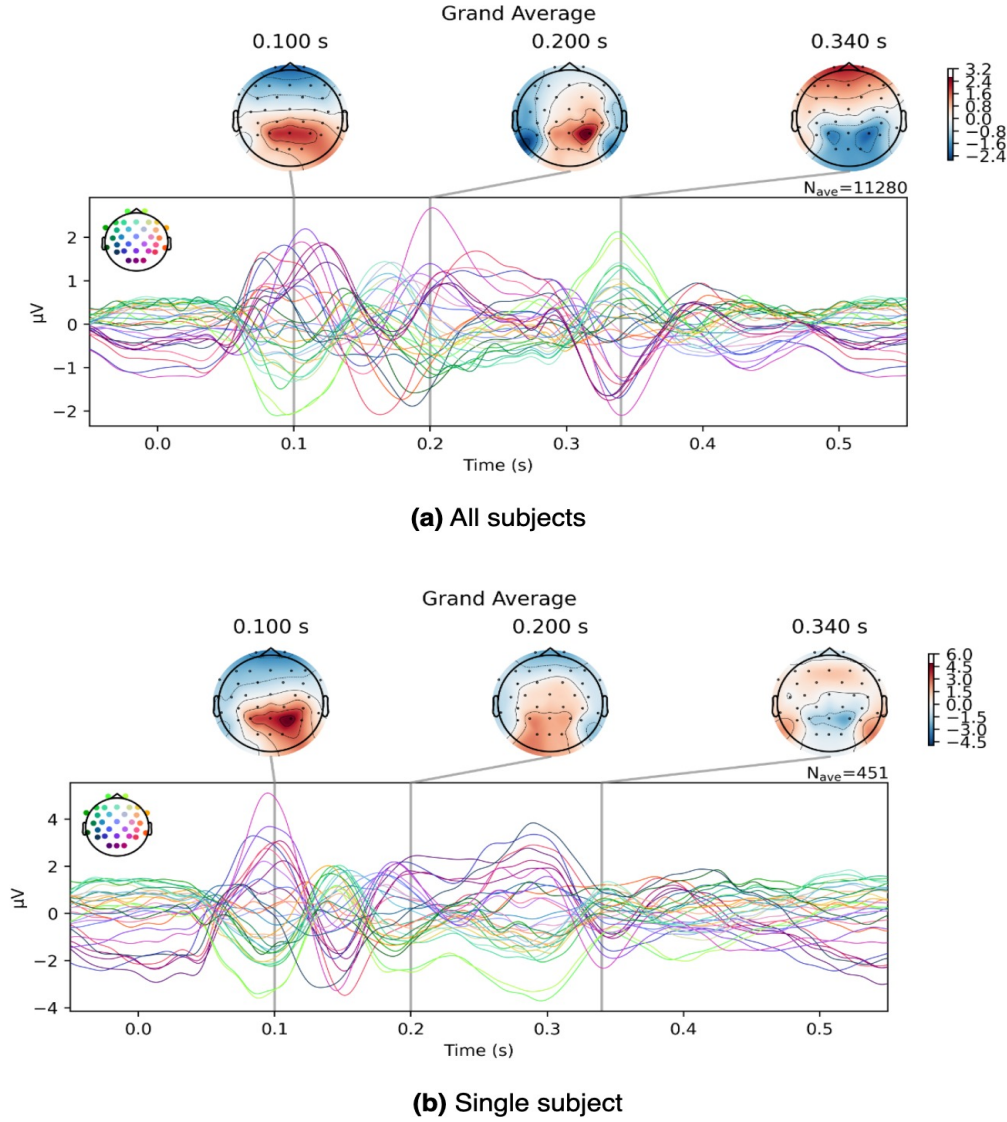


Figure 4.5: Butterfly plots for EEG averages across all words in article 1 **(a)** across subjects **(b)** single subject (JPY86). Each time-series line graph corresponds to an electrode.

phies of response differences between conditions across time, computed averages of 300 - 500 ms range in RSVP and RSVP with flankers, respectively. The color intensity indicates the mean ERP amplitude across all participants for the low and high predictability groups. As shown in Figure 4.6, the effect of word predictability actually modulated the ERP evoked after the word onset.

In order to capture variability in cloze values, we applied an independent samples t-test on the high and low predictability, and plotted the ERP waveforms between these conditions to examine ERP activity in the time window of 300 - 500 ms.

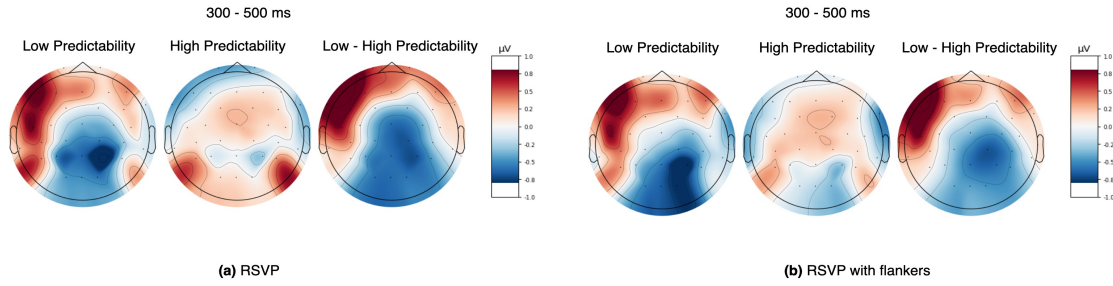


Figure 4.6: Scalp distributions of the response for low predictable words, high predictable words, and the difference between low and high predictable words (low minus high) in **(a)** RSVP experiment and **(b)** RSVP-with-flanker experiment. Amplitudes (μV) were averaged in 300-500 time window (for all electrodes).

Results from the t-test on midline electrodes and the corresponding visualisations are presented in Figure 4.7. Post-hoc t-tests revealed significant differences ($p < 0.05$) between high and low predictability at all mid-line electrodes in between 300 and 500 ms post stimulus in both RSVP and RSVP-with-flankers paradigms. If conducting t-tests for all channels, approximately 70% electrodes showed significant differences ($p < 0.05$) in both experiments. In the RSVP paradigm, nine electrodes (FC1, C3, TP9, P7, TP10, T8, FT10, FC6, and F4) had no predictability effect in the 300 - 500 ms range between two conditions. The predictability effect was not found for ten channels (Fp1, FC1, C3, P7, TP10, T8, FT10, FC6, F4, and Fp2) in the RSVP-with-flanker experiment at the same time window.

These validation results show one potential analysis trajectory that necessitates the use of both the behavioural word-prediction dataset and the EEG data. Moreover, it validates that the word-level predictability measures calculated in the behavioural word-prediction experiment dataset can be meaningfully integrated with the EEG-based reading experiment dataset. Importantly, such analyses could not be performed with one dataset alone.

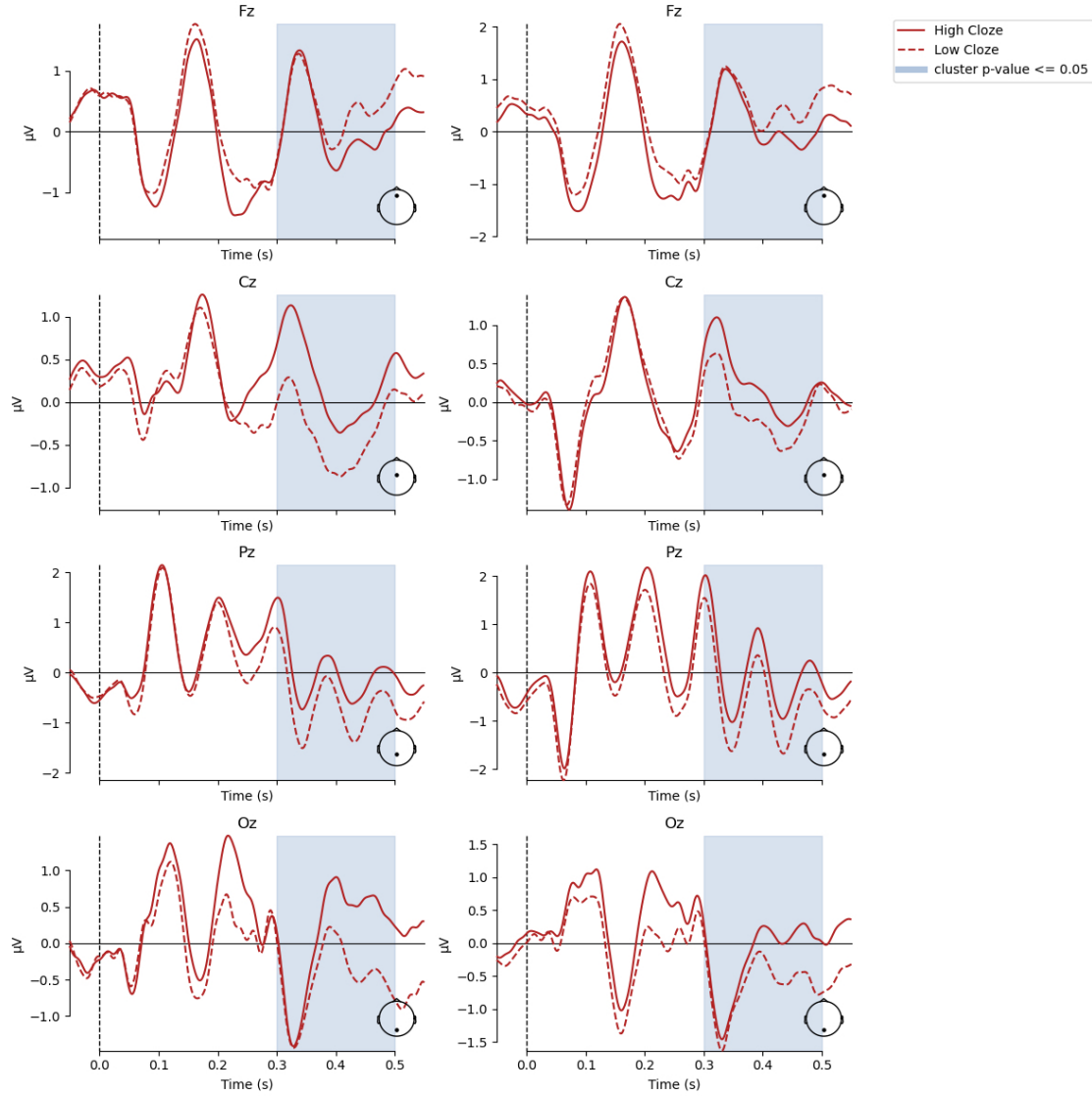


Figure 4.7: ERP activity for midline electrodes in RSVP experiment (left) and RSVP with flankers experiment (right). The time range of significant differences ($p < 0.05$) between 300ms and 500ms post-stimulus are indicated.

4.4 Discussion

4.4.1 Conclusion

In this chapter, we presented a new, freely available corpus of electrical brain activity recordings during reading and annotation. This dataset demonstrates the relationship between these two reading paradigms. We described the materials and experimental design in detail and conducted an extensive validation to ensure the quality of the recorded data. Since this corpus is tailored for next-word prediction

examination, the applications and reusability of this data are extensive. While this work does not directly address a specific research question in the dissertation, the DERC0 dataset results from the first research question (RQ1) and serves as a crucial foundation for investigating the subsequent research questions (RQ2 and RQ3) in the thesis.

The provided word-level EEG features can be utilised to enhance and evaluate LMs, for example, by assessing linguistic phenomena in neural models using neurolinguistic data. Additionally, because the words include semantic relation labels and annotations from all participants, the dataset can be broadly applied to tasks such as relation extraction and classification. Finally, the two carefully designed reading paradigms enable comparisons between normal reading and reading during annotation, which may contribute to improving the manual labelling process and enhancing the quality of annotations for language models

We believe that this dataset represents a valuable resource for natural language processing. As demonstrated in Chapter 5, this dataset is useful for advancing research into the human reading and language understanding process at the level of brain activity. Moreover, in Chapter 6, the EEG data is used to explore human-like reading predictability in various LMs.

4.4.2 Limitations

Although the EEG and behavioral word-prediction datasets provide a novel perspective on reading behaviors, some limitations still exist. Firstly, while the experimental design for EEG collection simulates a closer approximation of the natural reading scenario, it is still not completely naturalistic. However, because we focused on word-level EEG responses, a free-viewing paradigm was deemed unsuitable, and moreover, due to how eye movements generate artefacts in EEG. This is because people tend to skip easily predictable[128] or high-frequency words[129]. Additionally, temporal overlap occurs between the ERPs during longer saccades or regressions in a sentence when using free-viewing paradigms[111]. Thus, researchers need to

consider this when proposing hypotheses using this dataset.

Secondly, another limitation of the experimental setup is that instead of using two different groups of participants for the EEG experiment, one for the classic RSVP experiment and another for the RSVP-with-flankers experiment, we used the same participants for both. We used RSVP for the first three stories and RSVP-with-flankers for the remaining two, which reduced the total number of words available for analysis in each condition. This decision allowed for the EEG dataset to support analysis of the two RSVP conditions, whilst maintaining a sufficient number of participants for each.

Lastly, our EEG experiment participant pool was limited due to an unbalanced division of gender (approximately 68% male and 32% female). This gender imbalance should be remedied in future to achieve equal representation of each gender, because this gender distribution could have skewed results or prevented an effect from becoming significant due to lack of power [240]. However, our research primarily aimed to analyse next-word prediction behavior in adults aged 18 years and older, rather than focusing on gender differences. It should be acknowledged that several similar brain activity studies in reading experiments have also had unbalanced sample sizes for gender [201, 214, 223–225].

Chapter 5

Decoding EEG Signals to Explore Next-Word Predictability in the Human Brain

This chapter aims to address the research question “*Are neural responses to upcoming words sensitive to quantitative differences in lexical categories and grammatical categories?*” While traditional ERP analyses, such as univariate approaches, have limitations in providing interpretable visualisations, capturing individual variability, and recognising complex patterns, advanced EEG decoding techniques offer enhanced sensitivity, multivariate analysis capabilities, and greater interpretability. This study demonstrates the efficacy of EEG decoding techniques while providing empirical evidences of the neural mechanisms underlying bottom-up (related to high-order linguistic structure) and top-down (related to next-word predictability) processes, which interact to guide comprehension during reading. We examined how predictability, measured by cloze probability, influences brain response differences between content words (nouns, verbs, adjectives, and adverbs) and function words (pronouns, determiners, auxiliaries, and adpositions), particularly focusing on the N400 time window (300-500 ms post-stimulus), known for its sensitivity to semantic incongruities and predictability in reading. Our results in-

indicated that highly predictable words generate more significant differences in N400 effects across lexical classes compared to less predictable words. Moreover, significant differences in N400 responses between high and low cloze probability levels were greater for content words than function words. Among grammatical categories, verbs exhibited greater N400 differences than nouns in content categories, but nouns carried more distinct information about their predictability. In contrast, determiners showed more significant N400 effects than other function categories, reflecting their varying roles and predictability in narrative contexts.

5.1 Introduction

5.1.1 Next-Word Prediction Behaviour Observation

The cloze probability measure serves as a proxy for word predictability within a given context (as described in 4.3.1). Words with high cloze probability typically involve a reduced cognitive load because they align closely with contextual expectations. In such cases, the brain can efficiently predict and integrate these words based on the surrounding context, often resulting in attenuated neural responses, particularly in regions associated with effortful processing [241]. Conversely, words with low cloze probability require increased cognitive effort for processing. When encountering these less predictable words, the brain might need to engage additional neural resources to successfully integrate them into the context, potentially resulting in enhanced or distinct patterns of neural activity [84].

Traditionally, words are grouped into two basic types of words: content (open-class) and functions (closed-class) words. Content words include nouns, verbs, adjectives, and most adverbs. These word types convey specific semantic information about objects and events. Function words on the other hand comprise pronouns, determiners, auxiliaries, and adpositions, and serve as structural elements in phrase construction. Importantly, the different word classes differ on a variety of other dimensions as well [242, 243]. For example, content words, on average, are less fre-

quent, have more letters, are more concrete, have higher ratings for imageability, etc. Furthermore, the semantic representation of verbs is typically more complex than that of nouns because verb meanings are more flexible than noun meanings in different semantic contexts [244, 245].

5.1.2 ERPs Effects in Lexical Classes

Neville et al. [34] indicated that content words elicited a larger amplitude N400 component, peaking between 350-400 ms over posterior regions, whereas function words elicited a much smaller N400 response. Münte et al. [35] supported these findings, noting that both word classes could elicit N400 responses, but the amplitude was significantly smaller for function words, particularly high-frequency ones, which also showed a frontal negativity from 400 ms onwards (N400–700). Osterhout et al. [36] observed similar patterns, with a large negative peak at 400 ms (N400) that was most pronounced over midline and right-hemisphere electrode site for content words, while function words exhibited more enhanced negativity from 400-700 ms (N400-700). He et al. [208], focusing on high-frequency words, found that content words elicited more negative N400 responses at midline electrode sites compared to function words, consistent with previous findings [34, 35]. Collectively, these studies indicate that the N400 component varies significantly between content and function words, potentially reflecting different access processes for different lexicons for the two word classes.

5.1.3 ERPs Effects in Grammatical Categories

Besides research concerning word classes, grammatical categories and predictability also influence N400 amplitude, reflecting sensitivity to a common stage of bottom-up processing. The N400 amplitude is significantly more enhanced for nouns that are unexpected or inappropriate given their preceding context [246]. The N400 response was found for most nouns, and even the most highly unexpected nouns, even when they are contextually inappropriate. The more predictable the noun, the

less negative the N400 response [47]. Larger amplitudes for verb processing than noun processing has been observed in the N400 time window [247, 248]. Münte et al. [35] observed that the N400-700 effect was also sensitive to the very high frequency closed class words, such as determiners.

Additionally, Osterhout et al. [36] reported that different categories elicited a negative-going component peaking between 280 and 400 ms, with specific peaks for articles (N293), prepositions (N313), pronouns (N306), auxiliary verbs (N310), and later for verbs (N319) and nouns (N328).

5.1.4 Traditional ERPs Analysis Techniques

In major studies testing hypothesis on word classes and predictability with respect to brain language processes, researchers frequently use univariate analyses such as t-tests and analysis of variance (ANOVA) [34–36, 247]. These inferential statistics help researchers understand how neural signals differ between different lexical classes, enabling them to deduce the systematic relationships between word groups and predictability affecting ERPs. These analyses involve measuring grand-averaged ERP responses, characterised by a sequence of peaks and troughs within a specific time period for each condition at one or more electrode sites, and indicate whether the amplitude differences between conditions or groups are statistically significant. These conventional tests often fall short in capturing the complete range of cognitive processes involved due to certain limitations as mentioned below.

EEG data is typically recorded from 32, 64, 128, or even more electrodes [249]; however, the analyses often concentrate on a single electrode site or a subset of electrodes where the component of interest is largest to maximise the signal [250]. Researchers then apply univariate statistical analysis methods to examine different conditions on each electrode site. This classic approach significantly increases the risk of false negatives [251, 252], and is associated with low statistical power [158]. Although mass univariate analyses allows researchers to run tests at all time points and electrode sites between conditions, they also encounter reduced statisti-

cal power [253]. Another major shortcoming of these univariate approaches lies in their assumption of the same scalp distribution for a given ERP component across all subjects. However, individual subjects typically have unique cortical patterns [254], especially in reading.

5.1.5 Classification-Based Decoding Techniques

To overcome the limitations of traditional statistical techniques, researchers have increasingly adopted multivariate pattern analysis (MVPA) approaches in recent years [255, 256]. Several ML algorithms, including support vector machines (SVM), linear discriminant analysis (LDA), and random forests (RF), also known as MVPA classification or classification-based decoding, have demonstrated exceptional robustness and adaptability in distinguishing subtle distinctions between different stimulus groups by analysing voltage patterns across multiple electrode sites [257]. These decoding techniques have been proven more effective compared to conventional ERP analysis in classifying EEG signal patterns at the individual subject level, addressing the challenge of variability in individual neural responses [258]. Additionally, MVPA methods can analyse EEG signals before word onset in reading experiments without being as affected by ERP overlap issues [259, 260], which are difficult to adequately address with statistical and mathematical analysis techniques alone [158, 261]. This is evident from studies related to word information activation and prediction during reading [76, 204], as well as sentence processing [262–264].

In this chapter, in addition to using the conventional ERPs analysis method, we leverage the emerging decoding approach to add evidence about the nature of top-down processes and their relation to bottom-up information processing during reading comprehension. Using the DERCo dataset [39] (previously described in Chapter 4), a linguistic resource that integrates EEG data with next-word prediction information from participants reading narratives, we investigate influences of predictability on neural responses between linguistic units. This work’s purpose is also to address **RQ2** by examining various dimensions:

- Do differences in word class explain the N400 differences observed at various cloze probability levels?
- To what extent does the recognition of grammatical categories drive top-down processing?
- Do highly predictable words generate stronger N400 effects across lexical classes compared to less predictable words?

Furthermore, by using an EEG-based classification approach, we aim to investigate “Does the decoding-based approach increase effect sizes and statistical power in detecting N400 amplitude differences between two conditions in reading, compared to traditional ERP analysis methods?”

5.2 Method

5.2.1 EEG Recording and Data Preparation

This study made use of the DERCo dataset collected by Quach et al. [39]. For more detailed description of the participants, experimental design, recording procedures, data processing, please refer to Chapter 3 and Chapter 4. High-density EEG data was recorded using electrode 32-channels placed according to the international 10–20 system [193] over the left hemisphere electrodes (C3, CP1, CP5, F3, F7, FC1, FC5, Fp1, FT9, O1, P3, P7, T7, TP9), midline electrodes (Cz, Fz, Oz, Pz), and right hemisphere electrodes (C4, CP2, CP6, F4, F8, FC2, FC6, Fp2, FT10, O2, P4, P8, T8, TP10) locations.

As the purpose of these analyses was to examine neural responses across word classes, grammatical categories, and varying cloze probability levels, it was necessary to incorporate this information into the metadata. Parts of speech were identified using the Python library IPA [238]. This library allows us to extract and categories the grammatical function of each word within a sentence. For example, in the sentence “She quickly reviewed some detailed reports with her team,” the parts

of speech are categorised as follows: She (pronoun), quickly (adverb), reviewed (verb), some (determiner), detailed (adjective), reports (noun), with (adposition), her (pronoun), team (noun). In terms of cloze probability, we utilised the same two distinct groups (high versus low) extracted from the Amazon Mechanical Turk dataset as in the original study (see Chapter 4). This decision was based on their validation, which indicated significant differences in N400s (300-500 ms) across most electrodes. Two different sets of EEG epoch data and annotations were extracted according to word classes (content and function words) and cloze probability (high and low cloze word levels).

After preprocessing, the number of trials for high and low cloze probability levels in terms of lexical classes or each grammatical category varied between participants. To address this, we standardised the number of trials used for analysis across subjects by determining the smallest number of trials at each cloze probability level across participants and setting these as a baseline. We then randomly selected an equivalent number of trials from other participants, ensuring that the distributions of these selected trials (based on cloze values) closely matched the baseline distributions. This random selection process was guided by the Kolmogorov-Smirnov test, which assessed the similarity between distributions. The process iterated up to 5000 times, each time randomly selecting a subset and computing the KS statistic, with the subset yielding the lowest KS statistic considered the best match. For example, for content words, the smallest number of trials for high cloze probability was observed in subject “TRA37” with 212 trials, and that of low cloze probability was found in subject “QFT39” with 413 trials. We extracted the cloze values from each trial to build the distribution baseline, and then used this distribution to select the corresponding EEG data from other participants.

In our EEG epochs, the number of trials for adjective and adverb words was significantly small, less than 16 trials. This did not meet the minimum sample size required to produce statistically significant results with a reliable effect size [265]. Moreover, Rajput et al. [266] demonstrated that as sample size increased,

effect sizes and classification accuracies between two classes also increased, while variances in effect sizes decreased. This effect was observed when they examined five ML classifiers, starting with a sample size of 16. Consequently, we did not analyze these two grammatical categories to avoid potentially misleading outcomes. Nevertheless, we still included words from these classes in the content word category for the lexical group and cloze probability analyses.

5.2.2 Univariate Analysis

Traditional univariate analysis involves calculating a difference wave between two or more conditions for a given ERP component for each participant. In our experiment, the EEG epochs were recorded from 32 electrode sites over 20 participants, requiring multiple comparisons. The most common statistical approach to dealing with this issue uses data clustering combined with permutation testing [267]. By examining data across all electrode sites, these analyses ensure comprehensive spatial coverage, increase the statistical power, control for multiple comparisons, and allow for the identification of a cluster of electrodes where the component of interest is largest [251, 253, 267]. In the standard univariate approach, the difference in ERP amplitude between the two conditions is quantified at multiple time points. For the mass univariate analyses, we have a significantly larger set of (channel, time) pairs, also referred to as samples, wherein we aim to examine the N400 effect.

Difference waveforms reflecting next-word prediction effects were computed for two word classes by subtracting ERPs of function words from ERPs of content words. The same procedure was applied for lexical class effects, and grammatical category effects, on cloze probability levels. The time course of the difference waveforms between the two conditions was compared in the latency range from -100 to 500 ms after word onset, with a preliminary point-wise independent samples t-test identifying clusters that included data points falling below the cluster alpha level ($p < 0.05$). This p-value directly corresponds to two t-values given the number of participants, e.g. for 20 subjects the t-values corresponding to $p < 0.05$ are -2.093

and 2.093. After calculating the t-values for each time point, the sum of all t-values in each cluster was used to create several cluster-level statistics.

This preliminary clustering procedure was followed by a permutation procedure that involved randomly shuffling the cluster values N times to determine the size of clusters expected by random chance. For each permutation, new clusters were identified and the corresponding cluster values were derived. The largest absolute t value (t_{max}) across all time points and channels amongst all of the cluster-level statistics is used to identify where within that permutation distribution it falls. Clusters were considered significant if the probability of observing a larger cluster value in the new distribution was below the significance level on both sides. Regions of interest were defined based on the scalp topographies of the time windows identified by the permutation procedures.

It is obvious that the mass univariate analyses require computation of multiple sample-specific t-values to evaluate the effect's reliability. However, in comparison to single-sensor analyses, the multiple comparisons problem is significantly stronger in this context: with 32 EEG sensors and 200 temporal points (300 and 500 ms post-stimulus), we generate 6,400 t-values. The cluster-based permutation test (10,000 times) for multi-channel analyses is similar to that for single-channel analyses, differing primarily in one aspect: rather than clustering based solely on temporal adjacency, we now cluster the selected (channel, time) samples based on both spatial and temporal adjacency. Once clusters of (channel, time) samples are identified, those with p-values exceeded the critical two-sided alpha-level (0.05) are considered insignificant. The false discovery rate (FDR) is calculated and used to correct for multiple comparisons [268]. Each significant cluster represents a contiguous block of activity across time points and potentially across electrode channels. From these clusters, we can extract the optimal group of electrode sites within the N400 window. By averaging the difference waves for all participants across an optimal cluster of electrode sites, we can approximate the location of an equivalent current dipole that best accounts for the N400 effect.

Note that if the approach does not yield any optimal electrode sites using FDR correction, we will select optimal channels without applying the correction. Additionally, if no optimal cluster is identified in the N400 time window (labeled as “None” in Table 5.1), all electrode sites in that region will be used for the analyses. Table 5.1 outlines all parameters for the testing cases and the results of the optimal clusters of electrode sites in the N400 time window across three scalp regions: left hemisphere electrodes, midline electrodes, and right hemisphere electrodes. For comparison with the SVM-decoding technique, we will use the same measurement time window, from 100 ms pre-stimulus onset to 500 ms post-stimulus onset.

5.2.3 SVM-Based Decoding Analysis

SVM is particularly effective in decoding EEG/ERP data due to their ease of implementation, ability to handle non-linear relationships in high-dimensional spaces, and strong performance with small training sets [257]. Compared to other ML models, such as linear discriminant analysis and random forests, SVM has demonstrated superior performance in examining N400 effects of prediction and semantic relatedness [258]. Therefore, we applied SVM in this study to compare its performance with that of the traditional univariate technique.

SVM was initially developed for binary classification problems [269] and has grown over the years [270]. The SVM algorithm aims to find an optimal hyperplane that separates two classes of data points. For our analyses, these classes would be content vs. function word classes, and high vs. low cloze probability words. Given a set of EEG epochs $\{x_i, i = 1, \dots, N\}$, where $x_i \in \mathbb{R}^N$ represent the feature vector for the i^{th} epoch, and their corresponding labels from two classes $y_i \in \{-1, 1\}$. The classification problem is formulated as:

$$y(x) = w^T \phi(x) + b, \quad (5.1)$$

where w is the weight vector, $\phi(x)$ is the mapping of the input feature space to a

Group	N	Time window	Baseline no. of trials		Optimal electrode cluster (300-500 ms)			Cluster p-value
			Condition 1	Condition 2	LH	M	RH	
Content	20	-100-500 ms	212 (High cloze)	413 (Low cloze)	P3, P7, CP1, Fp1, F7, F3, FC5, T7, CP5	Cz, Pz, Oz, Fz	T8, CP6	0.05
Function	20	-100-500 ms	351 (High cloze)	200 (Low cloze)	CP5, FT9, FC1, Fp1, F7, F3, FC5	Cz, Pz, Oz	FT10, FC2, C4, CP2, P4	0.05
Noun	20	-100-500 ms	135 (High cloze)	94 (Low cloze)	P3, P7, CP1, Fp1, F7, F3, FC5, T7, CP5	Cz, Pz, Oz, Fz	None	0.05
Verb	20	-100-500 ms	43 (High cloze)	192 (Low cloze)	P3, P7, CP1, O1, FC5, T7, CP5	Cz, Pz, Oz	T8, CP6	0.05
Pronoun	20	-100-500 ms	101 (High cloze)	48 (Low cloze)	Fp1, F7, F3	Cz, Pz, Oz	Fp2, F4	0.05
Determiner	20	-100-500 ms	152 (High cloze)	39 (Low cloze)	Fp1, F7, F3, FC5, T7, CP5	Cz, Pz, Oz	T8, CP6, FT10, FC2, C4, CP2, Fp2, F4, F8	0.05
Adposition	20	-100-500 ms	51 (High cloze)	64 (Low cloze)	None	None	T8, CP6	0.05
Auxiliary	20	-100-500 ms	41 (High cloze)	34 (Low cloze)	Fp1, F7, F3, FC5	None	None	0.05
High cloze	20	-100-500 ms	212 (Content)	351 (Function)	FC1, C3, P3	Cz, Pz, Oz	FC6, T8, CP6, FT10, FC2, C4	0.05
Low cloze	20	-100-500 ms	413 (Content)	200 (Function)	None	None	FT10, FC2, C4	0.05

Table 5.1: Results of the univariate analysis procedure. **N** refers to the number of participants. **Baseline no. of trials** refers to the number of trials used for analyses in each participant. **LH, M, RH** represent the Left Hemisphere Electrode, Midline Electrode, and Right Hemisphere electrode sites, respectively. **Cluster p-value** denotes the p-value threshold used for cluster-based permutation tests.

higher-dimensional space, and b is the linear classification bias.

SVM searches for the optimal hyperplane that has a maximum margin between the nearest positive and negative samples. This optimisation problem can be formulated as:

$$\arg \min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \quad \text{subject to: } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad (5.2)$$

where ξ_i are the slack variables for misclassification, and C acts as a regularisation parameter controlling the misclassification level for the training samples.

The introduction of the Lagrange multipliers $\alpha = \{\alpha_i \mid i = 1, \dots, N\}$ converts the problem 5.2 into a maximisation problem with respect to α , as explained in [271]. However, when the problem is very noisy, even though kernels can represent non-linear decision boundaries, the problem may become very hard to solve computationally. To address this issue, an approach is proposed that allows SVMs to make certain number of misclassifications while maintaining a maximally wide margin, so that other points are still classified correctly. This is obtained simply by introducing a control parameter, C , which allow the violation of the margin constraints [272]. This is called the soft-margin problem, for which the optimisation problem can be expressed using the following dual formulation:

$$\max_{\alpha} D_{\gamma}(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K_{\gamma}(x_i, x_j) \quad (5.3)$$

subject to:

$$\begin{cases} 0 \leq \alpha_i \leq C, \quad \forall i \\ \sum_{i=1}^N y_i \alpha_i = 0, \quad \forall i \end{cases}$$

where K_{γ} denotes the kernel function.

One of the most advantageous features of SVMs is their ability to use a nonlinear kernel function to transform input data into a high-dimensional feature space. In EEG studies, we typically have more than two electrode sites. Our EEG data was

recorded from 32 channels, and all sites were input simultaneously for decoding, resulting in a 32-dimensional space. The iterative learning process of the SVM devises an optimal hyperplane with the maximal margin between each class in the high-dimensional feature space. Consequently, the maximum-margin hyperplane forms the decision boundary for distinguishing the neural responses associated with different stimulus classes [272].

The radial basis function (RBF) kernel is one of the most widely used nonlinear kernels and can be regarded as a restricted version of general Gaussian distribution. The RBF kernel function for two points computes the similarity or how close they are to each other. This kernel is defined as:

$$K_{\gamma}(x, x_i) = \exp(-\gamma\|x - x_i\|^2), \quad (5.4)$$

where γ is the variance and our hyperparameter, and $\|x - x_i\|$ is the Euclidean distance between two feature vectors.

For better generalisation performance, it is sometimes beneficial to allow a few misclassifications when setting up the decision boundary. The trade-off between accuracy and generalisation is controlled by an adjustable regularisation parameter C . Noted that small values of C cause the optimizer to look for a hyperplane with a large-margin separation, probably result in misclassify some points. Conversely, large values of C lead to a search for a smaller-margin hyperplane to better classify all the training points.

5.2.4 Analysis Procedures

The decoding process has separate training and testing phases. The flowchart in Figure 5.1 illustrates the approach to EEG data decoding using an SVM with an RBF kernel. The process begins with the collection of neural data from multiple subjects, where each participant's data is organised into a 3-dimensional array corresponding to trials, electrode sites, and time points. To reduce the potential bias,

the trials are randomly shuffled, mitigating order effects and ensuring better generalisation. Single-trial EEG epochs are often too noisy to effectively decode subtle stimulus classes. An increased signal-to-noise ratio can be achieved by randomly dividing the data into M sets of approximately a certain number of trials each, creating an averaged ERP for each set [257, 273]. Many previous studies [257, 273, 274] have used multiple sets of 10–20 trials and achieved the highest or near-highest performance in classifier accuracy. Therefore, averaging around 15 trials was utilised in this study. For example, if there were 80 high cloze trials and 80 low cloze trials for the content word class, we could average each set of 16 trials to create 5 ERPs for each condition. This approach also helps capture within-subject variability and assess decoding reliability.

The averaged ERPs are then subjected to stratified K -fold cross-validation [271], a technique that addresses the unbalanced distribution of classes and minimizes overfitting risks [257, 275], ensuring each class receives fair representation in both training and validation phases. Therefore, the SVM-based decoding is repeated K times. For each round, an SVM classifier with an RBF kernel is trained on $K-1$ folds and tested on the remaining fold. As in the main decoding procedure, this procedure was applied to each time point independently. Decoding accuracy is defined as the proportion of test cases that are correctly classified. In order to increase the resolution of the decoding accuracy, the entire process is iterated L times for each participant. For each iteration, we re-randomised the assignment of trials to averages and trained new SVMs.

If the neural pattern over the 32 electrodes contains information about lexical classes or predictability, then decoding accuracy should be greater than chance, which was 0.5 because we focus on binary classifications (e.g., content versus function words, high versus low probability). To compare decoding accuracy to chance at each time point while controlling for multiple comparisons, we used a cluster-based permutation technique similar to that in Section 5.2.2. We used one-sample t test to compare the mean accuracy across participants to chance. However, one-tailed tests

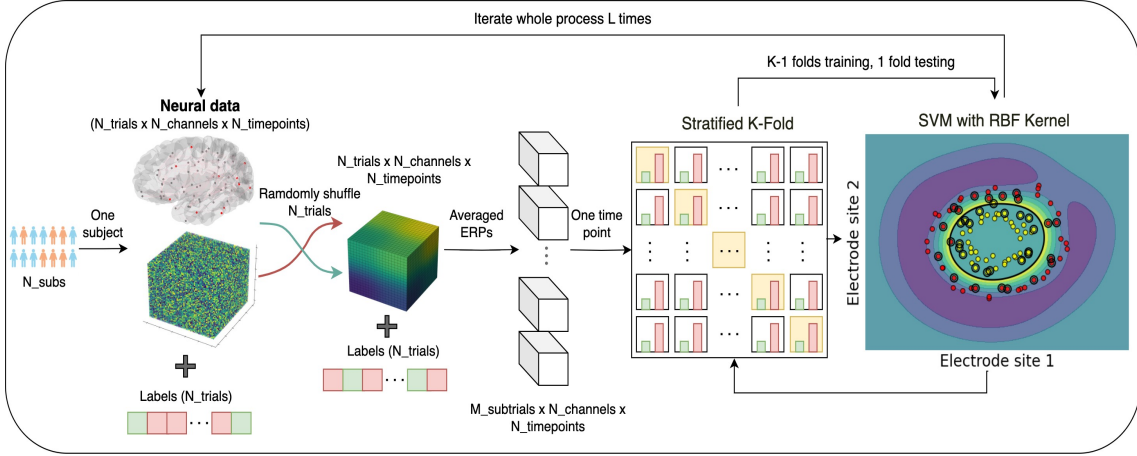


Figure 5.1: Flowchart for SVM-based decoding

were used for the SVM decoding accuracies in this case instead of two-tailed tests in the univariate technique because the SVM classifier should not produce meaningful below-chance decoding.

For a given score (i.e., a voltage in a univariate difference wave or a decoding accuracies), we selected a random sample of n of the N participants (sampling from the set of N participants with replacement) and computed the effect size for this random sample. We then conducted 5,000 iterations of the testing procedure, making it possible to construct the null distribution of the maximum cluster-level t_{max} with a cluster p-value. The summary of parameter setting for both traditional univariate analysis and decoding analysis is outlined in Table 5.2. Note that we only conducted univariate analysis to examine predictability effects across different word classes and lexical effects at various cloze probability levels, and then compared the results to the performance of SVM-based decoding. Regarding grammatical categories, the decoding analysis specifically aimed to explore how the recognition of grammatical categories were influenced by next-word predictability, which relates to top-down processing.

5.2.5 Comparison of Performance Between Methods

The SVM-based decoding provided decoding accuracies (proportion of correct classification), while the univariate analysis yielded p-values for wave differences (number

Group	N	No. of channels	No. of folds	Time window	L iteration	Baseline no. of trials	
Content	20	32	5	-100-500 ms	100	Condition 1	Condition 2
Function	20	32	5	-100-500 ms	100	212 (High cloze)	413 (Low cloze)
Noun	20	32	5	-100-500 ms	100	351 (High cloze)	200 (Low cloze)
Verb	20	32	5	-100-500 ms	100	135 (High cloze)	94 (Low cloze)
Pronoun	20	32	5	-100-500 ms	100	43 (High cloze)	192 (Low cloze)
Determiner	20	32	5	-100-500 ms	100	101 (High cloze)	48 (Low cloze)
Auxiliary	20	32	5	-100-500 ms	100	152 (High cloze)	39 (Low cloze)
Adposition	20	32	5	-100-500 ms	100	41 (High cloze)	34 (Low cloze)
High cloze	20	32	5	-100-500 ms	100	51 (High cloze)	64 (Low cloze)
Low cloze	20	32	5	-100-500 ms	100	212 (Content)	351 (Function)
						413 (Content)	200 (Function)

Table 5.2: Information about the SVM-based decoding procedure for each case. **N** refers to the number of participants. **No. of folds** indicates the number of folds used in Stratified K-fold cross-validation. **L iteration** represents the number of decoding iterations for each participant. **Baseline no. of trials** refers to the number of trials used for decoding in each participant. Note that a p-value threshold of 0.05 was used for the cluster-based permutation tests.

of microvolts). Both approaches demonstrated significant N400 amplitude differences between word classes, cloze probability levels, and grammatical categories. However, these measures could not quantify the magnitude of the conditions' effects for the N400. Moreover, a direct comparative assessment of their efficacy in distinguishing between ERP amplitude variations across conditions is not feasible. One of the most common solution is to use effect size or the power of a statistical test via Cohen's d_z [276]. Effect size measures the magnitude of the difference between outcomes. This approach allows researchers to present the magnitude of reported effects in a standardised metric, which can be understood regardless of the scale used to measure the dependent variable [277].

This metric quantifies the ability of each approach to produce statistically significant results while accounting for variation in wave difference (in univariate analysis) and decoding accuracy (in SVM-based decoding analysis) across participants as well as the mean. A larger effect size can reduce the Type II error rate, thereby increasing the proportion of significant effects that are actually true effects [278]. Thus, the approach that generated a larger Cohen's d_z exhibited greater statistical power for detecting differences between experimental conditions in a within-subjects analysis. The Cohen's d_z is calculated by subtracting the chance level from the mean across participants (after time-window averaging) and then dividing by the standard deviation across participants as in the following formula:

$$d_z = \frac{\bar{X} - \text{chance}}{\sigma_X}$$

Regarding the SVM-based decoding approach, X represents the mean decoding accuracy across participants, with chance level at 0.5 (binary classification), and σ_X denotes the standard deviation of these accuracy values. For the univariate analysis, X represents the mean wave difference between two conditions across participants, with chance level at 0, and σ_X denotes the standard deviation of these difference values. Note that the univariate analyses utilised ERP difference voltages averaged across the cluster of electrode sites where significant differences were observed in

the N400 time window (see Table 5.1), whereas the SVM analyses used data from all electrode sites. Bootstrapping (10,000 iterations) was then used to estimate the standard error of the effect size.

5.3 Results

5.3.1 High vs. Low Cloze Probability: Analysis Over Word Classes

Traditional Univariate Analysis

Several key points emerge from the analysis of Figure 5.2, particularly concerning the effects of cloze probability and the hemispheric distribution of N400 responses for content and function words. In general, the next-word prediction effect is most pronounced between 300-500 ms post-stimulus onset across the three regions (left, midline, and right hemisphere electrode sites) for different lexical classes, which is the typical time window associated with N400 activity.

The data shows that words with a higher cloze probability, which are more predictable within a given context, elicit a smaller N400 response (less negative amplitude) compared to words with lower cloze probabilities. This is consistent with the established understanding that the N400 amplitude reflects the ease with which a word can be integrated into its preceding context: more predictable words are processed more easily, resulting in a reduced N400 response [47, 52, 57].

In Figure 5.2 (a), the N400 effects are most pronounced larger over the midline electrodes region and then the right hemisphere electrodes compared to the left hemisphere electrodes, consistent with previous findings [36]. This hemispheric asymmetry is illustrated in the waveforms, particularly in the right hemisphere electrodes where the difference between high and low cloze conditions is most significant. Figure 5.2 (b) further support this observation, showing larger absolute differences in N400 amplitude in the right hemisphere electrodes, particularly for content words.

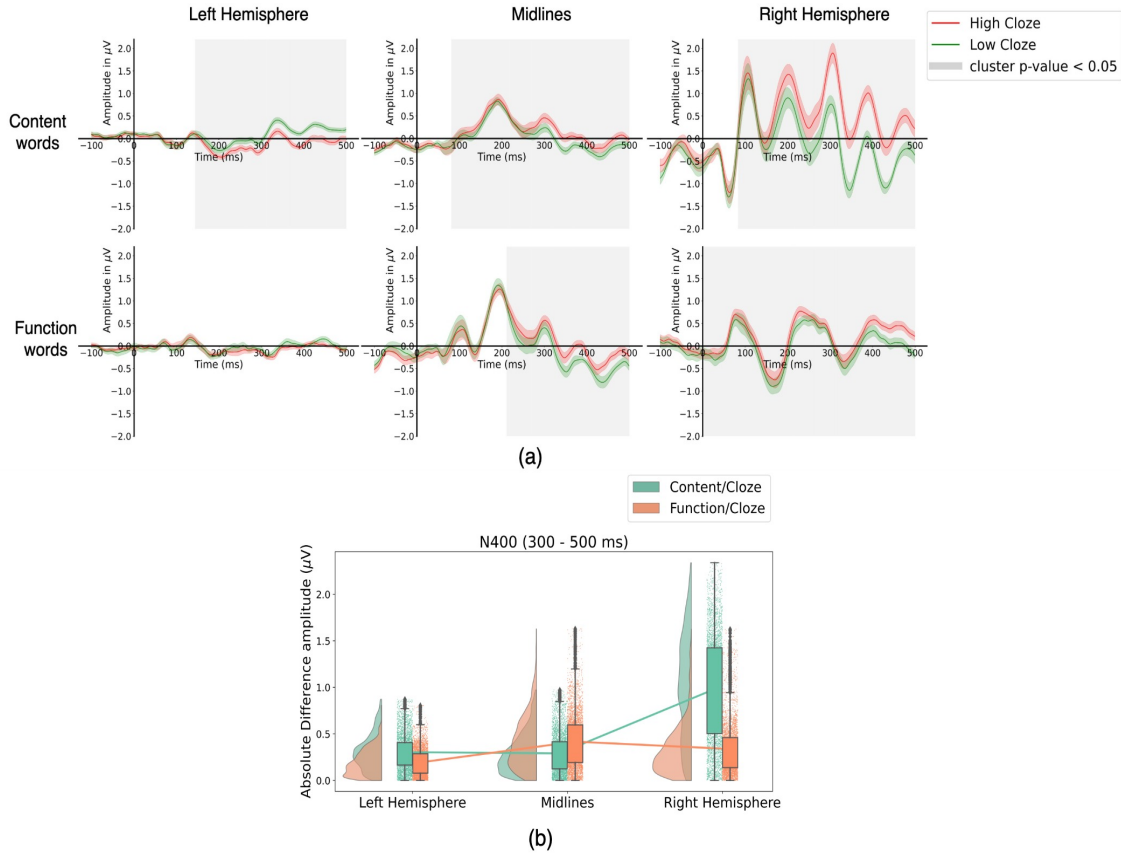


Figure 5.2: (a) ERPs plotted separately for content words (top row) and function words (bottom row) across the left hemisphere, midline, and right hemisphere electrode sites. ERPs are shown for high cloze (red) and low cloze (green) conditions. Gray shading indicates time points at which the values are significantly different from chance (p -values < 0.05) with a two-tailed test in the direction of the expected effect. (b) Raincloud plots displaying the absolute difference in N400 amplitude (measured from 300-500 ms) between content and function words across the left hemisphere, midline, and right hemisphere electrode regions. The plots combine boxplots, density distributions, and individual data points for both conditions. Note that for the content/cloze condition (green), ERPs to content words were averaged across cloze levels, while function/cloze condition (orange) represents function words across cloze levels. The lines connect the mean amplitude differences across these regions.

This right-lateralized effect suggests that the right hemisphere electrode region may play a more prominent role in the semantic processing for these two classes of words [57].

For function words, the N400 effect was widespread across the left hemisphere electrodes, but it did not vary significantly with cloze probability, indicating a more uniform processing that is less influenced by contextual predictability. This aligns with the notion that function words, which primarily serve syntactic rather than

semantic roles, elicit smaller and less variable N400 responses [34]. In contrast, content words exhibited a larger N400 effect in the left hemisphere electrodes, likely due to its role in leveraging semantic memory to preactivate the meaning of predictable words, leading to reduced N400 amplitudes for high cloze probability words. This distinction underscores the different neural processes engaged in the semantic processing of content words versus the syntactic processing of function words.

To conclude, the findings illustrated in the figure align with established theories in psycholinguistics and cognitive neuroscience. The data suggest that content words, which carry more semantic weight, show a stronger difference in N400 effect in response to predictability. Function words, in contrast, show a much smaller N400 effect, reflecting their reduced role in conveying semantic content.

Decoding Analysis

Figure 5.3 presents a detailed of SVM-based decoding accuracy over time for content and function words, measured across the left hemisphere, midline, and right hemisphere electrode sites. The analysis focuses on how accurately neural signals can differentiate between these conditions, particularly within the N400 time window. This decoding accuracy is significantly greater than chance over a broader range of time points compared to traditional univariate analyses of ERP difference waves. For instance, the decoding accuracy of content words (in Figure 5.3a) was slightly above the chance from around 100 ms to 200 post-stimulus (approximately 0.5 to 0.55), and increased significantly during the 300-500 ms window (reaching around 0.65), which corresponds to the N400 component. This broad range of significant decoding contrasts with the more limited time windows typically identified in univariate ERP analyses.

In Figure 5.3 (a), decoding accuracy for content words shows a noticeable increase over time, particularly around 300-500 ms, which corresponds to the N400 component known for reflecting semantic processing. The peak decoding accuracy reaches approximately 0.7 and 0.65 in the left hemisphere and midline electrode

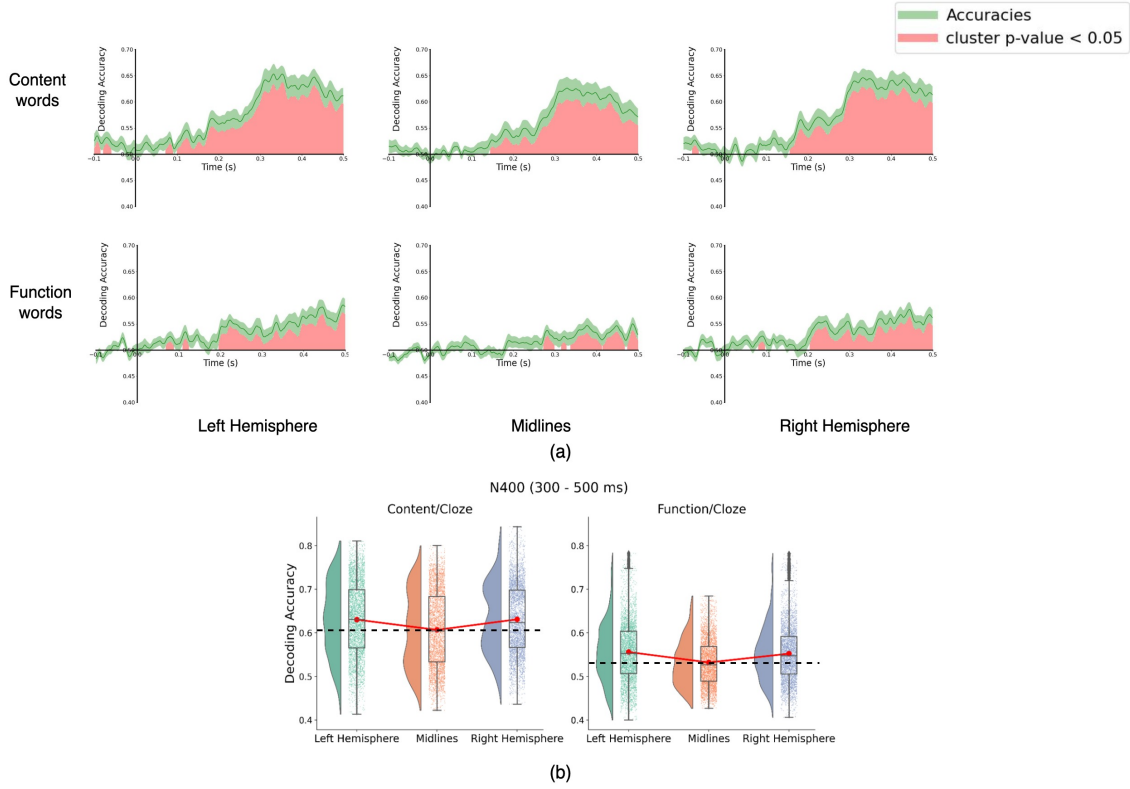


Figure 5.3: (a) Decoding accuracy for distinguishing between high and low cloze probability conditions plotted separately for content words (top row) and function words (bottom row) across the left hemisphere, midline, and right hemisphere electrode sites over time. Red shading indicates time points at which the value is significantly different from chance (p -values < 0.05) with a one-tailed test in the direction of the expected effect. (b) Raincloud plots showing decoding accuracy for content words (left) and function words (right) across the left hemisphere, midlines, and right hemisphere electrode regions during the N400 time window (300-500 ms). The plots include boxplots, density distributions, and individual data points. The black dashed line represents chance-level decoding (0.5), while the red lines indicate the mean decoding accuracy across regions.

sites, indicating that these regions contain substantial information about whether a content word is of high or low cloze probability. This suggests that as semantic processing unfolds, the brain's neural activity increasingly differentiates between more and less predictable content words, with statistically significant clusters ($p < 0.05$) marked by the red shaded areas. The right hemisphere electrode region also shows an increase in decoding accuracy, though slightly lower, peaking at around 0.6. In contrast, function words exhibit generally lower decoding accuracy across all three regions, peaking at around 0.55, and the time-course analysis shows fewer significant clusters, suggesting less differentiation between high and low cloze probability

function words in the neural signals. This indicates that function words, which are more syntactically focused and carry less semantic weight, generate less distinctive neural responses based on cloze probability.

Figure 5.3 (b) supports these findings by showing that during the N400 time window, the left hemisphere electrodes have the highest mean decoding accuracy for content words, emphasizing its critical role in processing semantically rich information. The accuracy slightly decreases in the midline and right hemisphere electrode regions, consistent with the left hemisphere’s dominance in language processing [32, 279]. For function words, however, decoding accuracy remains close to chance across all regions, reinforcing the idea that function words are processed more uniformly across the brain, reflecting their primarily syntactic rather than semantic function. The distribution analyses in this figure thus corroborate the trends observed in the time-course decoding accuracy, indicating distinct processing pathways for content and function words.

Both univariate ERP analysis and SVM-based decoding effectively detect neural differences between high and low cloze probability content words, but they offer different aspects. The univariate ERP analysis is particularly useful for highlighting the maximum amplitude difference of the N400 component, especially in the midline and right hemisphere electrode regions, where larger N400 effects are observed for less predictable content words. This suggests that these regions are more responsive to semantic incongruities. However, this traditional approach has limitations in discerning which word class—content or function words—shows a greater difference between high versus low cloze conditions. In contrast, SVM-based decoding offers a more dynamic and precise measure how the brain processes word predictability, particularly in the left hemisphere electrodes. While the univariate ERP analysis fails to capture the rich information about predictability that the left hemisphere electrode region, SVM-based decoding excels by showing that this electrode region contains more significant information for distinguishing between high and low cloze conditions.

In conclusion, both techniques confirm that content words generate more distinctive neural patterns compared to function words, which elicit weaker N400 effects and lower decoding accuracy. While the univariate analysis provides a static snapshot of ERP differences, SVM-based decoding adds a dynamic dimension by tracking how well neural signals classify word types over time, offering a more robust measurement of semantic predictability processing.

Combining these methods not only confirms the greater neural distinction for content words but also elucidates the temporal dynamics of these distinctions across different scalp regions, providing deeper insights into the interaction between top-down processes and lexical properties in language comprehension. These findings highlight the importance of using complementary approaches to fully capture the neural activity associated with word predictability.

5.3.2 High vs. Low Cloze Probability: Analysis Over Grammatical Categories

Content Class

Figure 5.4 presents an in-depth examination of how nouns and verbs are processed under varying levels of predictability (high and low cloze probability) across different electrode regions, as reflected in both ERPs and decoding accuracy. The analysis offers insights into the neural mechanisms underlying word recognition and integration, highlighting differences between these grammatical classes.

In Figure 5.4 (a), the grand average ERPs for nouns and verbs across left hemisphere, midline, and right hemisphere electrode sites reveal distinct patterns. Both nouns and verbs elicit N400 responses, with larger (more negative) amplitudes observed in low cloze conditions, reflecting the greater neural effort required to integrate less predictable words into the sentence context [52]. Verbs, especially under low cloze conditions, exhibit a more substantial N400 amplitude, particularly over midline and right hemisphere sites. This finding aligns with the semantic complexity theory, which suggests that verbs are more challenging to process due to their

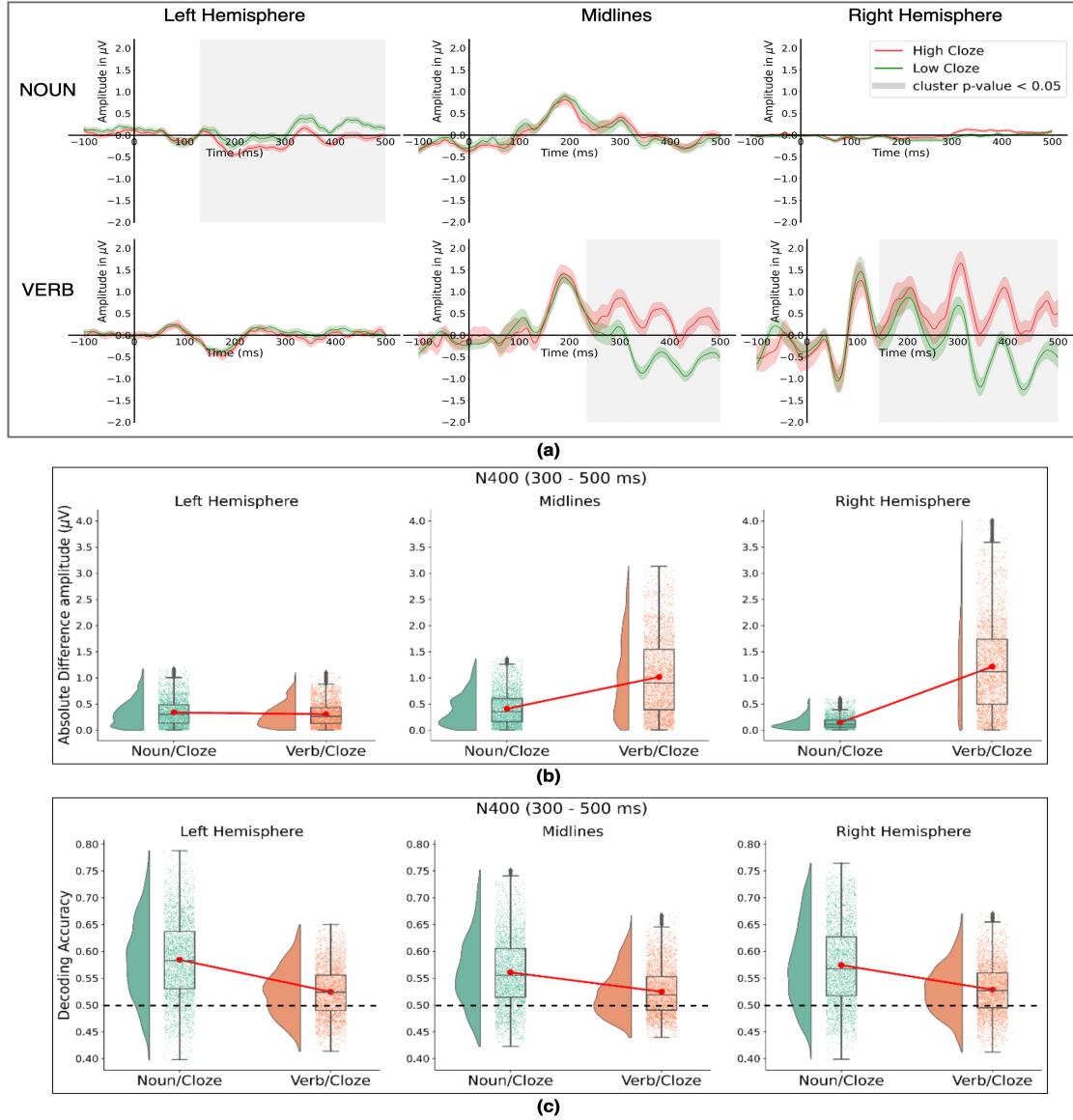


Figure 5.4: (a) Grand average ERPs plotted separately for nouns (top row) and verbs (bottom row) across the left hemisphere, midline, and right hemisphere electrode sites. ERPs are shown for high cloze (red) and low cloze (green) conditions. Gray shading indicates time points at which the values are significantly different from chance (p -values < 0.05) with a two-tailed test in the direction of the expected effect. (b) Raincloud plots displaying the absolute difference in N400 amplitude (measured from 300-500 ms) for nouns and verbs across the left hemisphere, midline, and right hemisphere electrodes. The plots combine boxplots, density distributions, and individual data points for both conditions. Note that for the Noun/Cloze condition (green), ERPs to nouns were averaged across cloze levels, while Verb/Cloze condition (orange) represents verbs across cloze levels. The lines connect the mean amplitude differences across these regions. (c) Raincloud plots showing decoding accuracy for distinguishing between high and low cloze conditions during the N400 time window (300-500 ms) for nouns and verbs across the left hemisphere, midline, and right hemisphere electrode regions. The plots include boxplots, density distributions, and individual data points. The black dashed line represents chance-level decoding (0.5), while the red lines indicate the mean decoding accuracy across regions.

lower imageability and fewer perceptual features compared to nouns [245, 280]. Notably, nouns show a stronger impact on predictability in the left hemisphere electrode region during the N400 time window, as evidenced by larger differences in N400 amplitude. However, Figure 5.4 (b) reveals that verbs show nearly equivalent predictability effects in the left hemisphere electrode region during the N400 time window. Indeed, verbs engage the left inferior frontal and middle temporal cortex, whereas nouns primarily involve temporal and parietal areas [281].

In contrast, the results in Figure 5.4 (c) indicate that nouns generally have higher decoding accuracy, especially in the left hemisphere electrode site, suggesting that the neural signals for nouns carry more distinct information about their predictability. This richer representation might be due to the generally more stable and predictable nature of nouns in language, making them easier to decode at the neural level, even when such distinctions are not evident in the traditional ERP analysis (as shown in Figure 5.4 (a)). Verbs, however, show lower overall decoding accuracy, reflecting more complex and distributed processing required for integrating verbs into the sentence structure [245, 280].

More importantly, this analysis provides distinct evidence when using traditional ERP analysis compared to decoding techniques. While the ERP differences in Figure 5.4 (b) initially suggest that verbs are more strongly affected by predictability than nouns, the decoding analysis in Figure 5.4 (c) reveals that nouns, despite showing less pronounced ERP effects, carry more distinct and accessible neural information compared to verbs [282]. This discrepancy arises because traditional ERP analyses quantify differences in specific components (e.g., the N400) across cloze conditions, assuming a consistent scalp distribution across participants. In contrast, decoding analyses measure the amount of information about a condition present in the recorded signal for each participant individually, without assuming that the underlying brain activity is the same across participants. This approach allows decoding to uncover more aspects that might be missed by applying traditional ERP methods.

Function Class

In Figure 5.5 (a) demonstrates that all categories within the function word class elicit a negative-going N400 component peaking between 280 and 400 ms, consistent with the findings of Osterhout et al. [36]. The N400 component reflects the brain's response to semantic processing challenges associated with function words, which, although less semantically rich than content words, still require integration into the sentence context. The N400 peak observed here is indicative of the cognitive effort involved in processing these grammatical categories within the function word class.

Figure 5.5 (b) illustrates the absolute differences in N400 amplitude between high and low cloze conditions across brain regions for each grammatical category. However, it is challenging to conclude which specific category dominates in N400 differences across the electrode regions. The variability observed suggests that no single grammatical category consistently exhibits larger N400 differences across all regions. For example, while pronouns show a significant N400 difference in amplitude in the left hemisphere electrode region, determiners exhibit more variability in the midline electrode region, and prepositions in the right hemisphere electrode region. This lack of a consistent pattern indicates that the neural processing of function words might depend on a complex interplay of factors, including sentence context and the specific grammatical role of each word category, rather than on semantic meaning when considered in isolation.

Figure 5.5 (c) indicates that all function word categories display lower decoding accuracy, suggesting that these words are more challenging to distinguish at the neural level. Nevertheless, we observe determiners' dominance across brain regions, likely due to their more abstract or context-dependent meanings. Indeed, ambiguous words can become minimally disambiguating in contexts when used with a determiner [283]. For example, consider the word "bowl," which can be ambiguous, referring to either a verb or a noun. However, when the determiner "the" is used as in "the bowl," it clarifies the meaning, indicating that "bowl" is being referred to as an object, not an action.

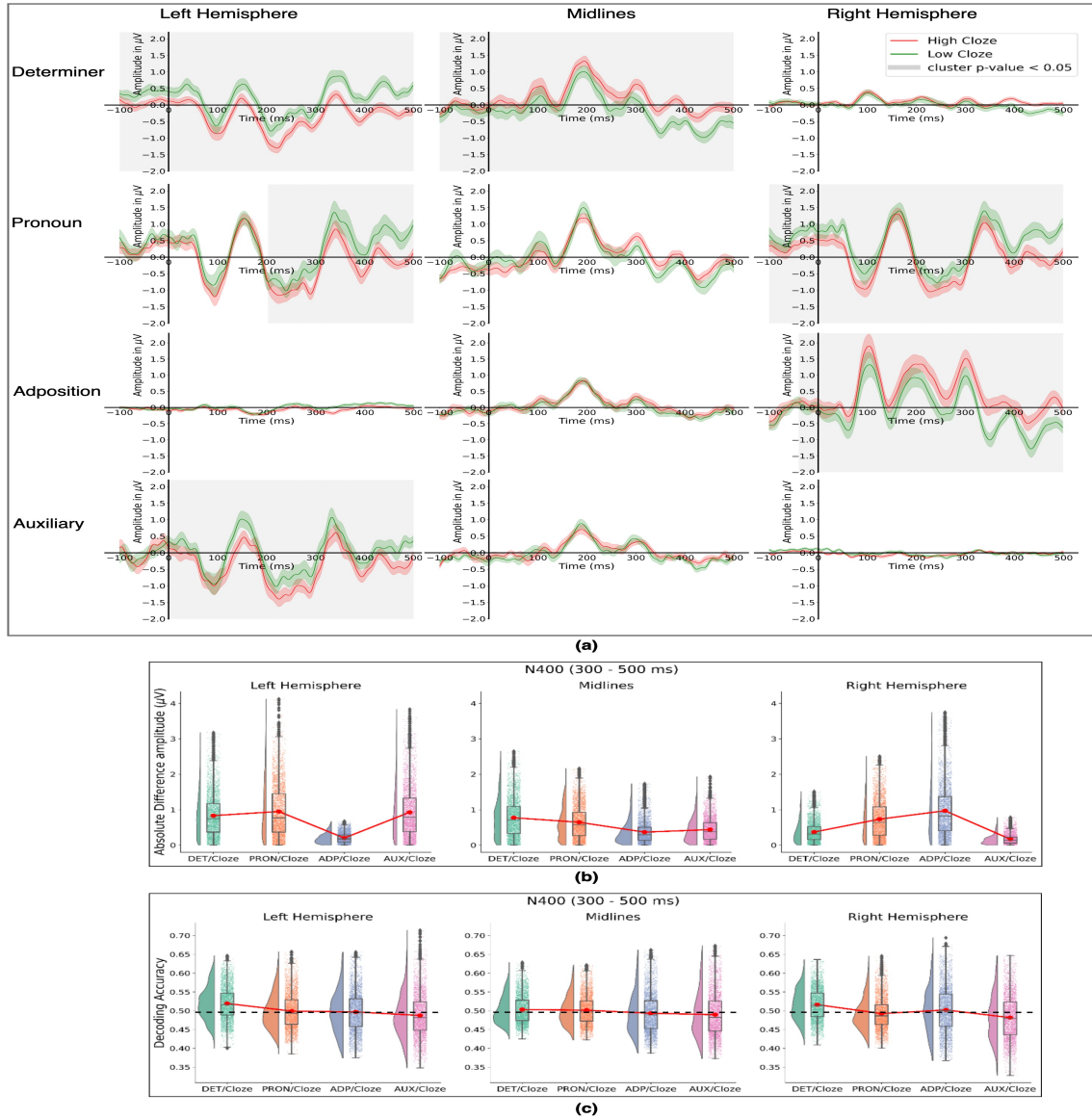


Figure 5.5: (a) Grand average ERPs plotted separately for determiners, pronouns, adpositions, and auxiliaries under high cloze (red) and low cloze (green) conditions across left hemisphere, midline, and right hemisphere electrode sites. Gray shading indicates time points at which the values are significantly different from chance (p -values < 0.05) with a two-tailed test in the direction of the expected effect. (b) Raincloud plots displaying the absolute difference in N400 amplitude (measured from 300-500 ms) for for each word class (determiner, pronoun, adposition, auxiliary) across the left hemisphere, midline, and right hemisphere electrodes. The plots combine boxplots, density distributions, and individual data points for both conditions. Note that for the Det/Cloze condition (green), ERPs to determiners were averaged across cloze levels, similar to Pron/Cloze (orange), Adp/Cloze (purple), and Aux/Cloze (pink). The lines connect the mean amplitude differences across these regions. (c) Raincloud plots showing decoding accuracy for distinguishing between high and low cloze conditions during the N400 time window (300-500 ms) for for each word class across the left hemisphere, midline, and right hemisphere electrodes. The plots include boxplots, density distributions, and individual data points. The black dashed line represents chance-level decoding (0.5), while the red lines indicate the mean decoding accuracy across regions.

To summarise, the analysis of function word categories reveals that all grammatical categories within this class elicit a significant N400 response, reflecting the cognitive effort required for their integration into sentence context. Additionally, while traditional ERP analysis does not show any clear dominance among grammatical categories in N400 amplitude differences across electrode regions, the SVM-based decoding approach provides a more detailed view. This advanced method uncovers determiners as the dominant category in neural processing, highlighting their role in disambiguating context-dependent meanings. These findings demonstrate the limitations of traditional ERP analysis and the added value of SVM-based decoding in identifying subtle distinctions in how the brain processes different categories of function words.

5.3.3 Results of Size Effects

Figure 5.6 provides a comparative analysis of the N400 time-window effect sizes between SVM decoding and traditional ERP analysis across ten distinct cases involving high versus low cloze probabilities for content words, function words, and their respective grammatical categories (nouns, verbs, determiners, pronouns, adpositions, and auxiliaries). In the standard univariate approach, the difference in ERP amplitude between the two conditions was quantified at single electrode site. In order to increase the size effect in the convention ERP analysis, we averaged the difference waves across an optimal cluster of electrode sites for each case.

Our findings indicate that in 6 out of the 10 cases, SVM-based decoding demonstrated superior sensitivity and effect sizes compared to conventional ERP analyses. In the remaining 4 cases, both methods exhibited approximately equivalent, with the decoding approach occasionally yielding slightly smaller effect sizes than the traditional ERP method. These results highlight the potential advantages of employing multivariate decoding techniques in neurocognitive language research.

SVM-based decoding demonstrated a much larger effect size compared to traditional ERP analysis, with the effect size for SVM being more than three times that

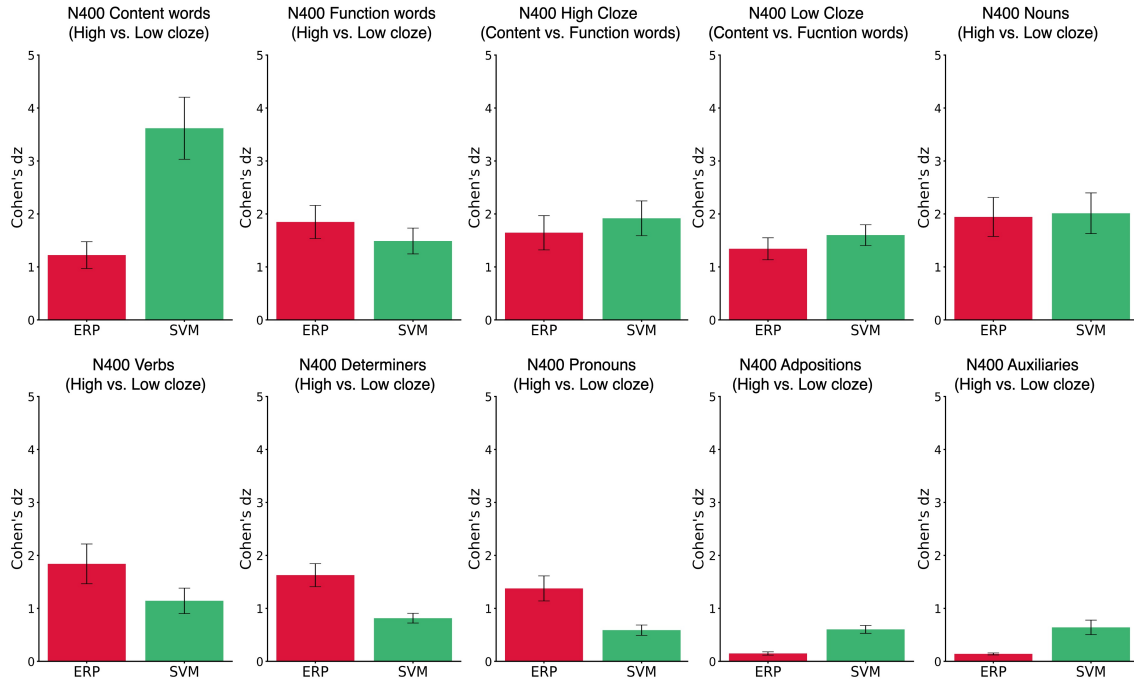


Figure 5.6: Effect size (Cohen's d_z) drawn on the same scale for each analysis in 300-500 ms time-window comparisons. The error bars show ± 1 standard error.

of the ERP effect size. The results indicate that SVM decoding is especially effective at distinguishing between high and low cloze probability in content words, potentially revealing subtle differences that may not be detected by traditional univariate ERP analysis.

In comparing content and function words under high cloze or low cloze conditions, SVM decoding similarly demonstrated outperformed result, with effect sizes exceeding those obtained from ERP analysis. However, both approaches were highly effective at detecting profound differences in neural processing between high predictable content and function words. This is presumably because of its ability to integrate information across multiple electrodes and time points, effectively capturing the semantic richness of content words, and the context-dependent nature of function words, which gain their significance largely from the surrounding context.

When distinguishing between high and low cloze probability nouns, SVM decoding exhibited a pronounced increase in effect size relative to ERP analysis. This indicated that SVM's effectiveness in capturing the dynamic neural responses involved in processing noun semantics under varying contextual expectations, offering

a more sensitive measure of semantic processing efficiency.

In the cases of verbs, determiners, and prepositions across high and low cloze conditions, ERP analysis slightly outperformed SVM decoding. Verbs, with their central role in sentence structure and meaning, typically evoke strong and distinct neural responses, which ERP captures more effectively. Similarly, determiners, due to their consistent usage and limited semantic content, result in uniform neural processing that ERP is better suited to detect. Pronoun also follow this pattern, with ERP analysis providing a more accurate representation of their predictable neural responses in language processing contexts compared to SVM decoding.

However, when analyzing adpositions and auxiliaries across high and low cloze conditions, the limitations of traditional ERP analysis become apparent. This approach relies on the assumption of a consistent scalp distribution of neural activity across participants, which can lead to smaller effect sizes. In contrast, SVM decoding does not make this assumption, instead quantifying the information present in each participant’s signal individually. This allows SVM decoding to capture more information about distinctions, resulting in significantly higher effect sizes for these grammatical categories. Additionally, the smaller sample sizes for adpositions and auxiliaries might have impacted the representativeness of the grand averaged ERPs across electrode sites, while SVM decoding has shown resilience and promise in performing well with limited data [284, 285].

Overall, our analysis demonstrates that SVM-based decoding techniques generally offer greater sensitivity and larger effect sizes compared to traditional univariate ERP analyses, particularly in detecting attenuated and distributed neural differences associated with semantic and grammatical processing. However, even in cases where both methods perform similarly, SVM decoding shows no disadvantage, proving its robustness across diverse linguistic contexts.

Moreover, traditional ERP analyses, which assume a relatively constant scalp distribution of neural activity across participants, can complement SVM-based decoding by providing consistent baseline measurements across conditions. In contrast,

SVM decoding offers an individual-centric approach, quantifying condition-specific information in each participant’s neural signals without relying on uniform brain activity patterns. Together, these methods can enhance our understanding of neural mechanisms by combining the strengths of consistent, broad-based ERP measurements with the detailed insights offered by multivariate decoding techniques. This integration enables a more comprehensive analysis of the neural mechanisms underlying language comprehension and cognition.

5.4 Discussion

This study provides substantial evidence for uncovering the neural mechanisms underlying next-word prediction in reading, particularly through the analysis of EEG signals and the N400 component. From this, we have addressed the second research question of the thesis *“Are neural responses to upcoming words attributable to quantitative differences in lexical categories (content versus function words) and grammatical categories (e.g., nouns, verbs, adjectives, pronouns, etc.)?”*. Additionally, by comparing traditional univariate ERP analyses with more advanced MVPA techniques such as SVMs, we have demonstrated that EEG decoding can reveal more detailed and robust patterns of brain activity associated with word predictability.

5.4.1 Key Findings and Interpretation

One of the primary findings of this study is the differential neural responses elicited by content and function words. Our results confirm that content words, which carry more semantic information, produce stronger N400 effects than function words, which are primarily involved in syntactic structuring. This aligns with previous research indicating that the brain’s semantic processing mechanisms are more active when dealing with content words due to their higher cognitive demands. Additionally, this finding also addresses the first main question: “Do differences in word

class explain the N400 differences observed at various cloze probability levels” as mentioned in Section 5.1.

Furthermore, the study reveals significant differences in how the brain processes various grammatical categories, such as nouns and verbs, in relation to word predictability. Verbs, which often have more abstract meanings and are less contextually grounded than nouns, elicited more substantial N400 amplitudes, especially in low cloze probability conditions. This suggests that verbs pose greater cognitive challenges for semantic integration, a finding consistent with theories that highlight the complexity of verb processing [247, 248]. Therefore, this result provides an empirical evidence for the second main research question: “To what extent does the recognition of grammatical categories drive top-down processing?”.

Importantly, our analysis of grammatical categories contributes to a relatively underexplored area of research. While much attention has been given to the distinction between content and function words, fewer studies have delved into the neural mechanisms differentiating specific grammatical categories, such as pronouns, determiners, or auxiliary verbs. This study, therefore, fills a critical gap by providing evidence that these grammatical categories also elicit distinct neural responses, reflecting their unique roles in language processing. The scarcity of research in this area underscores the need for further studies to better understand how these different categories are processed by the brain, particularly in relation to word predictability.

Moreover, the study reveals that predictability, as measured by cloze probability, significantly influences these neural responses. High cloze probability words, which are easier to predict within a given context, elicit smaller N400 amplitudes, reflecting a reduced cognitive load. This effect is particularly noticeable in content words, suggesting that the brain efficiently utilizes contextual information to anticipate and process semantically rich words. In contrast, function words show less variation in N400 amplitude based on predictability, likely due to their more stable and less context-dependent roles in sentence construction. As a result, the third main question, “Do highly predictable words generate greater N400 effects across

lexical classes compared to less predictable words?” has been addressed.

Last but not least, the comparison between traditional ERP analyses and SVM-based decoding further highlights the advantages of multivariate techniques in capturing the complexity of neural responses. While univariate approaches are useful for detecting broad patterns, they often overlook subtle but significant differences in brain activity across different conditions. The SVM-based approach, in contrast, demonstrated greater sensitivity and accuracy, provided detailed, individual-specific insights. This suggests that MVPA techniques provide a better understanding of the neural processes involved in language comprehension, making them a powerful tool for future research.

5.4.2 Theoretical Implications

These findings uncover several theoretical implications related to word classes, grammatical categories in reading, which, to our knowledge, have not been demonstrated using ML decoding techniques. First, our results support the interactive nature of top-down and bottom-up processes in reading comprehension. The ability of the brain to predict upcoming words based on context (a top-down process) is reflected in the N400 component, which also responds to the semantic congruity of words (a bottom-up process). This interplay suggests that reading comprehension is a dynamic process, where prediction and semantic integration occur simultaneously, each influencing the other.

Second, the differential effects observed between grammatical categories, such as nouns and verbs, provide evidence for the complexity of lexical processing. The fact that verbs, which often have more abstract meanings and are more context-dependent than nouns, show stronger neural differentiation based on predictability, highlights the cognitive challenges involved in processing different word types. This aligns with theories suggesting that verbs require more complex neural computations due to their syntactic flexibility and the variety of roles they play in sentence structure.

Moreover, our findings reveal the dominance of determiners over other grammatical categories in influencing N400 responses. This suggests that determiners play a crucial role in structuring sentence meaning and guiding the reader’s expectations, more so than other function words like pronouns, adpositions or auxiliaries. The significant neural response to determiners indicates that they are pivotal in helping the brain anticipate the semantic content of subsequent words, particularly nouns. Given this dominance, future research should investigate the N400 predictability effects on word combinations such as “Determiner + Noun.” Exploring these combinations could provide further insights into how these grammatical elements work together to facilitate language processing and comprehension, offering a more nuanced understanding of the interaction between top-down predictions and bottom-up integration during reading.

5.4.3 Limitations

Despite the strengths of this study, there are some limitations that should be acknowledged. One limitation is the relatively small sample size, particularly in the analysis of certain grammatical categories, which may affect the generalisability of the findings. Future research should aim to replicate these findings with larger and more diverse samples to ensure their robustness.

Additionally, while the study focused on the N400 component, other ERP components, such as the P600, which were proven to be indicative of syntactic processing [286, 287], could also be explored in future studies. Examining a broader range of ERP components could provide a more holistic understanding of the neural mechanisms involved in reading comprehension.

Finally, while this study employed SVM-based decoding, future research should consider using additional decoding techniques to compare their results and potentially gain a more comprehensive understanding of the neural data. For instance, Carrasco et al. [257] applied two MVPA approaches i.e. support vector machine decoding and the cross-validated Mahalanobis distance, which could provide a differ-

ent perspective on the data. Additionally, other machine learning classifiers such as LDA and RF could be explored to see how these techniques perform in comparison to SVM [258]. Future research should continue to refine these techniques, aiming to balance sensitivity with interpretability, and unfold how they can be integrated into broader models of cognitive processing.

5.4.4 Conclusion

In conclusion, this study advances our understanding of the neural processes involved in next-word prediction during reading. By leveraging both traditional and advanced analytical techniques, we provide new evidence for the complex interplay between predictability, word class, and grammatical category in shaping brain responses during language comprehension. These findings have important implications for cognitive neuroscience, highlighting the need for more sophisticated methods to fully capture the intricacies of human language processing.

Chapter 6

Encoding EEG Signals to Examine Human-Like Next-Word Prediction Behaviour in Language Models

While LMs are trained to excel at predicting the next word (or token) in a sequence given prior context, this represents only one aspect of intelligence. Although humans also share this predictability, our predictive capabilities are far more diverse, including reasoning, planning, imagination, working memory, factual recall, etc [288, 289]. Redesigning LMs to replicate the behaviour of the human brain is impractical due to our limited understanding of human cognition. Instead, we can focus on training LMs for next-word prediction tasks to be human-like by studying how humans perform this task and developing appropriate loss functions. Neuroscience research reveals that neural responses, such as ERPs, are critical indicators of brain activity during language processing. Particularly, lexical surprisal correlates robustly with brain responses, influencing ERPs after word onset. In this chapter, to address the research question RQ3 “*Which type of language model most closely aligns with human next-word predictability during reading comprehension?*”, we generate regressors based on two information measures, including top-1 prediction and surprisal, to predict ERP patterns in humans and LMs. Our results suggest that

while the more advanced LMs show a close correspondence to human performance in word prediction accuracy, only surprisal potentially correlates with language-processing ERPs, especially for open-class words with high semantic content. More importantly, our findings also challenge the assumption that increasing model size and training will consistently lead to improved convergence with human-like linguistic processing. This research sheds light on for improvements in LM training to be more human-like. LMs can be thought of as next-word predictive machines, but by taking advantage of neural responses with information theory alongside accuracy in training, may enable them to exhibit more human-like reading processing.

6.1 Introduction

In Chapter 2, we mentioned both surprisal estimated from LMs and human are able to predict semantic neural responses. According to Chapter 5, our findings demonstrated that highly predictable words (high cloze probability) generate more distinct N400 effects across lexical classes compared to less predictable words (high cloze probability). But how do these statistical language models perform relative to human cloze performance? In a next-word prediction paradigm, Goldstein et al. [76] identified a significant overlap between the predictions of various LMs, ranging from simple n-gram models to more sophisticated ones like GPT-2 with human prediction during podcast listening. Moreover, their findings indicate that GPT-2 embeddings capture information that is useful for modelling predictive brain activity in language comprehension.

Following the same general approach that examined surprisal as a predictor of N400 amplitude, Michaelov et al. [290] indicated that transformer-based LMs outperform other LMs with other architectures, such as RNNs, in predicting metrics related to human language comprehension. Subsequently, Michaelov et al. [291] found that the predictions of three advanced language models—GPT-3, RoBERTa, and ALBERT—aligned more closely with N400 responses than human predictions. This finding suggests that predictive mechanisms underlying the N400 might be

more closely tied to the statistical properties of language than previously believed.

Predicting neural responses is challenging, not only because it is technically difficult, but also because it raises important questions about the significance of such predictions. Our work systematically evaluates which language models most closely capture human cognitive processes in reading by comparing their performance in predicting EEG signals. Therefore, this chapter aims to address these concerns while addressing the final research question in the thesis: “Which type of language model most closely aligns with human next-word predictability during reading comprehension?” Ultimately, this contributes toward developing more human-like language models in the future.

Our investigation is organised as follows. We first calculate the lexical surprisal and top-1 prediction at the word level for content and function words across three predictors: human subjects, n-gram models, and GPT-family models (GPT2 and GPT-Neo), using the DERCO dataset [39]. Next, we encode neural responses by using regression-based deconvolution to estimate predictability effects on neural activity. Following this, we compare the correlations between neural response predictions derived from surprisal estimates of language models and those obtained from human cloze responses. The purpose of this comparison is to identify which model most closely mirror human-like predictability in reading behaviour. To provide deeper insights, these correlations will be visualised within significant time windows and across significant electrode clusters.

Through multiple experiments, we provide evidence to solve the research question via two primary findings. First, from a methodological standpoint, bigger is not always better. Although GPT-2 Large was trained on a smaller corpus and has fewer parameters than GPT-Neo 2.7B, surprisal estimates from GPT-2 Large show stronger correlations with neural prediction reconstruction ability, particularly for the N400 during reading. Second, while accuracy is a standard metric for evaluating language models (LMs), it is not a reliable indicator of human-like language processing. Instead, our findings highlight the value of information-theoretic surprisal

as a more meaningful metric, raising important questions about the future role of language models in psycholinguistic modelling.

6.2 Method

6.2.1 Stimuli and Data Preparation

EEG data

Again, all research questions are addressed in this thesis using the DERCo dataset [39]. In brief, EEG data were recorded from 22 native English speakers while they were reading The Grimm Brothers' Fairy Tales. Additionally, word-by-word cloze values for each word were collected using a cloze procedure via Mechanical Turk crowdsourcing platform. For more details on the EEG acquisition and data preprocessing, please refer to Chapter 3 and Chapter 4. Since the analysis used the preprocessed EEG data rather than the raw data, the number of EEG trials corresponding to each word in the transcript (see Appendix C.3) was reduced due to artifact removal. All remaining words after preprocessing served as stimuli for encoding brain signals. In Chapter 5, the Python library IPA was used to extract the parts of speech for all the words, which were then grouped into content and function word categories. Our results in Chapter 3 indicated that the significant differences in N400 responses between high and low cloze probability levels were more pronounced for content words than function words. Therefore, we will conduct our comparisons based on these word classes throughout the rest of this chapter.

Human Prediction

Regarding accuracy, top-1 prediction in this research refers to the highest percentage of participants who guessed the same next word. A top-1 prediction value of 100% indicates that all participants guessed the same next word, while a value of 0% indicates that no participant predicted that upcoming word. This can be simply defined as the maximum cloze probability among the possible words that could

appear in the upcoming position. Lexical surprisal, by contrast, is defined as the cloze probability of the **correct word** for the upcoming position in the transcript. It is important to note that the cloze value of the correct next word does not necessarily equal the top-1 prediction value. These values are equal only if the word is exceptionally easy to predict, meaning that the word predicted by all participants is also the correct word in the transcript.

N-gram models

In this chapter, we trained bigram to quadgram models using the NLTK Python package. N-gram models have a limitation in capturing very long-range dependencies in language, compared to more advanced language models such as transformer-based LMs [292, 293]. To compensate for this shortcoming, we used a corpus that has a similar domain as used in the DERCo dataset to provide relevant domain knowledge for our n-gram models. Thus, all n-gram models were trained on the Fairy Tale Corpus [294], which consists of 453 fairy tales extracted from Project Gutenberg [295]. However, to avoid bias, we removed the five stories from the Grimm Brothers’ Fairy Tales included in the transcript. All punctuation was removed, and the letters were converted to lowercase. To address data sparsity, we trained separate models using no smoothing, Laplace smoothing, and Kneser-Ney smoothing (see Appendix B for details). Each n-gram model then used a fixed-sized context window of $n - 1$ words, locating all matching windows within the problem instance, and counting the number of occurrences of each possible next token across those matches in order to calculate the probability of a word given.

GPT2 and GPT-Neo Family

In the transformer-based architecture, input tokens $U = (u_{i-k}, \dots, u_{i-1})$ are first mapped through a token embedding matrix W_e . After that, a position embedding

W_p is added corresponding to each word vector, resulting in the first hidden layer:

$$h_0 = UW_e + W_p. \quad (6.1)$$

Activities are then passed through a stack of transformer blocks, consisting of a multi-headed self-attention layer (as mentioned in Sub-section 2.3.2), a position-wise feedforward layer, and layer normalisation. This is repeated n times for each block b , after which (log) probabilities are obtained from a (log) softmax over the transposed token embedding of h_n :

$$h_b = \text{transformer block } (h_{b-1}) \quad \forall i \in [1, n] \quad (6.2)$$

In Equation 6.2, each transformer block takes the output from previous block h_{b-1} and produces a new hidden state h_b .

$$P(u_i|U) = \text{softmax}(h_n W_e^\top) \quad (6.3)$$

As in Equation 6.3, the final hidden state h_n is projected back to the vocabulary space by multiplying it with the transposed token embedding matrix $W_e^\top$. A softmax function is then applied to obtain the probability distribution $P(u_i|U)$ over the vocabulary for the next token u_i given the input sequence token U .

The input to the transformer model consists of a sequence of tokens, which can be words, phonemes, or punctuation. The number of tokens that can fit into a context window depends on the input story. If a story exceeds the length of one context window (e.g., 1024 tokens), the probability estimates for the remaining tokens are calculated by conditioning on the second half of the previous context window.

Predictions were performed on a run-by-run basis for each story in the transcript. This means that when computing the probability of a token in a new story, the entire preceding context was restarted, rather than carrying over the context from the previous story. More importantly, the topic of the story was included as part of the prediction, as it was disclosed to participants at the beginning of the Mechanical

Turk experiment (please refer to Sub-section 4.2.2 for further details.).

For words spanning multiple tokens, the word probability was calculated as the joint probability of the tokens using the chain rule. For surprisal estimation, we summed the surprisal values of the joint tokens, as this approach aligned with the online experiment presentation to participants, which included both the word and related punctuation. In contrast, for top-1 prediction estimation, since the focus was only on correct word prediction, participants were instructed to type only the word (excluding punctuation). Therefore, the probability used for calculating top-1 prediction was averaged over the joint tokens.

The model was implemented in PyTorch with the transformer modules via the HuggingFace Hub ¹. The examined GPT family variants differed primarily in their size, with the specific hyperparameters outlined in Table 2.1.

6.2.2 Top-1 Prediction Estimation

The objective for most LMs is to compute a probability distribution over the model’s vocabulary, V , for the likely next-word $w_i \in V$ at position i , given the context $w_1, w_2 \dots w_{i-1}$ consisting of the sequence of preceding words in the document. The highest probability, also known as top-1 prediction, to the next token is calculated as the following formula:

$$\text{Top-1 prediction } P_{w_i} = \max_{w_i \in V} P(w_i \mid w_1, w_2, \dots, w_{i-1}) \quad (6.4)$$

The main difficulty in comparing human and LM performance is that, unlike language models, it is infeasible to retrieve an entire probability distribution for the next words from the human mind during reading. Therefore, regarding human performance, the traditional approach to assess the contextual probability of a word w_i involves presenting participants with that word’s context and asking them to generate an appropriate continuation [37]. The calculation of cloze values is completely described in Chapter 2. In this case, the highest cloze probability will be

¹<https://huggingface.co/>

corresponding to human top-1 prediction.

A significant problem with cloze procedures is that they often produce zero probability estimates for some items. In reading experiments, there are cases where the target word w_i is not predicted by any participant as a possible continuation of $w_1, w_2, \dots, w_{i-1}$. With realistic sample sizes, words with $p(w_i|w_1, w_2, \dots, w_{i-1}) < 0.001$ will be completely absent from the participants' responses. Intuitively, this problem indicates the necessity of expanding the probability distribution over more words than those actually observed. Furthermore, zero-probability events are problematic for computing other information measures such as surprisal because the logarithm of zero is undefined. This limitation is challenging from both theoretical and methodological perspectives. In order to address the issue of zero probability estimates, we followed the approach used by Lowder et al. [296], which involved replacing these cloze values of zero with half the value of the lowest nonzero cloze value before converting them to surprisal values. For example, since the minimum non-zero cloze value in our study was .005, we converted all zero cloze values to .00225 before calculating surprisal values.

6.2.3 Surprisal Estimation

For more information about surprisal explanation and examples of surprisal, please refer to Subsection 2.4.1. Generating surprisal values for each word is straightforward because all three model families (n-grams, GPT-2, and GPT-Neo) directly output probability estimates for each word. An LM's surprisal estimates also serve to quantify how effectively that model has captured the statistical patterns of sentences. The more accurately a language models forms expectations about upcoming words, the higher probability it generally assigns to correct next words (low surprisal). In terms of human performance, the cloze probability of the correct word is utilised to calculate surprisal.

The comprehension process for a K -word document can be assumed to comprise a sequence of comprehension events for K words: $w_{1...k} = w_1, w_2, \dots, w_k$. After the

first t words of the sentence, $w_{1...t}$, will be processed, the identity of the upcoming word, w_{t+1} , is still unknown and can therefore be viewed as a random variable. The surprisal is defined as the negative log probability of the actual next word, given its preceding context:

$$\text{Surprisal } S_{w_{t+1}} = -\log P(w_{t+1}|w_{1...t}) \quad (6.5)$$

6.2.4 Brain Encoding Models

Figure 6.1 illustrates a method for predicting EEG data from measures derived from human and LMs, specifically from the GPT-family. The hypothesis is that with the same materials that human and LMs process in reading, they must share many of the same abilities, particularly in next-word predictability. Thus, to effectively investigate this hypothesis, we used two measures: Top-1 prediction and Surprisal as mentioned in Sub-sections 6.2.2 and 6.2.3. These measures serve as proxies for human and computational models' expectations and processing effort in language comprehension, capturing different aspects of cognitive load associated with word prediction. For more information about the cloze procedure and EEG data collection, please refer to Chapter 4.

To summarise, EEG recordings were conducted on a group of 20 participants, with signals collected from 32 electrodes distributed across the scalp. The words within each sentence were aligned with the recorded EEG signals, designating each word onset as time point 0 ms to standardise timing across trials. The two measures derived from the cloze task and LM predictions are utilised as regressors to model the EEG signal in a word-by-word, time-resolved manner. This approach, known as an encoding model, is designed to assess the extent to which each measure can explain variability in neural responses. At each time point within the EEG signal, separate regression models are fitted, predicting the amplitude of the EEG response based on these two linguistic measures. By applying this encoding model, we aim to capture the temporal dynamics of neural responses and to determine how predictive features such as semantic surprisal influence brain activity during reading processing.

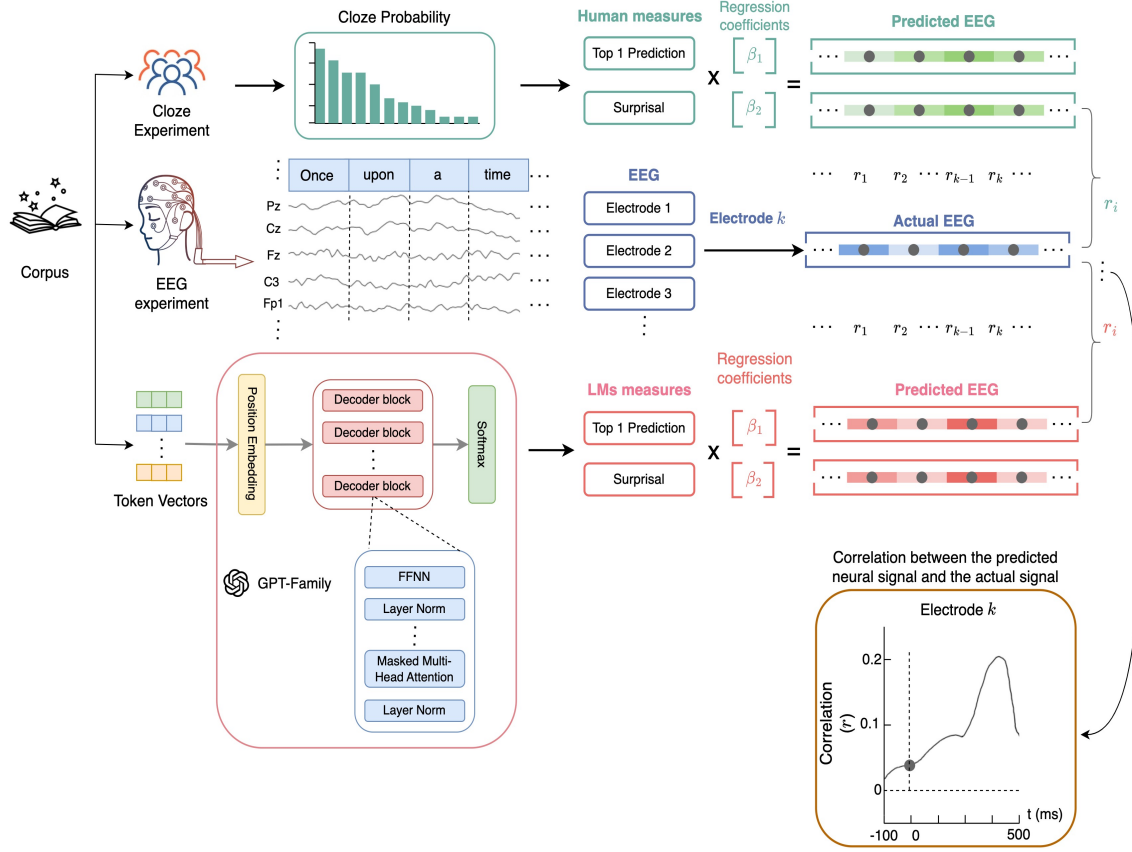


Figure 6.1: Linear encoding model used to predict the neural responses to each word in the narrative before and after word-onset.

The final step in the diagram involves evaluating the similarity between predicted and actual neural signals. For each electrode and time point, the predicted EEG signals (derived from the encoding model) and the actual recorded EEG signals are represented as vectors of voltage values, capturing both spatial (across electrodes) and temporal (across time points) dimensions. Pearson correlation coefficients (r) were then used to assess the quality of these predictions between predicted and observed ERPs.

In our study, after running regressions for each measure, we estimated a series of predicted EEG amplitudes from t_{min} (-100 ms) to t_{max} (500 ms), which are estimated for each subject per electrode. To quantify the similarity between the predicted and actual neural responses, we compute the Pearson correlation coefficient (r) for each word. This correlation-based evaluation produces a set of correlation values that reflect the alignment between model-predicted and observed EEG signals, thereby

validating the model’s effectiveness in capturing neural representations of language processing. For evaluation, we used a 5-fold cross-validation for each regressor. On each fold, the regression model used 80% of the data set as training trials to learn a set of 600 coefficients for each combination of time point and sensor. Subsequently, EEG activity (i.e., the pattern of signal amplitudes across sensors) was predicted for words not seen during training (i.e., the 20% of trials held out in the respective fold).

6.2.5 Regression Modelling

An approach commonly referred to as encoding models, as discussed in Section 2. This approach entails fitting a regression model where each time sample of the EEG is predicted by the semantic surprisal and top-1 prediction. However, running separate mixed effects models as in prior studies [89, 96, 297] for each time point would require estimating an excessive number of parameters. This would create a complex model that would be prone to significantly overfitting the data by capturing noise rather than real patterns. To address this issue, we adopt a regularised regression method using ridge regression, which adds constraints to the model to control its variability and make predictions more reliable [271].

Ridge regression was introduced by Hoerl and Kennard [298] as a generalisation of ordinary least-square regression. In comparison with the traditional least mean square regression, ridge regression provides better generalisation to unseen data by regularising the coefficient estimates, particularly in the presence of a large number of predictor variables. Ridge regression can be formulated as the following optimisation problem, solving for the regression coefficients b^* independently at each channel:

$$b^* = \arg \min_{b \in \mathbb{R}^p} (\|y - Xb\|_2^2 + \lambda \|b\|_2^2), \quad (6.6)$$

where $X \in \mathbb{R}^{n \times p}$ represents the matrix of stimulus features, where n corresponds to the number of time samples and p to the number of features. The target vector $y \in \mathbb{R}^n$ denotes EEG data, with n time points at single spatial location (electrode). The

$\|\cdot\|_2$ is the ℓ^2 norm of a vector which is applied to regularise the model’s coefficients, and the hyper-parameter λ controls the penalty weight in the loss function. Optimal λ is estimated from a set of candidates via cross-validation. If the value of λ is too low, the model may overfit to noise in the EEG data, while if λ is too high, the brain encoder model may underfit, reducing its ability to capture relevant brain response patterns [299].

In order to ensure each subject’s regression parameters equally, we determine λ by fitting a single regression model on the data from all participants for each value of λ over a log-spaced interval $[10^{-5}, 10^{-4}, \dots, 10^5]$. We select the λ value yielding the lowest generalisation error, as estimated through 5-Fold cross-validation (with shuffling enabled) over trials, using the Scikit-learn library implementation [300].

6.2.6 Significance Tests

All statistical analyses applied two-tailed tests with a significance level of $\alpha = 0.05$, corresponding to a 95% confidence interval. Before implementing any univariate tests (e.g., comparing model performance within and between participants), we first ensured that our data met two main assumptions: 1. Normal distribution, and 2. Absence of outliers. Normality was assessed using the D’Agostino and Pearson test as implemented in SciPy ², while outlier detection was based on the 1.5 IQR (interquartile range) method. When the data met both criteria, we applied one-sample t-test; otherwise, non parametric alternatives (e.g., Wilcoxon signed-rank tests) were used. This approach helps prevent misleading conclusions and statistical errors in the results.

In EEG research, the more statistical tests performed, the higher the probability of getting a false positive result due to random chance [301]. For the data in this study, the multiple comparisons problem is significantly more pronounced, with 32 EEG sensors and 600 time points resulting in 19,200 t-values per subject. To address this, we implemented cluster-based permutation tests [267] with 5,000 permutations

²<https://docs.scipy.org>

per test to identify significant time windows for the encoding models. While the clustered effects shown in Figures 6.3 and Figure 6.7 provide a partial view, we also reported a more comprehensive picture of the correlation across all channels (e.g., Figure 6.8 for GPT-2 Large). To visualise this, a mass-univariate testing approach was used, performing one-sample t-tests with a ‘hat’ variance adjustment method ($\sigma = 10^{-3}$), and correcting for multiple comparisons using the Bonferroni-Holm method for each time point and electrode to indicate the statistical consistency of the effects. The resulting t-statistics were subsequently converted into multiple-comparison corrected p-values. Additionally, multiple comparison correction for the explained variance across electrodes was conducted using Threshold-Free Cluster Enhancement (TFCE) [302].

6.3 Results

6.3.1 Performance and Human Alignment in Different Language Models

Accuracy Performance

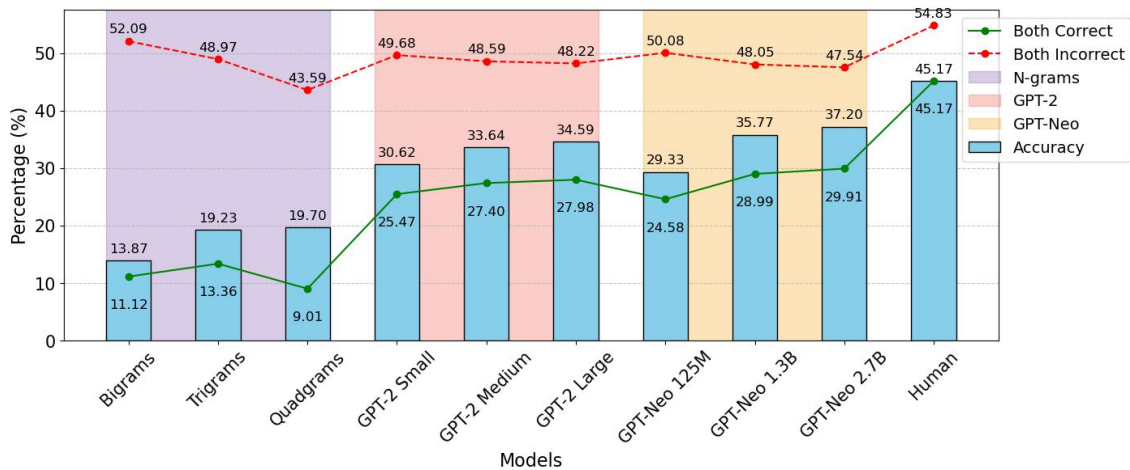


Figure 6.2: Performance comparison of various language model families (N-grams, GPT-2, GPT-Neo) and human benchmarks in a next-word prediction task, evaluated by per-model accuracy and joint prediction percentages “both correct” (green line) and “both incorrect” (red line).

Figure 6.2 highlights the performance of various LMs in the next-word prediction task, and revealing significant differences in accuracy and joint accuracy between LMs and human prediction. Among the n-gram models, Quadgrams achieved the highest accuracy (19.70%), slightly higher than Trigrams (19.23%). However, only 9.01% of the words were correctly predicted jointly by Quadgrams and humans, which is lower than Trigrams (13.36%) and even Bigrams (11.12%). This indicates the limitations of these models in capturing the complexity of natural language patterns.

The GPT-Neo models outperformed the GPT-2 models, with GPT-Neo 2.7B achieving the best accuracy (37.20%) and the highest percentage of joint correct predictions (29.91%) among the language models. Humans, however, still outperformed all LMs, achieving the highest accuracy (45.17%), which underscores the existing gap between language models and human predictive capabilities. The results suggest that LMs predict next words similarly to human prediction in narrative contexts, with their performance gradually approaching that of humans; larger LMs are more accurate than smaller ones, but they have not surpassed human accuracy.

Correlation Performance

Several methods were developed for quantifying relationship between time series, particular in EEG data analysis [303, 304]. Two most widely used correlation techniques are Pearson product-moment correlation, which measures linear relationships, and the Spearman rank-order correlation, which evaluates monotonic relationships [304] (see Appendix A for detailed explanations and formulas).

The similarity of neural patterns between pairs of words was quantified using Pearson correlation in previous studies [76, 208], as it effectively captures the similarity of two neural patterns regardless of the amplitude of the neural response. To strengthen our analysis, in addition to examining Pearson correlation, we also investigated Spearman correlation to provide more robust evidence for the observed relationships.

Model Variants	Top-1 Prediction (vs. Human)		Surprisal (vs. Human)		Top-1 Prediction (vs. Human)		Surprisal (vs. Human)	
	Pearson r	p	Pearson r	p	Spearman r	p	Spearman r	p
Bigrams (KN)	-0.02	0.31	0.08	$< p^*$	0.08	$< p^*$	0.18	$< p^*$
Trigrams (KN)	0.13	$< p^*$	0.03	0.17	0.11	$< p^*$	0.05	0.01
Quadgrams (KN)	0.04	0.02	-0.01	0.59	0.02	0.30	-0.02	0.36
GPT-2 Small	0.52	$< p^*$	0.72	$< p^*$	0.51	$< p^*$	0.78	$< p^*$
GPT-2 Medium	0.56	$< p^*$	0.75	$< p^*$	0.55	$< p^*$	0.80	$< p^*$
GPT-2 Large	0.59	$< p^*$	0.76	$< p^*$	0.57	$< p^*$	0.81	$< p^*$
GPT-Neo 125M	0.52	$< p^*$	0.73	$< p^*$	0.51	$< p^*$	0.78	$< p^*$
GPT-Neo 1.3B	0.59	$< p^*$	0.76	$< p^*$	0.58	$< p^*$	0.82	$< p^*$
GPT-Neo 2.7B	0.61	$< p^*$	0.76	$< p^*$	0.59	$< p^*$	0.82	$< p^*$
Human	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00

Table 6.1: Correlation comparisons of various language model families (N-grams, GPT-2, GPT-Neo) against human benchmarks in a next-word prediction task. Metrics include Pearson and Spearman correlation coefficients r for top-1 prediction and surprisal. Statistically significant correlations with $p^* = 0.001$ are indicated in two-sided test.

Table 6.1 illustrates the correlation between these LMs and human predictions in terms of top-1 prediction and surprisal. The n-gram models (bigram, trigram, and quadgram) demonstrate poor performance compared to human benchmarks. Regarding top-1 prediction correlation with human predictions, the trigram model demonstrates a statistically significant but weak correlation, with a Pearson value of approximately 0.13 and a Spearman value of nearly 0.11, while the bigram and quadgram models exhibit negligible correlations. Similarly, surprisal values across all n-gram models show weak correlations with human judgments. Only the bigram model shows a significant correlation for surprisal measures ($p < 0.001$), with a Pearson correlation of around 0.08 and a Spearman correlation of about 0.18. Overall, the correlations between n-gram models and human predictions are generally weak and statistically non-significant. These results indicate that n-gram models, while historically important, struggle to capture the complexity of human-like next-word prediction.

GPT-2 and GPT-Neo models demonstrate robust correlations with human performance across both top-1 prediction and surprisal metrics, consistently improving as model size increases. Importantly, the consistent increasing patterns observed in both Pearson r and Spearman r values, regardless of assumptions about the data,

indicate a robust relationship between the variables. This alignment suggests that the correlation is both monotonic and, to some extent, linear, demonstrating the reliability of the observed relationship which is not influenced by outliers or deviations from linearity.

In the GPT-2 family, top-1 prediction correlations range from a Pearson correlation of approximately 0.52 and a Spearman value of about 0.51 for GPT-2 Small to a Pearson value of around 0.59 and a Spearman correlation of nearly 0.57 for GPT-2 Large. Surprisal correlations are moderately higher, ranging from 0.72 to 0.76 in Pearson correlation and from 0.78 to 0.81 in Spearman correlation. Similarly, GPT-Neo models show increasing top-1 prediction correlations, rising from approximately 0.52 to 0.61 in Pearson correlation and from about 0.51 to 0.59 in Spearman correlation. Surprisal correlations also improve, reaching a Pearson correlation of about 0.76 and a Spearman correlation of approximately 0.82 for the GPT-Neo 2.7B model.

Notably, all these correlations are statistically significant at the $p < 0.001$ level, suggesting a robust correlation between these transformer-based LMs and human prediction. Moreover, the results indicate that larger models are increasingly capable of capturing complex linguistic structures through predicting next words in a manner more consistent with human predictions. This is reflected in the gradual improvement of both predictive accuracy and joint performance metrics.

While state-of-the-art language models show significant improvements, particularly as model size increases, the underlying mechanisms driving their correlation with human behaviour remain unclear. The question arises whether these LMs truly reflect human reading processes in the human brain or just exhibit surface-level convergence with human performance in next-word prediction paradigms. To explore this further, we delve into the analysis and comparison of results from brain encoding models, using information measures computed by these language models as described in the following sections.

6.3.2 Neural Encoding Using Predictive Metrics from Human Cloze

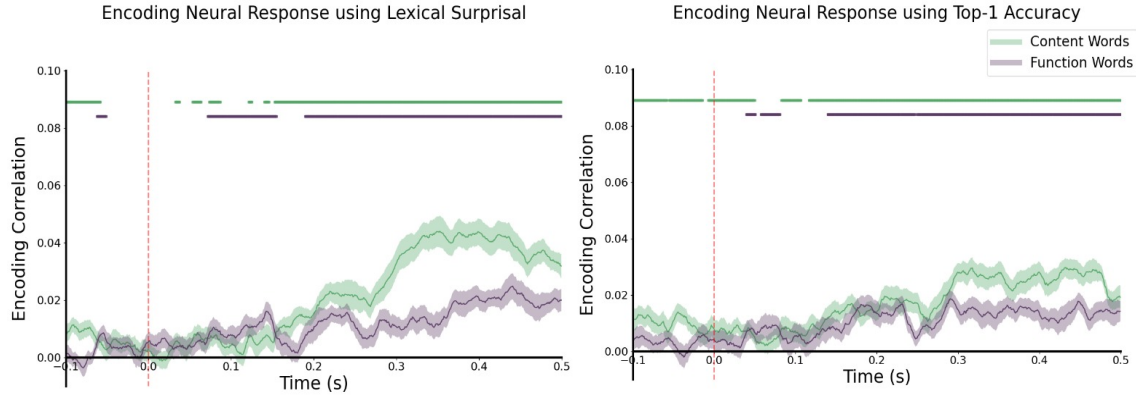


Figure 6.3: Correlations between regression model results estimated from human cloze probabilities and neural responses for content and function words across time, shown for both lexical surprisal (left) and top-1 prediction (right) metrics. The encoding analysis was conducted for each electrode with significant encoding of predictive metrics and then averaged across electrodes. Asterisks indicate window time at which the value is significantly different ($p < 0.001$) with cluster-based permutation tests in the direction of the expected effect for content words (green line) and function words (purple line). The shaded regions represent the between-subject SEM (standard error of the mean) of the encoding models.

Figure 6.3 illustrates that lexical surprisal derived from human cloze probabilities serves as a more robust predictor of neural responses, particularly in the N400 time window (300-500 ms), compared to top-1 prediction measurements. These results provide strong evidence supporting our findings in Chapter 5, in line with the well-established influence of semantic effects on the N400 (see Chapter 2 for more references). Additionally, the encoding correlations are stronger for content words relative to function words, suggesting that lexical surprisal more effectively captures neural processes associated with semantically rich lexical words. These findings further validate the utility of surprisal metrics in modelling neural responses during language comprehension.

Therefore, we take the human cloze as a baseline for quantifying the linguistic predictive abilities of different LMs, and surprisal as the most explanatory representation of predictive metrics for the rest of our analyses.

6.3.3 Surprisal Estimate for All Encoding Models

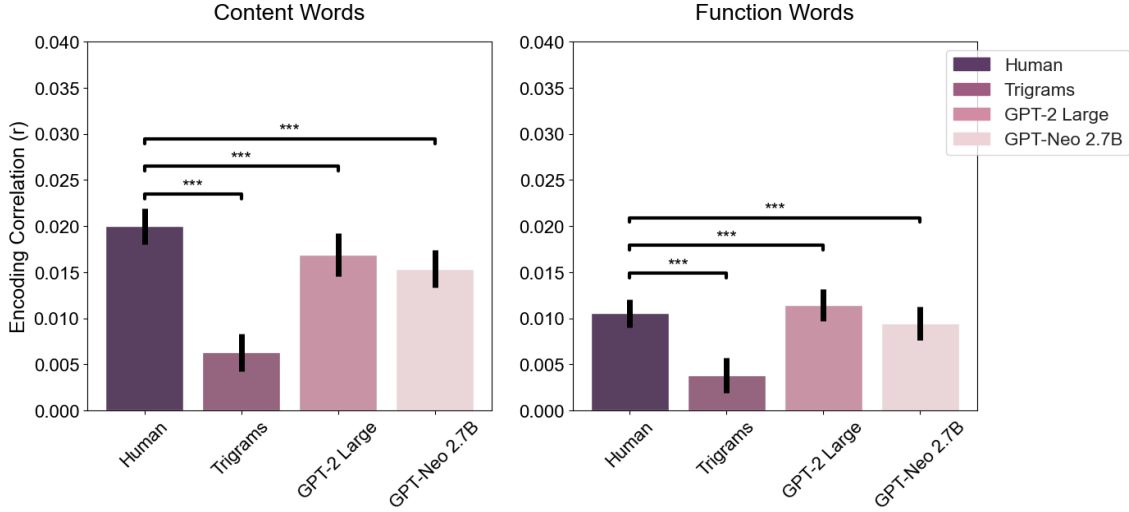


Figure 6.4: EEG encoding model comparison results using surprisal for content words (left) and function words (right) averaged across all channels over time (from -100 to 500 ms) for all participants. Significance levels indicate statistical differences between models, with $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***) based on two-tailed paired t-tests or Wilcoxon signed-rank tests. Error bars represent the SEM.

We selected Trigrams, GPT-2 Large and GPT-Neo 2.7B as the representatives for each language model family due to their best performance within their respective groups, as shown in Section 6.3.1. Figure 6.4 illustrates EEG encoding model comparison results using surprisal for content and function words across all channels and subjects within the -100 to 500 ms time window, with “Human” as the control model.

Model Comparison Averaged Across Subjects

Across all channels and subjects, among computational models, transformer-based LMs show significant performances over the traditional statistical language model i.e. trigram. Particularly, the numerically best performing language model overall is GPT-2 Large, which significantly outperforms all other LMs in both content and function classes. We thus find a strong evidence that surprisal-based regression estimated from the larger GPT-Neo 2.7B model do not provide stronger prediction correlations than those from the LM with a smaller in size and lower prediction accuracy i.e. GPT-2 Large variant, consistent with much prior work [79].

In terms of lexical classes, while language models exhibit relatively lower correlation values for function words compared to the content word category (less than approximately 50%), the overall patterns in predictive performance remain broadly similar. This observation clearly indicates that the top-down expectation effect, as captured by the surprisal metric, is not limited to only highly semantic contexts. Rather, it applies much more generally, in line with the notion of continuous prediction underlying reading processing.

Single-Subject Model Comparison

Since Figure 6.4 only shows averaged prediction correlations across all subjects, we further examine individual subjects to gain deeper insights. In the content group, as shown in Figure 6.5, the GPT-2 Large model demonstrates the highest neural encoding performance in 11 out of the 20 participants (ACB71, JPY86, LYY64, MNY88, MRB47, NXB64, QPI83, SAB93, UJM92, WHR08, and WJX11). The results of these participants are statistically significant, indicating a strong alignment between the GPT-2 Large model and human predictive modelling. This performance surpasses other models, such as trigrams and GPT-Neo 2.7B, whose results occasionally achieve the highest correlations but sometimes fail to significantly correlate with human predictive modelling. For instance, while the trigram model shows the highest encoding performance for subject RRO98, its results do not correlate significantly with the control model.

In the function group (as depicted in Figure 6.6), 11 out of 20 participants (ACB71, HMK96, JPY86, LRK01, MNY88, MRB47, OLW10, QFT39, RRO98, UJM92, and WAL74) exhibit the highest predictive correlation scores for GPT-2 Large compared to other models. However, the correlation between GPT-2 Large and the control model is less stable in this group compared to the content group. For example, while GPT-2 Large achieves the highest prediction scores for subjects MNY88, MRB47, and OLW10, these results may not reflect a significant relationship with human predictive modeling, as indicated by p-values greater than 0.05. This



Figure 6.5: Surprisal-based encoding model comparison results for **content words** across all channel for individual participants. Significance levels indicate statistical differences between models, with $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***) based on two-tailed paired t-tests or Wilcoxon signed-rank tests. Error bars represent the SEM.

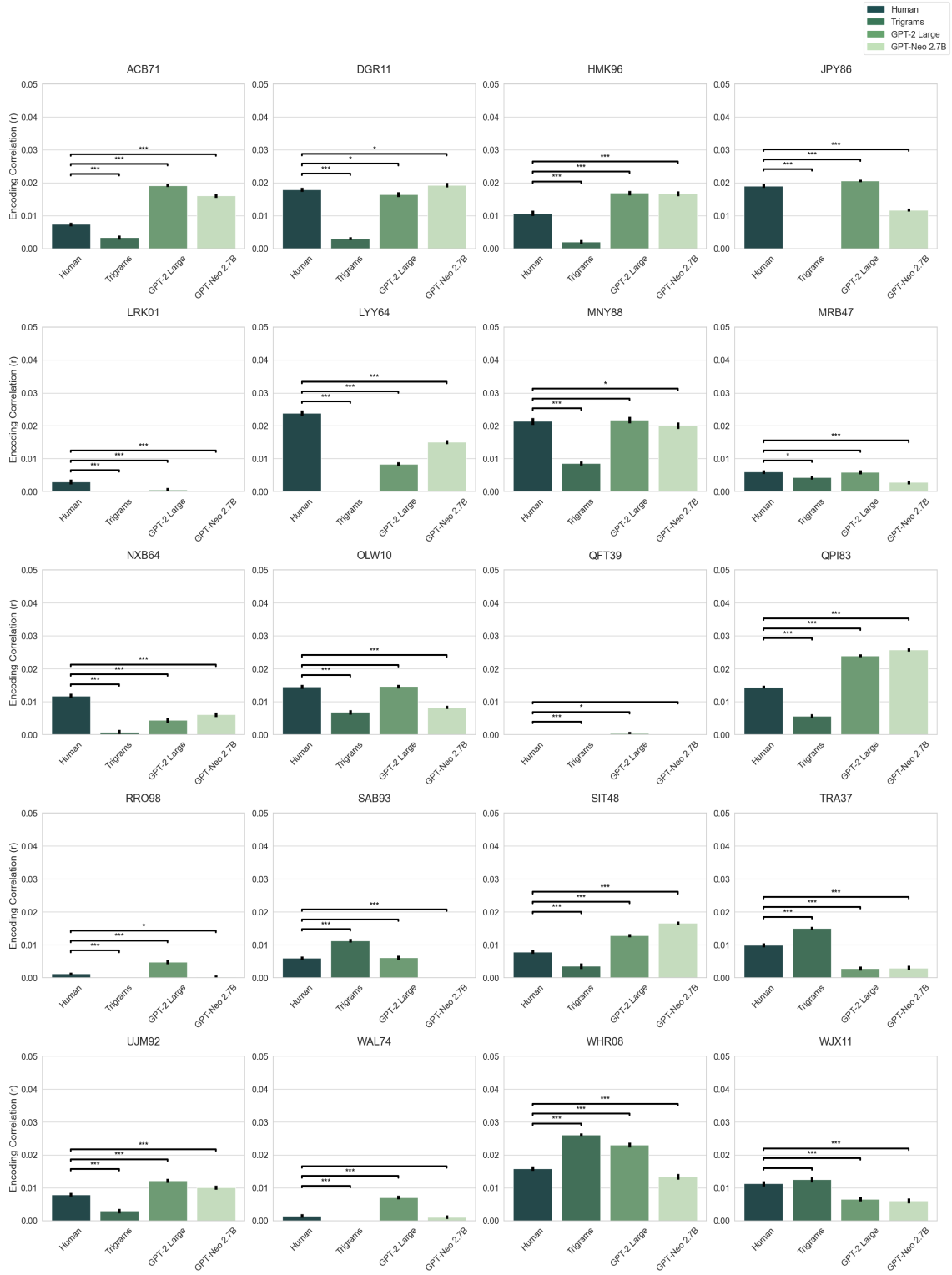


Figure 6.6: Surprisal-based encoding model comparison results for **function words** across all channel for individual participants. Significance levels indicate statistical differences between models, with $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***) based on two-tailed paired t-tests or Wilcoxon signed-rank tests. Error bars represent the SEM.

unstable correlation could be caused from the differences in processing function words by human and GPT-2 Large. Indeed, humans depend on grammatical rules and habitual patterns to predict the function words, whereas GPT-2 Large makes its predictions by using statistical probabilities based on large datasets.

6.3.4 Neural Correlates of Surprisal in GPT-2 Large Regression

Grand-Averaged Neural Encoding Correlations for Surprisal-Based Regressions

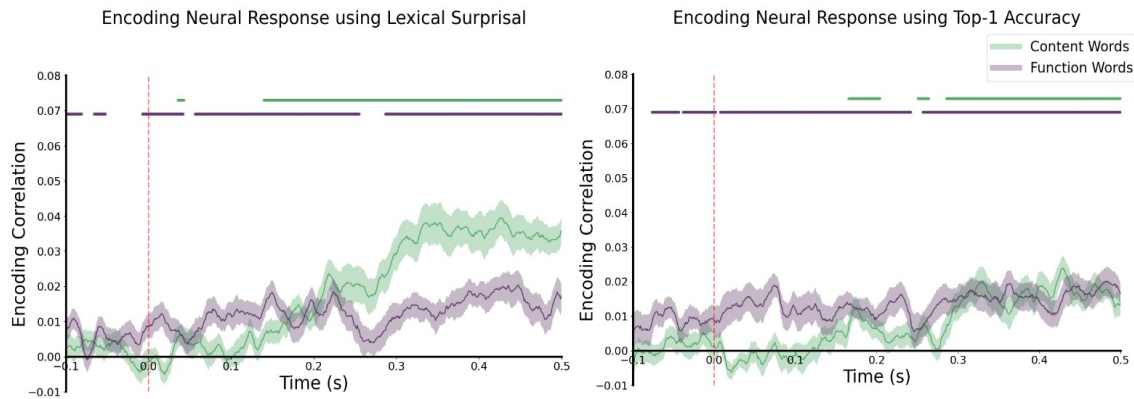


Figure 6.7: Correlations between regression model results estimated from GPT-2 Large model and neural responses for content and function words across time, shown for both lexical surprisal (left) and top-1 prediction (right) metrics. The encoding analysis was conducted for each electrode with significant encoding of predictive metrics and then averaged across electrodes. Asterisks indicate window time at which the value is significantly different ($p < 0.001$) with cluster-based permutation tests in the direction of the expected effect for content words (green line) and function words (purple line). The shaded regions represent the between-subject SEM (standard error of the mean) of the encoding models.

Figure 6.7 illustrates the comparison of word-by-word regressors using both surprisal-based and top-1-accuracy measurements built from GPT-2 Large against EEG signals in terms of content and function groups. As can be seen, content words show stronger neural correlations that last longer, particularly during the N400 time window. Function words, while still showing significant correlations, demonstrate weaker encoding correlation during this same period. The lexical surprisal metric

is especially effective at highlighting these differences between word types, suggesting it better captures how the brain processes the meaning and context of content words.

However, although top-1 prediction presents a similar trend between content and function words, the difference becomes less apparent. As a result, even though these results come from GPT-2 Large model, the patterns mirror those seen in human cloze, as shown in Figure 6.3.

While Figure 6.7 only shows the encoding correlation scores averaged over channels participating in the cluster (thereby only visualising the effect), these topographies, as illustrated in Figure 6.8, visualise the results comprehensively across all channels, over time.

Topographic Maps of Grand-Averaged Neural Encoding Correlations

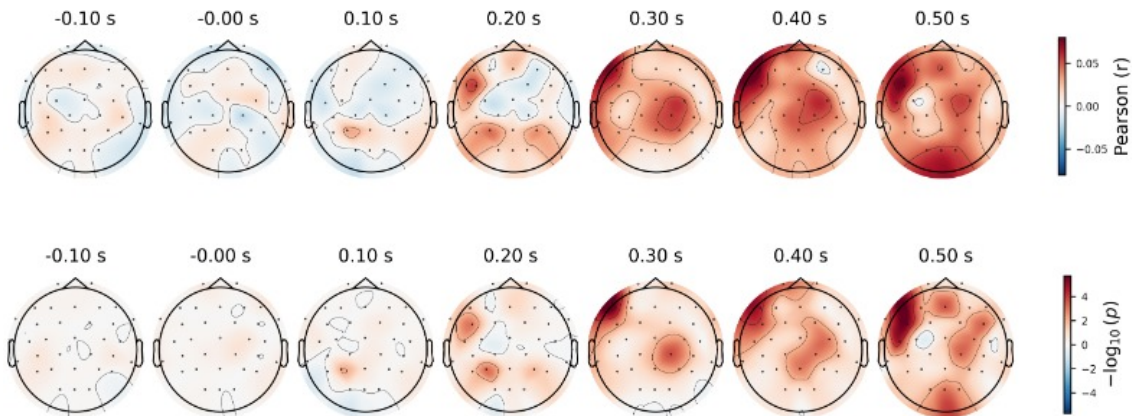


Figure 6.8: Full EEG Topographies of the effects of lexical surprisal. The topographies show the Pearson r of encoding prediction correlations (upper row) and the respective TFCE-corrected significance $-\log_{10}(p)$ (lower row).

As shown in Figure 6.8, the full topographies across all 32 scalp channels revealed distinct spatial patterns at different time lags. During the early time windows (100-200 ms), no significant spatial pattern was discernible. However, a clear and structured spatial pattern emerged at 300, 400, and 500 ms, with particularly strong weightings observed in the 300-500 ms time window. We take this highly structured pattern of encoded correlations as strong evidence supporting our claim

that the neural response, particularly the N400 component, is best explained by a probabilistic model of prediction, as estimated by the surprisal measure.

In summary, our analysis of encoding lexical surprisal offers multiple pieces of evidence suggesting that transformer-based language models can serve as a theoretical framework for understanding language comprehension processes. Moreover, we also provide evidence that surprisal-based GPT-2 Large can effectively track human neural signals, especially in predicting the next content words.

6.4 Discussion

The overarching research question, “*Which type of language model most closely aligns with human next-word predictability during reading comprehension?*” is addressed by our findings (as mentioned in Sub-section 6.4.1). Surprisal-based measure, rather than prediction accuracy, provides a more effective approach for understanding the neural correlates of language processing, offering deeper insights into how LMs can better approximate human predictive behaviour in naturalistic reading contexts.

Although language plays a central organising role in human cognition, linguistic competence is insufficient to fully capture the complexity of thinking. Unlike humans, LMs cannot think, understand, or generate long-term scenarios by integrating prior knowledge. Instead, they simply reflect statistical patterns from their training data. This research opens avenues for future research at the intersection of cognitive neuroscience and machine learning is how LMs might leverage different forms of linguistic prediction (e.g., syntactic versus semantic), like those learned by human for generating and articulating thoughts.

6.4.1 Key Findings and Interpretation

This chapter investigated the correlation between LM predictions and neural responses during reading comprehension estimated by surprisal and top-1 prediction

measurements and their alignment with human next-word predictability. The findings indicated that surprisal estimates, derived from various transformer-based language models, serve as significant predictors of neural responses, particularly in the N400 component. This result is consistent with findings from prior studies, which have linked the N400 to semantic processing and predictability [88, 89, 99, 291]. Additionally, surprisal measures demonstrated stronger predictive power for neural responses compared to top-1 prediction, especially for open-class content words, even though both metrics reflect next-word predictability.

Our work extended and clarified the discourse of neural processing difference in how content and function words during reading comprehension. Surprisal-based predictors showed significant differences in neural responses for these two lexical categories, a distinction not captured by top-1 prediction. Furthermore, larger and more advanced language models typically show a close correspondence to human productions in next-word prediction, according to information-theoretic measures. However, our findings indicated that bigger is not always better. In other words, increasing model size, complexity, or training data does not necessarily guarantees better performance in predicting word-by-word neural responses. As shown in our experiments, surprisal-based GPT-2 Large regression predictions substantially outperformed those of the larger GPT-Neo 2.7B model when comparing grand-averaged predictions within or between subjects, despite the latter’s increased size and computational resources.

6.4.2 Limitations

Despite the valuable insights gained from this research, several limitations must be acknowledged to guide future improvements and broader applicability. First, this chapter primarily focused on the GPT family of transformer-based language models. While GPT models are widely popular, the landscape of transformer-based LMs has rapidly evolved. Other unidirectional models, such as LLaMA [305] and Falcon LLMs [306], also leverage transformer architectures but introduce unique training

objectives and capabilities. These differences may significantly influence cognitive modelling. To develop more comprehensive insights into language model generation, future studies should broaden the scope to include a diverse set of transformer-based models.

Secondly, we have explored next-word predictability primarily through the lens of lexical groups. While our results provided valuable insights, it represents a limited aspect of human reading behaviour. Evidence suggests that human sensitivity to surprisal extends to highly predictable words [88, 307] or frequent words [297, 308]. Consequently, the lack of aspects resulted in the surprisal effect may not generalise across varying levels of word difficulty.

Finally, we concentrated our experiments on top-down processes, specifically those involving semantic processing. However, estimates of word predictability costs derived from language models significantly underestimate the contribution of bottom-up processes, particularly those driven by syntactic structures [309–311]. Therefore, future studies should aim to incorporate syntactic processing to develop more cognitively plausible predictions of neural responses based on language models.

Chapter 7

Conclusion and Future Work

A better understanding of how the reading process works in the human brain is a crucial step toward modifying and developing human-like language models. An important aspect to be discussed in this chapter is the ethical considerations associated with this line of research, which are discussed in Section 7.1. Next, we will summarise the thesis’s primary contributions in Section 7.2. Finally, we will outline future directions to achieve comprehensive human-level language understanding in a computational language model in Section 7.3.

7.1 Ethical Considerations

In concluding this thesis, we should make attention to the ethical considerations of human language processing signal research. As our research involves the collection of brain signal datasets, we encountered numerous ethical protocols and requirements. Despite this, we realised that several crucial ethical dimensions still demand careful examination and reflection.

7.1.1 Dual Use

Dual use research of concern (DURC), as defined by the World Health Organization, refers to scientific research originally intended to benefit society but with potential for malicious misapplication. In recent years, the growing intersection of human-

machine interactions, particularly in brain signal processing, has increasingly raised concerns about technologies that can potentially access, evaluate, and influence thought, emotion, and behaviour.

Critical questions emerge: What are the potential implications of these research results? What underlying objectives drive the final applications? How might it affect individuals and society? These questions highlight the vulnerability to dual-use risks. Machine learning applications that leverage cognitive insights (i.e., decoding and encoding modelling) have the potential to be utilised in both beneficial and harmful ways [312, 313].

7.1.2 Observational Bias

It is important to critically consider where the data comes from and identify any biases present within the dataset. Henrich et al. [314] claimed that researchers frequently publish dataset about human psychology based exclusively on samples from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies. This assumption of universality has resulted in a heavily biased collection of psychological data. The exclusion or misrepresentation of demographic groups has significant consequences [315], because this in itself already represents an ethical problem for research purposes, threatening the universality and objectivity of scientific knowledge [316].

Beyond sampling biases, Caliskan et al. [317] demonstrated that text corpora retain recoverable imprints of historical biases, which can be extracted and are even reflected in EEG recordings [318]. Therefore, researchers must exercise extreme caution. The reuse of similar data corpora risks perpetuating and amplifying existing biases across multiple experimental contexts. Carefully interpreting results is essential to uphold scientific integrity and capture the complexity of human experience.

7.1.3 Protecting Privacy

The emerging capability to decode human thoughts introduces profound ethical challenges regarding privacy. Society must confront the ethical implications of thought-decoding technologies. Researchers have a responsibility to protect the rights and dignity of research participants, who willingly share their brain activity for scientific purposes. Thus, both the participants and their data must be treated respectfully, even if NLP practitioners are reusing rather than recording it themselves. Despite data anonymisation protocols, researchers should avoid attempts to reconstruct or trace insights back to individual participants such as our emotions, preferences, and mental health.

The responsible deployment of cognitive-inspired language models requires a comprehensive and collaborative approach involving developers, regulators, and users. Developing effective policies will require collaboration between governments, researchers, AI developers, and ethicists. Through this collaborative approach, stakeholders can ensure that thought-decoding technologies are implemented ethically, balancing the potential benefits with the need to protect privacy.

7.2 Summary

7.2.1 Contributions

With the blooming of AI technology, language models have made tremendous progress in recent years. However, we may be overlooking the most fundamental question of all: *What does it truly mean to have a mind?* This was the question posed by the legendary British mathematician Alan Turing in 1950 with his provocative query: “Can machines think?” Humans, unlike language models, do not navigate the world by scanning vast databases. Our thinking is not simply algorithmic, measuring correlations between variables. Moreover, advanced LMs, with their billions of adjustable parameters, have become too complex to fully decipher. In these and other respects, their functionality is fundamentally different from how human minds

work.

Building a LM that can think as human is a long-term and overwhelming challenge for neuroscience or AI researchers. Until now, our understanding of reading comprehension in the human mind is incomplete, making the development of human-like language models a goal that is far from reach. Given this complexity, we proposed approaching the challenge through the lens of behaviour. Reading is not a natural process; we learn to read by observing and mimicking the reading behaviours of previous generations, which then shape the reading processes in our minds. LMs are trained to excel at predicting the next word in the sequence given prior context, and humans also share this predictability in reading comprehension. It could be said that an important bridge between LMs and human in reading comprehension lies in next-word predictability. In this dissertation, we pursued this goal through an interdisciplinary approach to understanding reading behaviour, focusing on three main threads:

- Implementing and sharing datasets designed for EEG-based reading processing in order to research reading processing in the human mind and LMs.
- Leveraging brain signals to uncover the mechanisms underlying reading processing in the human mind.
- Evaluating next-word predictability in both language models and humans to examine the degree of cognitive-like LMs can display beyond the mindless identification of statistical correlations.

7.2.2 Research Question 1

“How do we design a standard framework for EEG-based reading experiment?”

We introduced the DERCo, which combines EEG recordings and next-word prediction data from narrative text reading. As detailed in Chapter 3 and Chapter 4,

the DERCo dataset comprises behavioural data from 500 online participants and EEG recordings from 22 native English speakers. By integrating behavioural predictions and neural responses, this dataset revealed significant differences in brain activity between high and low predictable words. This dataset addressed the shortage of public brain activity data for semantic-level reading and provides a valuable resource for both neurolinguistics and AI research in reading comprehension. This data collection effort established the foundation of this thesis, allowing us to pursue primary research questions.

In addition to introducing the DERCo dataset, we provided a step-by-step guide for designing EEG-based reading experiments and developed an EEG-based reading processing pipeline. This effort was motivated by the lack of a common framework for conducting EEG-based reading experiments. Most existing experiments in this field require researchers to have prior knowledge, making it difficult for beginners. Moreover, preprocessing pipelines for EEG signals often depend on the specific experimental design, such as those for speech, motor, or reading tasks. There are several proposed pipelines, but not much work has been conducted on EEG signals for reading experiments. Additionally, the comparison of different cleaning algorithms is ambiguous due to the lack of a standardised evaluation metric, and the selection often relies on researchers' purpose.

7.2.3 Research Question 2

“Are neural responses to upcoming words attributable to quantitative differences in lexical categories (content versus function words) and grammatical categories (e.g., nouns, verbs, adjectives, pronouns, etc.)?”

Recent advances in machine learning have motivated researchers to test neurolinguistic hypotheses using neural decoding methods. Prediction in language has traditionally been studied relied on ERP analysis to examine brain responses to expected and unexpected words are compared in a categorical fashion. However, this

method had limitations in producing clear visual representations, failed to capture personal variations among individuals, and detecting complex patterns. Currently, EEG decoding approach based on ML algorithms has provided better pattern recognition, multi-dimensional analysis, and clearer interpretations. In Chapter 5, our study successfully used EEG decoding to investigate how the brain processes language through two mechanisms: bottom-up processing (handling complex language structure) and top-down processing (dealing with word predictability). We explored how different word types (content words and function words) and grammatical categories affect brain activity during reading comprehension, particularly when people try to predict upcoming words.

In detail, we indicated how word predictability (measured by cloze probability) affects brain responses to two main word groups: content words (i.e., nouns, verbs, adjectives, and adverbs) versus function words (i.e., pronouns, determiners, auxiliaries, and adpositions). Our results revealed several key patterns: (1) highly predictable words generated more distinct N400 patterns across different word categories, with content words displaying larger N400 variations between high and low predictability states than function words; (2) within content words, verbs exhibited more pronounced N400 differences compared to nouns, although nouns conveyed more clear-cut information about their predictability; and (3) for function words, determiners stood out with the most significant N400 effects, presumably due to their diverse contextual roles in sentences. These findings demonstrate how EEG decoding techniques can illuminate the intricate interplay between language structure and word predictability in neural processing.

7.2.4 Research Question 3

“Which type of language model most closely aligns with human next-word predictability during reading comprehension?”

Language models have demonstrated impressive proficiency in natural language tasks, such as text generation, coding, and answering questions. However, excelling

in these tasks does not necessarily mean that LMs can model human intelligence. Even theoretical attempts to replicate the structure of the human brain would not guarantee that an LM have a thought like a human. This limitation stems from a fundamental aspect of human cognitive development: language comprehension is not an innate capability but rather an acquired skill. Humans develop sophisticated language processing mechanisms through progressive learning and exposure. This developmental trajectory suggests that to achieve more anthropomorphic artificial intelligence, the focus should shift from architecture to learning behaviour modelling.

As a starting point, LMs are primarily trained to predict the next word (or token) in a sequence based on prior context, which captures a fundamental aspect of intelligence. Nevertheless, evaluating LMs only based on accuracy fails to integrate the multifaceted nature of human intelligence. We can capture a bigger knowledge beyond that accuracy such as reasoning, planning, imaginative thinking, common sense, working memory, and factual recall. In seeking solutions, we adapted brain encoding models based on information-theoretic surprisal to analyse neural response patterns during reading comprehension. Empirical evidence suggests robust correlations between distributional lexical surprisal and neural responses, particularly in N400 time window. This approach offers a promising avenue for developing more sophisticated evaluation metrics that better align with human cognitive processes.

In Chapter 6, to address the final research question RQ3, we generated regressors based on two information measures—top-1 accuracy and surprisal—to predict ERP patterns in humans and LMs. Our findings indicate that while more advanced LMs closely align with human performance in word prediction accuracy, only surprisal shows a robust correlation with language-processing ERPs, particularly for open-class words with high semantic content. Moreover, our results challenge the assumption that increasing model size and training will consistently enhance cognitive convergence with human-like linguistic processing. This research highlights the potential for improving LM training to make them more human-like. Although

LMs are fundamentally designed as next-token prediction systems, integrating neural responses informed by information theory, alongside accuracy, may facilitate the development of more cognitively plausible reading processes that better approximate human language comprehension mechanisms.

7.3 Direction for Future Research

Through our past and ongoing projects, we have made contributions to this field and gained valuable insights into the interaction between humans and language models in the context of reading comprehension. Some of the remaining open challenges in this area include uncovering theoretical hypotheses related to reading and language processing in the human brain, as well as extending research to encompass a broader range of languages. Based on the current state of this research field and the work presented in this thesis, we have identified several promising avenues for future research, which are outlined below.

7.3.1 Unfolding Reading Processing in the Human Brain and LMs

The intersection of brain activity analysis and language modelling has provided unprecedented insights into human perception, action, and cognitive processes. In the context of reading processes, researchers have employed decoding and encoding techniques to analyse the information represented in brain activity. Briefly, decoding involves using neural activity to predict stimuli [319, 320], while encoding involves using stimuli to predict brain activity [321, 322]. The line of research predominately leverage brain encoding and decoding models to examine human intelligence with machine intelligence in reading processing. We anticipate that future research directions can be developed from two key perspectives.

First, this research presents an important opportunity for the scientific community to gain a deeper understanding of how language is processed in the brain during

reading and how it forms the foundation of human intelligence. For centuries, neuroscientists have studied brain function by linking neural activity to specific inputs and outputs. However, when attempting to measure thoughts, there is no direct benchmark for comparison with neural activity. Instead, researchers often analyse behaviour to infer the beliefs or thoughts that best explain that behaviour. Previous studies have indicated that the difficulty of comprehending words in context is closely linked to their predictability [85, 87, 323], which is a critical aspect of reading behaviour. However, the cognitive mechanisms underlying these predictability effects have been a topic of debate for many years [7, 77, 324, 325]. Therefore, future research could leverage encoding-decoding models to provide empirical evidence supporting probabilistic inference as a fundamental component of human language comprehension.

Not limited to information-theoretic metrics, decoding and encoding models have also been applied to investigate word representations in the human brain. Researchers have been particularly interested in understanding how processing depth influences brain representations of words and concepts. Existing research demonstrated that the contextual embeddings generated by LLMs more accurately predict shared neural activity between speakers and listeners compared to non-contextual word embeddings [76, 326, 327]. Furthermore, investigations have revealed common geometric patterns between neural embeddings in human brain and contextual embeddings derived from transformer-based models. [76, 328, 329]. Looking forward, a promising research trajectory involves fine-tuning language models using synthetic brain data during the word embedding stage. This approach could potentially enhance the human-like characteristics of computational language representations, bridging the gap between artificial and neural information processing.

7.3.2 Translation of Thoughts to Text

Content-Driven EEG-to-Text Translation

The conversion of brain activity into text using brain signals has gained considerable attention in recent years. Researchers are increasingly focused on developing advanced models to decode EEG signals for direct brain-to-text communication [330]. This field incorporates various methodologies for data collection, including EEG, fMRI, and Electrocorticography (ECoG), each providing unique insights into brain activity. Deep neural networks, particularly RNNs [331], and autoregressive transformer-based models [207, 217], have shown significant potential in translating thought patterns into text. This progress signifies a promising intersection of cutting-edge neural network architectures and neuroscientific data, paving the way for more accurate and efficient EEG-to-text conversion techniques.

Notable architectures include DeWave, an end-to-end model for EEG-to-text translation, which employs self-supervised and contrastive learning to establish robust links between brain activity and language [207]. Additionally, the MDADenseNet-AM model introduces a sophisticated deep learning architecture for converting EEG signals into text [332]. The EEG-To-Text model addresses the limitations of predefined open-vocabulary EEG-to-text decoding by integrating convolutional neural networks (CNNs) for feature extraction and LSTM networks for sequence-to-sequence decoding [217].

Emotion-Enhanced EEG-to-Text Models

Recent breakthroughs in EEG-to-text conversion represent a significant milestone in neuroscientific research, unlocking unprecedented communication possibilities. However, current models predominantly focus on literal content translation, overlooking a critical dimension of human communication: emotional context. Addressing this limitation requires advancing the field toward conveying emotions through EEG signals. Achieving this goal demands a deeper understanding of the neural correlates of emotions and the development of sophisticated algorithms to process and integrate

emotional information.

Enhancing the emotional sensitivity of EEG-to-text systems will not only improve their fidelity and usability but also hold transformative implications for fields such as mental health. Recognising and interpreting emotions through EEG signals can facilitate more accurate diagnoses and more effective therapies. The ultimate goal is to develop technologies that not only translate thoughts but capture the rich, complex emotional landscape underlying human expression.

Bilingual EEG-to-Text Models

An emerging and compelling research trajectory involves cross-language brain decoding, offering unprecedented insights into multilingual neural representation. Recent studies have demonstrated remarkable progress in understanding how the brain processes linguistic information across different languages. Silva et al. [333] utilized ECoG alongside deep learning and statistical natural language models for English and Spanish. Their models achieved 88% accuracy in distinguishing between the two languages from the first word and decoded entire sentences with 75% accuracy.

Similarly, Yang et al. [334] extended cross-language semantic decoding by leveraging a concept’s fMRI signature to decode sentences across three languages: English, Portuguese, and Mandarin. In their study, a classifier trained on word-activation mappings from one or two languages was tested on its ability to decode semantic content in a third language, demonstrating the feasibility of cross-language decoding. Dehghani et al. [335] further explored this area by exposing native speakers of English, Mandarin, and Farsi to translated story narratives during fMRI scanning. Their findings showed that cross-language story predictions were not significantly affected by language-specific lexical or syntactic variations, indicating the presence of shared semantic constructs across languages.

To our knowledge, there have been no EEG-to-text systems designed specifically for bilinguals. Developing cross-language EEG decoding systems—leveraging models trained on brain activity from one language to decode stimuli in another—presents

a promising avenue for future research. This approach could shed light on how the brain represents and processes multiple languages, offering new insights into bilingualism and the neural basis of language comprehension.

Appendix

A Correlation

A.1 Pearson Correlation

The Pearson r correlation is one of the most widely used statistical measures to quantify the dependence between pairs of time series, particularly in the analysis of bio-signals [336]. For instance, if we aim to measure the relationship between two signals, the Pearson r correlation quantifies the degree of relationship between the two. The correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative linear correlation, 0 signifies no correlation, and +1 represents a perfect positive linear correlation.

A point-biserial correlation is conducted with the Pearson correlation formula except that one of the variables is dichotomous. The Pearson correlation coefficient, denoted as r_{xy} , is defined as:

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

where:

- r_{xy} : Pearson correlation coefficient between x and y
- n : number of observations
- x_i : value of x (for the i -th observation)
- y_i : value of y (for the i -th observation)

A.2 Spearman Correlation

Spearman rank correlation is a non-parametric test that is used to measure of the strength of a monotonic relationship between two independent variables [337]. Compared to Pearson correlation coefficient, Spearman correlation coefficient operates on the ranks of the data rather than the raw data, and it does not require the relationship between variables is linear. Since it is based on the ranks of the data, it can well represent the similarity of the trend of the time series.

Spearman correlation analysis ranks each variable separately from lowest to highest and record the difference between ranks of each data pair [338]. The sum of the square of the difference between ranks denotes the strength of the correlation between variables. If the data is strong correlated, then the sum will be small, vice versa. Besides, the magnitude of the sum is related to the significance of the correlation. The Spearman ranks correlation coefficient can be calculated using the following equations:

$$r_s = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)} \quad (1)$$

where d_i is the difference between ranks for each data pair, and N is the number of data pairs.

B Smoothing Techniques in N-Grams

When dealing with **n-gram** models, smoothing refers to the practice of adjusting empirical probability estimates to account for insufficient data.

B.1 Laplace Smoothing

Laplace smoothing, also known as add-one smoothing, assumes that each n-gram in a corpus occurs exactly one more time than it actually does. The simplest way to do this smoothing is to add one to all the n-gram counts, before we normalize them into probabilities. All the counts that used to be zero will now have a count of 1,

the counts of 1 will be 2, and so on. In the equation below, we use the notation $w_i^j, i < j$, to denote the $(j - i)$ -gram $(w_i, w_{i+1}, \dots, w_j)$.

$$p(w_i \mid w_{i-n+1}^{i-1}) = \frac{1 + c(w_{i-n+1}^i)}{|V| + \sum_{w_i} c(w_{i-n+1}^i)}$$

where $c(a)$ denotes the empirical count of the n -gram a in the corpus, and $|V|$ corresponds to the number of unique n -grams in the corpus.

B.2 Kneser–Ney Smoothing

Unlike Laplace smoothing, Kneser-Ney smoothing not only accounts for the frequency of observed n -grams but also considers the diversity of contexts word w has appeared in. For example, a word like “the” will have a high raw frequency but occurs in many repetitive contexts (e.g., “the dog,” “the cat”). In contrast, a word like “Francisco” might appear less frequently but in more unique contexts i.e., “San Francisco”, making it more informative.

This method is an extension of absolute discounting with a clever way of constructing the lower-order (backoff) model. The lower-order model is significant only when count is small or zero in the higher-order model, and so should be optimised for that purpose. The probability of word w_i given its previous context w_{i-n+1}^{i-1} (the previous $n - 1$ words) using Kneser-Ney smoothing is expressed as follows:

$$P_{KN}(w_i \mid w_{i-n+1}^{i-1}) = \frac{\max(c_{KN}(w_{i-n+1}^i) - d, 0)}{c_{KN}(w_{i-n+1}^{i-1})} + \lambda(w_{i-n+1}^{i-1})P_{KN}(w_i \mid w_{i-n+2}^{i-1})$$

where,

$c_{KN}(w_{i-n+1}^{i-1})$: represents the count of the preceding context w_{i-n+1}^{i-1} . The definition of the count c_{KN} depends on whether we are counting the highest-order N -gram being interpolated (for example, trigram if we are interpolating trigram, bigram, and unigram) or one of the lower-order N -grams (bigram or unigram if we are interpolating trigram, bigram, and unigram):

$$c_{KN}(\cdot) = \begin{cases} \text{count}(\cdot) & \text{for the highest order,} \\ \text{continuation count}(\cdot) & \text{for lower orders.} \end{cases}$$

$\max(c_{KN}(w_{i-n+1}^i) - d, 0)$: applies a discount d of the observed count $c_{KN}(w_{i-n+1}^i)$.

If the discounted value becomes negative, the max function ensures it is clipped to zero.

$\lambda(w_{i-n+1}^{i-1})$: is known as the back-off weight. It is simply the amount of probability mass we left for the next lower order model.

$P_{KN}(w_i \mid w_{i-n+2}^{i-1})$: represents the backoff probability, calculated recursively for a lower-order $n - 1$ -gram model.

C DERCo Dataset

C.1 DERCo Ethic Approval

Ollscoil Chathair Bhaile Átha Cliath
Dublin City University



Ms Boi Mai Quach
School of Computing

10th May 2023

REC Reference: DCUREC/2023/085

Proposal Title: Using EEG Signals To Examine Next-Word Prediction Ability of NLP Models in Reading

Applicant(s): Ms Boi Mai Quach, Dr Graham Healy, Dr Cathal Gurrin

Dear Colleagues,

Thank you for your application to DCU Research Ethics Committee (REC). Further to notification review, DCU REC is pleased to issue approval for this research proposal.

DCU REC's consideration of all ethics applications is dependent upon the information supplied by the researcher. This information is expected to be truthful and accurate. Researchers are responsible for ensuring that their research is carried out in accordance with the information provided in their ethics application.

Materials used to recruit participants should note that ethical approval for this project has been obtained from the Dublin City University Research Ethics Committee. Should substantial modifications to the research protocol be required at a later stage, a further amendment submission should be made to the REC.

Yours sincerely,



Dr. Melrona Kirrane
Chairperson
DCU Research Ethics Committee



Taighde & Nuálaíocht Tacaíocht
Ollscoil Chathair Bhaile Átha Cliath,
Baile Átha Cliath, Éire

Research & Innovation Support
Dublin City University,
Dublin 9, Ireland

T +353 1 700 8000
F +353 1 700 8002
E research@dcu.ie
www.dcu.ie

Note: Please retain this approval letter for future publication purposes (for research students, this includes incorporating the letter within their thesis appendices).

C.2 Questionnaire for Five Articles

Question about the article you have just read

Congratulations! You have just finished the story "**The Mouse, the Bird, and the Sausage**". Please answer the following questions before moving on to the next story.

*Please respond to the following questions by placing a **check mark** (✓)*

General questions

1. Have you ever read this specific article before? ☐ Yes | ☐ No
2. Have you ever read a similar article before? ☐ Yes | ☐ No
3. Did the article use too many advanced/unfamiliar words? ☐ Yes | ☐ No
4. Estimate the percentage of content you still remember? Please indicate with a **check mark** (✓) on the scale below. You can also check between markings.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

Story-related questions

Who is responsible for bringing wood from the forest?

- ☐ The Mouse
☐ The Bird
☐ The Sausage

Why did the Bird refuse to go to the forest one day?

- ☐ Because it was tired.
☐ Because it wanted to cook instead.
☐ Because other birds made fun of it.

What happened to the Sausage when it went to collect wood in the forest?

- ☐ It got lost.
☐ It was attacked and eaten by a dog.
☐ It found a treasure.

How did the Mouse try to cook after the Sausage was gone?

- ☐ By grilling
- ☐ By baking
- ☐ By rolling itself in the pot to mix and season the food

What happened to the Bird in the end?

- ☐ It was attacked by the dog
- ☐ It drowned in the well
- ☐ It flew away and never returned

Question about the article you have just read

Congratulations! You have just finished the story "**Straw, Coal, and Bean**". Please answer the following questions before moving on to the next story.

*Please respond to the following questions by placing a **check mark** (✓) on the scale below.*

General questions

1. Have you ever read this specific article before? ☐ Yes | ☐ No
2. Have you ever read a similar article before? ☐ Yes | ☐ No
3. Did the article use too many advanced/unfamiliar words? ☐ Yes | ☐ No
4. Estimate the percentage of content you still remember? Please indicate with a **check mark** (✓) on the scale below. You can also check between markings.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

Story-related questions

How did the coal escape from the fireplace?

- ☐ It rolled out by itself.
- ☐ It was thrown out by the old woman.
- ☐ It jumped out on its own.
- ☐ It was carried out by the bean.

Why did the straw catch fire and fall into the brook?

- ☐ The old woman set it on fire.
- ☐ The coal accidentally ignited it.
- ☐ It got wet from the water in the brook.
- ☐ It burned due to the heat from the coal.

Who saved the bean after it burst from laughing?

- ☐ The old woman
- ☐ The tailor
- ☐ The straw
- ☐ The coal

How did the tailor fix the burst bean?

- ☐ He glued it back together.
- ☐ He used a needle and thread to sew it.
- ☐ He replaced it with a new bean.
- ☐ He wrapped it with a piece of cloth.

Why do all beans have a black seam since that time?

- ☐ The beans were burned in the fireplace.
- ☐ The old woman sewed them with black thread.
- ☐ The tailor used black thread to sew the bean.
- ☐ The water in the brook turned the beans black.

Question about the article you have just read

Congratulations! You have just finished the story " **Poverty and Humility Lead to Heaven**". Please answer the following questions before moving on to the next story.

*Please respond to the following questions by placing a **check mark** (✓)*

General questions

1. Have you ever read this specific article before? ☐ Yes | ☐ No
2. Have you ever read a similar article before? ☐ Yes | ☐ No
3. Did the article use too many advanced/unfamiliar words? ☐ Yes | ☐ No
4. Estimate the percentage of content you still remember? Please indicate with a **check mark** (✓) on the scale below. You can also check between markings.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

Story-related questions

What did the king's son do when he felt sad and full of thought?

- ☐ He prayed for wealth and power.
- ☐ He went to live in a cave.
- ☐ He looked at the sky and sighed.
- ☐ He started singing and dancing.

How did the poor gray-haired man suggest the king's son could reach heaven?

- ☐ By wearing luxurious clothes and jewels.
- ☐ By asking for food from compassionate hearts.
- ☐ By becoming a king and ruling a kingdom.
- ☐ By studying and gaining knowledge.

Why did the servants at the king's palace not believe that the beggar was the king's son?

- ☐ They thought he was too well-dressed.
- ☐ They didn't recognize him due to his ragged clothes.
- ☐ They were told not to let anyone enter the palace.
- ☐ They believed he was an imposter.

How did the ill-natured servant treat the food meant for the beggar?

- ☐ He shared it with the beggar.
- ☐ He kept it for himself.
- ☐ He gave it to the dogs.
- ☐ He threw it away.

What grew on either side of the beggar's grave after he died?

- ☐ Roses on both sides.
- ☐ Lilies on both sides.
- ☐ Roses on one side and lilies on the other.
- ☐ Nothing grew near his grave.

Question about the article you have just read

Congratulations! You have just finished the story " **The Death of the Little Hen**". Please answer the following questions before moving on to the next story.

*Please respond to the following questions by placing a **check mark** (✓) on the scale below.*

General questions

1. Have you ever read this specific article before? ☐ Yes | ☐ No
2. Have you ever read a similar article before? ☐ Yes | ☐ No
3. Did the article use too many advanced/unfamiliar words? ☐ Yes | ☐ No
4. Estimate the percentage of content you still remember? Please indicate with a **check mark** (✓) on the scale below. You can also check between markings.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

Story-related questions

What did the little hen find on Nut Mountain?

- ☐ A small pebble
- ☐ A large nut
- ☐ A colourful feather
- ☐ A shiny coin

What did the little hen want to do with the nut she found?

- ☐ Share it with the little rooster
- ☐ Bury it in the ground
- ☐ Plant it to grow a nut tree
- ☐ Eat the kernel herself

How did the little hen get into trouble with the nut kernel?

- ☐ She lost it on the way down Nut Mountain
- ☐ She swallowed it whole without chewing
- ☐ She gave it away to the fox
- ☐ She threw it into the well

Why did the little rooster need red silk from the bride?

- ☐ To make a scarf for himself
- ☐ To wrap around the nut kernel
- ☐ To trade with the well for water
- ☐ To fix the broken carriage

How did the little hen and the other animals meet their end?

- ☐ They were attacked by a group of mice
- ☐ They got lost in the forest
- ☐ They drowned in the brook
- ☐ They were turned to stone by a magical spell

Question about the article you have just read

Congratulations! You have just finished the story "**The Wolf and the Fox**". Please answer the following questions before moving on to the next story.

*Please respond to the following questions by placing a **check mark** (✓)*

General questions

1. Have you ever read this specific article before? ☐ Yes | ☐ No
2. Have you ever read a similar article before? ☐ Yes | ☐ No
3. Did the article use too many advanced/unfamiliar words? ☐ Yes | ☐ No
4. Estimate the percentage of content you still remember? Please indicate with a **check mark** (✓) on the scale below. You can also check between markings.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

Story-related questions

What was the main reason the fox decided to steal a lamb for the wolf?

- ☐ Because the fox was feeling hungry
- ☐ To please the wolf and avoid being eaten
- ☐ To play a prank on the wolf
- ☐ To teach the wolf a lesson about greed

How did the wolf end up getting caught while trying to steal the second lamb?

- ☐ The farmer's dog attacked him
- ☐ The fox betrayed him to the farmer's people
- ☐ The wolf made too much noise
- ☐ The sheep recognized the wolf

Why did the wolf want more pancakes after eating the first six?

- ☐ He was still hungry
- ☐ He wanted to share with the fox
- ☐ He wanted to take some back for later
- ☐ He accidentally dropped the first six

How did the farmer's wife discover the wolf when he tried to get more pancakes?

- ☐ The wolf knocked over a dish
- ☐ She heard the wolf howling
- ☐ The fox told her about the wolf
- ☐ She smelled the wolf's presence

What ultimately happened to the wolf when he tried to steal meat from the cellar?

- ☐ He successfully escaped with the meat
- ☐ He was caught by the farmer and killed
- ☐ The fox helped him get out safely
- ☐ He realized he couldn't fit through the hole and gave up

Questions about your experience

For the questions below, on a scale of 1-10, where 1 is **Strongly Disagree** and 10 is **Strongly Agree**. Please **circle** the number that best represents your answer.

To what extent do you disagree/agree with the following statements?

1. The instructions and explanations by the experimenters *alone* provided me with enough information to perform the experiment.

1 2 3 4 5 6 7 8 9 10

Strongly disagree

Strongly agree

2. The experimental room is too noisy to focus on the reading experiment.

1 2 3 4 5 6 7 8 9 10

Strongly disagree

Strongly agree

3. I frequently got distracted when reading the article.

1 2 3 4 5 6 7 8 9 10

Strongly disagree

Strongly agree

4. Current condition checklist. Please put a **check mark (✓)** at the number that represents your best answer, where 1 is **extremely low** and 10 represents **very high**.

Condition	1	2	3	4	5	6	7	8	9	10
Comfort										
Concentration										
Eye fatigue										
Drowsiness										

C.3 Transcripts of Five Articles

Story 1: “The Mouse, the Bird, and the Sausage”

Once upon a time a mouse, a bird, and a sausage formed a partnership. They kept house together, and for a long time they lived in peace and prosperity, acquiring many possessions. The bird’s task was to fly into the forest every day to fetch wood. The mouse carried water, made the fire, and set the table. The sausage did the cooking.

Whoever is too well off always wants to try something different! Thus one day the bird chanced to meet another bird, who boasted to him of his own situation. This bird criticized him for working so hard while the other two enjoyed themselves at home. For after the mouse had made the fire and carried the water, she could sit in the parlor and rest until it was time for her to set the table. The sausage had only to stay by the pot watching the food cook. When mealtime approached, she would slither through the porridge or the vegetables, and thus everything was greased and salted and ready to eat. The bird would bring his load of wood home. They would eat their meal, and then sleep soundly until the next morning. It was a great life.

The next day, because of his friend’s advice, the bird refused to go to the forest, saying that he had been their servant long enough. He was no longer going to be a fool for them. Everyone should try a different task for a change. The mouse and the sausage argued against this, but the bird was the master, and he insisted that they give it a try. The sausage was to fetch wood, the mouse became the cook, and the bird was to carry water.

And what was the result? The sausage trudged off toward the forest; the bird made the fire; and the mouse put on the pot and waited for the sausage to return with wood for the next day. However, the sausage stayed out so long that the other two feared that something bad had happened. The bird flew off to see if he could find her. A short distance away he came upon a dog that had seized the sausage as free booty and was making off with her. The bird complained bitterly to the dog

about this brazen abduction, but he claimed that he had discovered forged letters on the sausage, and that she would thus have to forfeit her life to him.

Filled with sorrow, the bird carried the wood home himself and told the mouse what he had seen and heard. They were very sad, but were determined to stay together and make the best of it. The bird set the table while the mouse prepared the food. She jumped into the pot, as the sausage had always done, in order to slither and weave in and about the vegetables and grease them, but before she reached the middle, her hair and skin were scalded off, and she perished.

When the bird wanted to eat, no cook was there. Beside himself, he threw the wood this way and that, called out, looked everywhere, but no cook was to be found. Because of his carelessness, the scattered wood caught fire, and the entire house was soon aflame. The bird rushed to fetch water, but the bucket fell into the well, carrying him with it, and he drowned.

Story 2: “Straw, Coal, and Bean”

An old woman lived in a village. She had gathered a serving of beans and wanted to cook them, so she prepared a fire in her fireplace. To make it burn faster she lit it with a handful of straw. While she was pouring the beans into the pot, one of them fell unnoticed to the floor, coming to rest next to a piece of straw. Soon afterward a glowing coal jumped out of the fireplace and landed next to them.

The straw said, “Dear friends, where do you come from?”

The coal answered, “I jumped from the fireplace, to my good fortune. If I had not forced my way out, I surely would have died. I would have burned to ash.”

The bean said, “I too saved my skin. If the old woman had gotten me into the pot I would have been cooked to mush without mercy, just like my comrades.”

“Would my fate have been any better?” said the straw. “The old woman sent all my brothers up in fire and smoke. She grabbed sixty at once and killed them. Fortunately, I slipped through her fingers.”

“What should we do now?” asked the coal.

“Because we have so fortunately escaped death,” answered the bean, “I think that we should join together as comrades. To prevent some new misfortune from befalling us here, let us together make our way to another land.”

This proposal pleased the other two, and they set forth all together. They soon came to a small brook, and because there was neither a bridge nor a walkway there, they did not know how they would get across it. Then the straw had a good idea, and said, “I will lay myself across it, and you can walk across me like on a bridge.”

So the straw stretched himself from one bank to the other. The coal, who was a hot-headed fellow, stepped brashly onto the newly constructed bridge, but when he got to the middle and heard the water rushing beneath him, he took fright, stopped, and did not dare to go any further. Then the straw caught fire, broke into two pieces, and fell into the brook. The coal slid after him, hissed as he fell into the water, and gave up the ghost.

The bean who had cautiously stayed behind on the bank had to laugh at the event. He could not stop, and he laughed so fiercely that he burst. Now he too would have died, but fortunately a wandering tailor was there, resting near the brook. Having a compassionate heart, he got out a needle and thread and sewed the bean back together. The bean thanked him most kindly. However, because he had used black thread, since that time all beans have had a black seam.

Story 3: “Poverty and Humility Lead to Heaven”

There was once a king’s son who went out into the world, and he was full of thought and sad. He looked at the sky, which was so beautifully pure and blue, then he sighed, and said, how well must all be with one up there in heaven. Then he saw a poor gray-haired man who was coming along the road towards him, and he spoke to him, and asked, how can I get to heaven. The man answered, by poverty and humility.

Put on my ragged clothes, wander about the world for seven years, and get to know what misery is, take no money, but if you are hungry ask compassionate hearts

for a bit of bread. In this way you will reach heaven. Then the king's son took off his magnificent coat, and wore in its place the beggar's garment, went out into the wide world, and suffered great misery. He took nothing but a little food, said nothing, but prayed to the Lord to take him into his heaven.

When the seven years were over, he returned to his father's palace, but no one recognized him. He said to the servants, go and tell my parents that I have come back again. But the servants did not believe it, and laughed and left him standing there. Then said he, go and tell it to my brothers that they may come down, for I should so like to see them again. The servants would not do that either, but at last one of them went, and told it to the king's children, but these did not believe it, and did not trouble themselves about it.

Then he wrote a letter to his mother, and described to her all his misery, but he did not say that he was her son. So out of pity, the queen had a place under the stairs assigned to him, and food taken to him daily by two servants. But one of them was ill-natured and said, why should the beggar have the good food, and kept it for himself, or gave it to the dogs, and took the weak, emaciated beggar nothing but water. The other, however, was honest, and took the beggar what was sent to him. It was little, but he could live on it for a while, and all the time he was quite patient, but he grew continually weaker.

As his illness increased, he desired to receive the last sacrament. When the mass was being celebrated, all the bells in the town and neighborhood began to ring of their own accord. After mass the priest went to the poor man under the stairs, and there he lay dead. In one hand he had a rose, in the other a lily, and beside him was a paper on which was written his history. When he was buried, a rose grew on one side of his grave, and a lily on the other.

Story 4: "The Death of the Little Hen"

One time the little hen and the little rooster went to Nut Mountain, and they agreed that whoever would find a nut would share it with the other one. Now the little hen

found a large, large nut, but wanting to eat the kernel by herself - she said nothing about it. However, the kernel was so thick that she could not swallow it down. It got stuck in her throat, and fearing that she would choke to death, she cried out, "Little Rooster, I beg you to run as fast as you can to the well and get me some water, or else I'll choke to death."

The little rooster ran to the well as fast as he could, and said, "Well, give me some water, for the little hen is lying on Nut Mountain. She swallowed a large nut kernel and is about to choke to death on it." The well answered, "First run to the bride, and get some red silk from her."

The little rooster ran to the bride: "Bride, give me some red silk, and I'll give the red silk to the well, and the well will give me some water, and I'll take the water to the little hen who is lying on Nut Mountain. She swallowed a large nut kernel and is about to choke to death on it." The bride answered, "First run and get my wreath. It got caught on a willow branch." So the little rooster ran to the willow and pulled the wreath from its branch and took it to the bride, and the bride gave him some red silk, which he took to the well, which gave him some water, and the little rooster took the water to the little hen, but when he arrived, she had already choked to death, and she lay there dead, and did not move at all.

The little rooster was so sad that he cried aloud, and all the animals came to mourn for the little hen. Six mice built a small carriage which was to carry the little hen to her grave. When the carriage was finished, they hitched themselves to it, and the little rooster drove. On the way they met the fox.

"Where are you going, little rooster?"

"I'm going to bury my little hen."

"May I ride along?"

"Yes, but you must sit at the rear, because my little horses don't like you too close to the front."

So he sat at the rear, and then the wolf, the bear, the elk, the lion, and all the animals in the forest. They rode on until they came to a brook. "How can we get

across?” said the little rooster. A straw was lying there next to the brook, and he said, “I’ll lay myself across, and you can drive over me.” But just as the six mice got onto the straw, it slipped into the water, and the six mice all fell in and drowned.

They did not know what to do, until a coal came and said, “I am large enough. I will lay myself across and you can drive over me.” So the coal laid itself across the water, but unfortunately it touched the water, hissed, and went out; and it was dead. A stone saw this happen, and wanting to help the little rooster, it laid itself across the water. The little rooster pulled the carriage himself. He nearly reached the other side with the dead little hen, but there were too many others seated on the back of the carriage, and the carriage rolled back, and they all fell into the water and drowned. Now the little rooster was all alone with the dead little hen. He dug a grave for her and laid her inside. Then he made a mound on top, and sat on it, and grieved there so long that he too died. And then everyone was dead.

Story 5: “The Wolf and the Fox”

A wolf and a fox once lived together. The fox, who was the weaker of the two, had to do all the hard work, which made him anxious to leave his companion. One day, passing through a wood, the wolf said, “Red-fox, get me something to eat, or I shall eat you.”

The fox answered, “I know a place where there are a couple of nice young lambs; if you like, we will go and fetch one.”

This pleased the wolf, so they went. The fox stole one, brought it to the wolf, and then ran away, leaving his comrade to devour it. This done, the wolf was not content, but wishing for the other, went himself to fetch it; and being very awkward, the old sheep saw him, and began to cry and bleat so horribly that the farmer’s people came running to see what was the matter. Of course they found the wolf there, and beat him so unmercifully, that, howling and limping, he returned to the fox.

“You had already shown me how, so I went to fetch the other lamb,” said he, “but the farmer’s people discovered me, and have nearly killed me.”

“Why are you such a glutton?” replied the fox.

The next day they went again into the fields. “Red-fox,” said the wolf, “get me something quickly to eat, or I shall eat you!”

“Well,” replied the fox, “I know a farm, where the woman is baking pancakes this evening; let us go and fetch some.”

They went accordingly, and the fox, slipping round the house, peeped and sniffed so long, that he found out at last where the dish stood, then quietly abstracting six pancakes, he carried them to the wolf.

“Here is something for you to eat,” said he, and then went away. The wolf had swallowed the six pancakes in a very short space of time, and said, “I should very much like some more.”

But going to help himself, he pulled the dish down from the shelf; it broke into a thousand pieces, and the noise, in addition, brought out the farmer’s wife to discover what was the matter. Upon seeing the wolf, she raised such an alarm, that all the people came with sticks or any weapon they could snatch. The consequence was that the wolf barely escaped with his life; he was beaten so severely that he could scarcely hobble to the wood where the fox was.

“Pretty mischief you have led me into,” said the wolf, when he saw him, “the peasants have caught, and nearly flayed me.”

“Why, then, are you such a glutton?” replied the fox.

Upon a third occasion, being out together, and the wolf only able with difficulty to limp about, he nevertheless said again, “Red-fox, get me something to eat, or I shall eat you!”

“Well,” said the fox, “I know a man who has been butchering, and has all the meat salted down in a tub in his cellar. We will go and fetch it.”

“That will do,” said the wolf, “but I must go with you, and you can help me to get off, if anything should happen.”

The fox then showed him all the by-ways, and at last they came to the cellar, where they found meat in abundance, which the wolf instantly greedily attacked, saying

at the same time to himself, “Here, there is no occasion to hurry.”

The fox also showed no hesitation, only, while eating, he looked sharply about him, and ran occasionally to the hole by which they had entered in order to try if he was still small enough to get out by the same way he had come in.

“Friend fox,” said the wolf, “pray tell me why you are so fidgety, and why you run about in such an odd manner.”

“I am looking out, lest anyone should come,” replied the cunning creature. “Come, are you not eating too much?”

“I am not going away,” said the wolf, “until the tub is empty; that would be foolish!”

In the meantime, the farmer, who had heard the fox running about, came into the cellar to see what was stirring, and upon the first sight of him, the fox with one leap was through the hole and on his way to the wood. But when the wolf attempted to follow, he had so increased his size by his greediness, that he could not succeed, and stuck in the hole, which enabled the farmer to kill him with his cudgel. The fox, however, reached the wood in safety, and rejoiced to be freed from the old glutton.

C.4 Online Consent Form for Amazon Mechanical Turk Experiment.

In this experiment, you will read text and try to make predictions for each upcoming word.

The study is expected to take approximately **25 minutes**.

Thanks for participating!

Clicking on the **I Agree** button below indicates that:

- You voluntarily agree to participate
- You are 18 years old or above
- You are a native speaker of English

Note: Please aware that your de-identified data will be made indefinitely available to researchers.

If you do not agree to all of these, please close this window in your browser now.

Figure 1: Online Consent Form for Behavioural Word-Prediction Experiment.

C.5 Human Next-Word Prediction Performance

Story	No. words	% nouns	% verbs	% adjs & advs	% function words
1	562	21.17%	16.01%	8.72%	50.18%
2	464	16.81%	16.59%	12.93%	50.43%
3	491	15.68%	14.46%	10.59%	55.60%
4	658	15.35%	15.96%	13.53%	52.28%
5	774	15.63%	17.83%	11.50%	50.39%
Total	2949	16.82%	16.31%	11.50%	51.64%

Table 1: Human next-word prediction performance. % shows the percentage of correctly predicted words for each story and part of speech.

Bibliography

- [1] Philip B Gough and William E Tunmer. “Decoding, reading, and reading disability”. In: *Remedial and special education* 7.1 (1986), pp. 6–10 (cit. on p. 3).
- [2] Wesley A Hoover and Philip B Gough. “The simple view of reading”. In: *Reading and writing* 2 (1990), pp. 127–160 (cit. on p. 3).
- [3] Maryellen C MacDonald, Neal J Pearlmutter, and Mark S Seidenberg. “The lexical nature of syntactic ambiguity resolution.” In: *Psychological review* 101.4 (1994), p. 676 (cit. on p. 4).
- [4] Maria Novary Ngabut. “Reading theories and reading comprehension”. In: *Journal on English as a Foreign Language* 5.1 (2015), pp. 25–36 (cit. on p. 4).
- [5] Kara D Federmeier. “Thinking ahead: The role and roots of prediction in language comprehension”. In: *Psychophysiology* 44.4 (2007), pp. 491–505 (cit. on p. 4).
- [6] Gerry TM Altmann and Jelena Mirković. “Incrementality and prediction in human sentence processing”. In: *Cognitive science* 33.4 (2009), pp. 583–609 (cit. on pp. 4, 77).
- [7] Gina R Kuperberg and T Florian Jaeger. “What do we mean by prediction in language comprehension?” In: *Language, cognition and neuroscience* 31.1 (2016), pp. 32–59 (cit. on pp. 4, 158).

- [8] Charles A Perfetti and Lesley Hart. “The lexical basis of comprehension skill.” In: (2001) (cit. on p. 4).
- [9] David E Rumelhart. “Schemata: The building blocks of cognition”. In: *Theoretical issues in reading comprehension*. Routledge, 2017, pp. 33–58 (cit. on p. 4).
- [10] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018) (cit. on pp. 4, 19).
- [11] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9 (cit. on pp. 4, 9, 19, 21).
- [12] Yinhan Liu et al. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019) (cit. on pp. 4, 19).
- [13] Claude Elwood Shannon. “A mathematical theory of communication”. In: *The Bell system technical journal* 27.3 (1948), pp. 379–423 (cit. on pp. 4, 8, 24).
- [14] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *nature* 323.6088 (1986), pp. 533–536 (cit. on p. 4).
- [15] Alec Radford et al. “Improving language understanding by generative pre-training”. In: (2018) (cit. on pp. 4, 20).
- [16] Tom Brown et al. “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf (cit. on p. 4).
- [17] Matthew E. Peters et al. “Deep Contextualized Word Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

- Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: <https://aclanthology.org/N18-1202> (cit. on p. 4).
- [18] Alex Wang et al. “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Tal Linzen, Grzegorz Chrupała, and Afra Alishahi. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: <https://aclanthology.org/W18-5446> (cit. on p. 4).
- [19] James WA Strachan et al. “Testing theory of mind in large language models and humans”. In: *Nature Human Behaviour* (2024), pp. 1–11 (cit. on pp. 4, 5).
- [20] Cameron R Jones, Sean Trott, and Benjamin Bergen. “Comparing Humans and Large Language Models on an Experimental Protocol Inventory for Theory of Mind Evaluation (EPITOME)”. In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 803–819 (cit. on p. 5).
- [21] Yufei Tian et al. “Are Large Language Models Capable of Generating Human-Level Narratives?” In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024, pp. 17659–17681 (cit. on p. 5).
- [22] Jean-Rémi King et al. “Encoding and decoding neuronal dynamics: Methodological framework to uncover the algorithms of cognition”. In: (2018) (cit. on p. 5).
- [23] Juliette Millet and Jean-Remi King. “Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech”. In: *arXiv preprint arXiv:2103.01032* (2021) (cit. on p. 5).

- [24] Omar Chehab et al. “Deep Recurrent Encoder: A scalable end-to-end network to model brain signals”. In: *arXiv preprint arXiv:2103.02339* (2021) (cit. on p. 5).
- [25] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. “Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects”. In: *EMNLP 2021-Conference on Empirical Methods in Natural Language Processing*. 2021 (cit. on p. 5).
- [26] Aditya R Vaidya, Shailee Jain, and Alexander G Huth. “Self-supervised models of audio effectively explain human cortical responses to speech”. In: *arXiv preprint arXiv:2205.14252* (2022) (cit. on p. 5).
- [27] Frank Keller. “Cognitively plausible models of human language processing”. In: *Proceedings of the ACL 2010 Conference Short Papers*. 2010, pp. 60–67 (cit. on p. 5).
- [28] Alexander G Huth et al. “Natural speech reveals the semantic maps that tile human cerebral cortex”. In: *Nature* 532.7600 (2016), pp. 453–458 (cit. on pp. 6, 30).
- [29] Jan-Mathijs Schoffelen et al. “A 204-subject multimodal neuroimaging dataset to study language processing”. In: *Scientific data* 6.1 (2019), pp. 1–13 (cit. on pp. 6, 30).
- [30] Micha Heilbron et al. “Word contexts enhance the neural representation of individual letters in early visual cortex”. In: *Nature communications* 11.1 (2020), pp. 1–11 (cit. on pp. 6, 30).
- [31] Peter Hagoort. “How the brain solves the binding problem for language: a neurocomputational model of syntactic processing”. In: *Neuroimage* 20 (2003), S18–S29 (cit. on p. 6).
- [32] Gregory Hickok and David Poeppel. “The cortical organization of speech processing”. In: *Nature reviews neuroscience* 8.5 (2007), pp. 393–402 (cit. on pp. 6, 110).

- [33] Angela D Friederici. “The brain basis of language processing: from structure to function”. In: *Physiological reviews* 91.4 (2011), pp. 1357–1392 (cit. on p. 6).
- [34] Helen J Neville, Debra L Mills, and Donald S Lawson. “Fractionating language: Different neural subsystems with different sensitive periods”. In: *Cerebral cortex* 2.3 (1992), pp. 244–258 (cit. on pp. 7, 91, 92, 108).
- [35] Thomas F Münte et al. “Differences in brain potentials to open and closed class words: Class and frequency effects”. In: *Neuropsychologia* 39.1 (2001), pp. 91–102 (cit. on pp. 7, 91, 92).
- [36] Lee Osterhout, Mark Allen, and Judith McLaughlin. “Words in the brain: lexical determinants of word-induced brain activity”. In: *Journal of Neurolinguistics* 15.3-5 (2002), pp. 171–187 (cit. on pp. 7, 91, 92, 106, 114).
- [37] Wilson L Taylor. ““Cloze procedure”: A new tool for measuring readability”. In: *Journalism quarterly* 30.4 (1953), pp. 415–433 (cit. on pp. 8, 25, 130).
- [38] Sid Black et al. “Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow”. In: *If you use this software, please cite it using these meta-data* 58.2 (2021) (cit. on pp. 9, 22).
- [39] Boi Mai Quach, Cathal Gurrin, and Graham Healy. “DERCo: A Dataset for Human Behaviour in Reading Comprehension Using EEG”. In: *Scientific Data* 11.1 (2024), p. 1104 (cit. on pp. 9, 53, 77, 79, 93, 94, 126, 127).
- [40] Hartmut Fitz and Franklin Chang. “Language ERPs reflect learning through prediction error propagation”. In: *Cognitive Psychology* 111 (2019), pp. 15–52 (cit. on pp. 15, 17).
- [41] A Guyton and J Hall. “Textbook of medical physiology, Elsevier Inc”. In: *Philadelphia PA* (2006) (cit. on p. 16).
- [42] Hallowell Davis et al. “Electrical reactions of the human brain to auditory stimulation during sleep”. In: *Journal of Neurophysiology* 2.6 (1939), pp. 500–514 (cit. on p. 16).

- [43] Dragoljub Gajic et al. “Detection of epileptiform activity in EEG signals based on time-frequency and non-linear analysis”. In: *Frontiers in computational neuroscience* 9 (2015), p. 38 (cit. on p. 17).
- [44] Melissa L-H Võ and Jeremy M Wolfe. “Differential ERP signatures elicited by semantic and syntactic processing in scenes”. In: *Psychological Science* 24.9 (2013), p. 1816 (cit. on p. 17).
- [45] Jos JA Van Berkum et al. “Anticipating upcoming words in discourse: evidence from ERPs and reading times.” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31.3 (2005), p. 443 (cit. on p. 17).
- [46] Ellen Lau et al. “A lexical basis for N400 context effects: Evidence from MEG”. In: *Brain and language* 111.3 (2009), pp. 161–172 (cit. on p. 17).
- [47] Marta Kutas and Steven A Hillyard. “Brain potentials during reading reflect word expectancy and semantic association”. In: *Nature* 307.5947 (1984), pp. 161–163 (cit. on pp. 18, 25, 26, 81, 92, 106).
- [48] Cyma Van Petten and Marta Kutas. “Interactions between sentence context and word frequency in event-related brain potentials”. In: *Memory & cognition* 18.4 (1990), pp. 380–393 (cit. on p. 18).
- [49] Marta Kutas. “In the company of other words: Electrophysiological evidence for single-word and sentence context effects”. In: *Language and cognitive processes* 8.4 (1993), pp. 533–572 (cit. on pp. 18, 19).
- [50] Sarah Laszlo and Kara D Federmeier. “A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context”. In: *Journal of Memory and Language* 61.3 (2009), pp. 326–338 (cit. on p. 18).
- [51] Sarah Laszlo and Kara D Federmeier. “The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects”. In: *Psychophysiology* 48.2 (2011), pp. 176–186 (cit. on p. 18).

- [52] Marta Kutas and Kara D Federmeier. “Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP)”. In: *Annual review of psychology* 62 (2011), pp. 621–647 (cit. on pp. 18, 19, 26, 106, 111).
- [53] Christoph Aurnhammer et al. “Retrieval (N400) and integration (P600) in expectation-based comprehension”. In: *Plos one* 16.9 (2021), e0257430 (cit. on p. 18).
- [54] MIREILLE Besson, MARTA Kutas, and CYMA Van Petten. “ERP signs of semantic congruity and word repetition in sentences”. In: *Psychophysiological brain research* 1 (1990), pp. 259–262 (cit. on p. 18).
- [55] Cyma Van Petten and Maria Kutas. “Electrophysiological evidence for the flexibility of lexical processing”. In: *Advances in psychology*. Vol. 77. Elsevier, 1991, pp. 129–174 (cit. on p. 18).
- [56] Peter Hagoort and Jos van Berkum. “Beyond the sentence given”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1481 (2007), pp. 801–811 (cit. on p. 18).
- [57] Gina R Kuperberg, Trevor Brothers, and Edward W Wlotko. “A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation”. In: *Journal of cognitive neuroscience* 32.1 (2020), pp. 12–35 (cit. on pp. 18, 26, 106, 107).
- [58] Phillip J Holcomb, Sharon A Coffey, and Helen J Neville. “Visual and auditory sentence processing: A developmental analysis using event-related brain potentials”. In: *Developmental Neuropsychology* 8.2-3 (1992), pp. 203–241 (cit. on p. 19).
- [59] Kara D Federmeier et al. “Sounds, words, sentences: age-related changes across levels of language processing.” In: *Psychology and aging* 18.4 (2003), p. 858 (cit. on p. 19).

- [60] Aaron J Newman et al. “The influence of language proficiency on lexical semantic processing in native and late learners of English”. In: *Journal of cognitive neuroscience* 24.5 (2012), pp. 1205–1223 (cit. on p. 19).
- [61] Katarzyna Jankowiak and Karolina Rataj. “The N400 as a window into lexico-semantic processing in bilingualism”. In: *Poznan Studies in Contemporary Linguistics* 53.1 (2017), pp. 119–156 (cit. on p. 19).
- [62] Xavier Aparicio et al. “Language effects in trilinguals: an ERP study”. In: *Frontiers in psychology* 3 (2012), p. 402 (cit. on p. 19).
- [63] Vicente Iragui, Marta Kutas, and David P Salmon. “Event-related brain potentials during semantic categorization in normal aging and senile dementia of the Alzheimer’s type”. In: *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section* 100.5 (1996), pp. 392–406 (cit. on p. 19).
- [64] John M Olichney et al. “Cognitive event-related potentials: biomarkers of synaptic dysfunction across the stages of Alzheimer’s disease”. In: *Journal of Alzheimer’s Disease* 26.s3 (2011), pp. 215–228 (cit. on p. 19).
- [65] Marilyne Joyal et al. “Semantic processing in healthy aging and Alzheimer’s disease: a systematic review of the N400 differences”. In: *Brain sciences* 10.11 (2020), p. 770 (cit. on p. 19).
- [66] Namita Kumar and J Bruno Debruille. “Semantics and N400: insights for schizophrenia”. In: *Journal of Psychiatry and Neuroscience* 29.2 (2004), pp. 89–98 (cit. on p. 19).
- [67] Kui Wang et al. “Semantic processing disturbance in patients with schizophrenia: a meta-analysis of the N400 component”. In: *PLoS One* 6.10 (2011), e25435 (cit. on p. 19).
- [68] Kevin Clark et al. “Electra: Pre-training text encoders as discriminators rather than generators”. In: *arXiv preprint arXiv:2003.10555* (2020) (cit. on p. 19).

- [69] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 20).
- [70] Mark Chen et al. “Generative pretraining from pixels”. In: *International conference on machine learning*. PMLR. 2020, pp. 1691–1703 (cit. on p. 21).
- [71] Yu-An Chung and James Glass. “Generative pre-training for speech with autoregressive predictive coding”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 3497–3501 (cit. on p. 21).
- [72] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. “ProtGPT2 is a deep unsupervised language model for protein design”. In: *Nature communications* 13.1 (2022), p. 4348 (cit. on p. 21).
- [73] Yingcong Li et al. “Mechanics of next token prediction with self-attention”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 685–693 (cit. on pp. 21, 22).
- [74] Woon Sang Cho et al. “Towards Coherent and Cohesive Long-form Text Generation”. In: *Proceedings of the First Workshop on Narrative Understanding*. 2019, pp. 1–11 (cit. on p. 23).
- [75] Tal Linzen and Marco Baroni. “Syntactic structure from deep learning”. In: *Annual Review of Linguistics* 7.1 (2021), pp. 195–212 (cit. on p. 23).
- [76] Ariel Goldstein et al. “Shared computational principles for language processing in humans and deep language models”. In: *Nature neuroscience* 25.3 (2022), pp. 369–380 (cit. on pp. 24, 69, 74, 93, 125, 137, 158).
- [77] Clara Meister et al. “Revisiting the Uniform Information Density Hypothesis”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 963–980 (cit. on pp. 24, 158).
- [78] Micha Heilbron et al. “A hierarchy of linguistic predictions during natural language comprehension”. In: *Proceedings of the National Academy of Sciences* 119.32 (2022), e2201968119 (cit. on pp. 24, 69, 72).

- [79] Cory Shain et al. “Large-scale evidence for logarithmic effects of word predictability on reading time”. In: *Proceedings of the National Academy of Sciences* 121.10 (2024), e2307876121 (cit. on pp. 24, 27, 141).
- [80] Moshe Bar. “The proactive brain: using analogies and associations to generate predictions”. In: *Trends in cognitive sciences* 11.7 (2007), pp. 280–289 (cit. on p. 24).
- [81] John Hale. “Information-theoretical complexity metrics”. In: *Language and Linguistics Compass* 10.9 (2016), pp. 397–412 (cit. on p. 24).
- [82] John Hale. “A probabilistic Earley parser as a psycholinguistic model”. In: *Second meeting of the north american chapter of the association for computational linguistics*. 2001 (cit. on p. 25).
- [83] Roger Levy. “Expectation-based syntactic comprehension”. In: *Cognition* 106.3 (2008), pp. 1126–1177 (cit. on p. 25).
- [84] Cyma Van Petten and Barbara J Luka. “Prediction during language comprehension: Benefits, costs, and ERP components”. In: *International journal of psychophysiology* 83.2 (2012), pp. 176–190 (cit. on pp. 25, 90).
- [85] Trevor Brothers and Gina R Kuperberg. “Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension”. In: *Journal of Memory and Language* 116 (2021), p. 104174 (cit. on pp. 25, 26, 158).
- [86] Nathaniel Smith and Roger Levy. “Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing”. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 33. 33. 2011 (cit. on p. 26).
- [87] Nathaniel J Smith and Roger Levy. “The effect of word predictability on reading time is logarithmic”. In: *Cognition* 128.3 (2013), pp. 302–319 (cit. on pp. 26, 158).

- [88] James Michaelov and Benjamin Bergen. “How well does surprisal explain N400 amplitude under different experimental conditions?” In: *Proceedings of the 24th Conference on Computational Natural Language Learning*. 2020, pp. 652–663 (cit. on pp. 26, 148, 149).
- [89] Stefan L Frank et al. “Word surprisal predicts N400 amplitude during reading”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2013, pp. 878–883 (cit. on pp. 26, 134, 148).
- [90] Alma Lindborg et al. “Semantic surprise predicts the N400 brain potential”. In: *Neuroimage: Reports* 3.1 (2023), p. 100161 (cit. on p. 26).
- [91] Jeffrey L Elman. “Distributed representations, simple recurrent networks, and grammatical structure”. In: *Machine learning* 7 (1991), pp. 195–225 (cit. on p. 26).
- [92] Kyunghyun Cho et al. “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. DOI: 10.3115/v1/D14-1179. URL: <https://aclanthology.org/D14-1179> (cit. on p. 26).
- [93] S Hochreiter. “Long Short-term Memory”. In: *Neural Computation MIT-Press* (1997) (cit. on p. 26).
- [94] Adam Goodkind and Klinton Bicknell. “Predictive power of word surprisal for reading times is a linear function of language model quality”. In: *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*. 2018, pp. 10–18 (cit. on p. 26).
- [95] Ethan Gotlieb Wilcox et al. “On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior”. In: *Proceedings of*

- the 42nd Annual Meeting of the Cognitive Science Society*. 2020, pp. 1707–1713 (cit. on p. 27).
- [96] Yiding Hao et al. “Probabilistic Predictions of People Perusing: Evaluating Metrics of Language Model Performance for Psycholinguistic Modeling”. In: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. 2020, pp. 75–86 (cit. on pp. 27, 134).
- [97] Danny Merks and Stefan L Frank. “Human sentence processing: Recurrence or attention?” In: *Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2021)*. Association for Computational Linguistics (ACL). 2021, pp. 12–22 (cit. on p. 27).
- [98] Martin Schrimpf et al. “The neural architecture of language: Integrative modeling converges on predictive processing”. In: *Proceedings of the National Academy of Sciences* 118.45 (2021), e2105646118 (cit. on p. 27).
- [99] Stefan L Frank et al. “The ERP response to the amount of information conveyed by words in sentences”. In: *Brain and language* 140 (2015), pp. 1–11 (cit. on pp. 27, 75, 148).
- [100] Roel M Willems et al. “Prediction during natural language comprehension”. In: *Cerebral Cortex* 26.6 (2016), pp. 2506–2516 (cit. on p. 27).
- [101] Kristijan Armeni et al. “Frequency-specific brain dynamics related to prediction during language comprehension”. In: *NeuroImage* 198 (2019), pp. 283–295 (cit. on p. 27).
- [102] Micha Heilbron et al. “Tracking naturalistic linguistic predictions with deep neural language models”. In: *2019 Conference on Cognitive Computational Neuroscience (CCN 2019)*. 2019, pp. 424–427 (cit. on p. 27).
- [103] James A Michaelov et al. “Strong Prediction: Language model surprisal explains multiple N400 effects”. In: *Neurobiology of language* 5.1 (2024), pp. 107–135 (cit. on p. 27).

- [104] Zihang Dai et al. “Transformer-XL: Attentive Language Models beyond a Fixed-Length Context”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2978–2988 (cit. on p. 27).
- [105] Zhilin Yang et al. “XLNet: Generalized Autoregressive Pretraining for Language Understanding”. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. 2019 (cit. on p. 27).
- [106] Fatma Deniz et al. “The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality”. In: *Journal of Neuroscience* 39.39 (2019), pp. 7722–7736 (cit. on pp. 30, 68).
- [107] Marie T Banich and Molly Mack. *Mind, brain, and language: multidisciplinary perspectives*. Taylor & Francis, 2003 (cit. on p. 31).
- [108] Florian Hutzler et al. “Parafoveal X-masks interfere with foveal word recognition: Evidence from fixation-related brain potentials”. In: *Frontiers in Systems Neuroscience* 7 (2013), p. 33 (cit. on p. 32).
- [109] Andrey R Nikolaev, Radha Nila Meghanathan, and Cees van Leeuwen. “Combining EEG and eye movement recording in free viewing: Pitfalls and possibilities”. In: *Brain and cognition* 107 (2016), pp. 55–83 (cit. on p. 32).
- [110] Aline Frey et al. “An eye fixation-related potential study in two Reading tasks: Reading to memorize and Reading to make a decision”. In: *Brain topography* 31.4 (2018), pp. 640–660 (cit. on p. 32).
- [111] Olaf Dimigen and Benedikt V Ehinger. “Regression-based analysis of combined EEG and eye-tracking data: Theory and applications”. In: *Journal of Vision* 21.1 (2021), pp. 3–3 (cit. on pp. 32, 37, 87).

- [112] Michael Dambacher and Reinhold Kliegl. “Synchronizing timelines: Relations between fixation durations and N400 amplitudes during sentence reading”. In: *Brain research* 1155 (2007), pp. 147–162 (cit. on pp. 32, 54).
- [113] Lisa Spitzer and Stefanie Mueller. “Using a test battery to compare three remote, video-based eye-trackers”. In: *2022 Symposium on Eye Tracking Research and Applications*. 2022, pp. 1–7 (cit. on p. 32).
- [114] Keith Rayner et al. “Eye movements and on-line language comprehension processes”. In: *Language and Cognitive Processes* 4.3-4 (1989), SI21–SI49 (cit. on p. 32).
- [115] Keith Rayner. “Eye movements in reading and information processing: 20 years of research.” In: *Psychological bulletin* 124.3 (1998), p. 372 (cit. on pp. 32, 33).
- [116] Mary C Potter. “Rapid serial visual presentation (RSVP): A method for studying language processing”. In: *New methods in reading comprehension research*. Routledge, 2018, pp. 91–118 (cit. on p. 32).
- [117] James F Juola et al. “Reading with and without eye movements”. In: *Eye movements from physiology to cognition*. Elsevier, 1987, pp. 499–508 (cit. on p. 32).
- [118] Thomas Kosch et al. “One does not simply RSVP: mental workload to select speed reading parameters using electroencephalography”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–13 (cit. on pp. 33, 70).
- [119] Michael Dambacher et al. “Stimulus onset asynchrony and the timeline of word recognition: Event-related potentials during sentence reading”. In: *Neuropsychologia* 50.8 (2012), pp. 1852–1870 (cit. on pp. 33, 70).
- [120] Horacio A Barber et al. “Parafoveal perception during sentence reading? An ERP paradigm using rapid serial visual presentation (RSVP) with flankers”. In: *Psychophysiology* 48.4 (2011), pp. 523–531 (cit. on p. 34).

- [121] Horacio A Barber, Maartje van der Meij, and Marta Kutas. “An electrophysiological analysis of contextual and temporal constraints on parafoveal word processing”. In: *Psychophysiology* 50.1 (2013), pp. 48–59 (cit. on pp. 34, 36, 70).
- [122] Wenjia Zhang et al. “Integration of sentence-level semantic information in parafovea: Evidence from the RSVP-flanker paradigm”. In: *PloS one* 10.9 (2015), e0139016 (cit. on p. 34).
- [123] Nan Li et al. “Parafoveal words can modulate sentence meaning: Electrophysiological evidence from an RSVP-with-flanker task”. In: *bioRxiv* (2021) (cit. on pp. 34–36).
- [124] Mallory C Stites, Brennan R Payne, and Kara D Federmeier. “Getting ahead of yourself: Parafoveal word expectancy modulates the N400 during sentence reading”. In: *Cognitive, Affective, & Behavioral Neuroscience* 17.3 (2017), pp. 475–490 (cit. on pp. 34, 35).
- [125] Brennan R Payne, Mallory C Stites, and Kara D Federmeier. “Event-related brain potentials reveal how multiple aspects of semantic processing unfold across parafoveal and foveal vision during sentence reading”. In: *Psychophysiology* 56.10 (2019), e13432 (cit. on p. 34).
- [126] Horacio A Barber et al. “Parafoveal N400 effect during sentence reading”. In: *Neuroscience letters* 479.2 (2010), pp. 152–156 (cit. on p. 36).
- [127] Elizabeth R Schotter. “Reading ahead by hedging our bets on seeing the future: Eye tracking and electrophysiology evidence for parafoveal lexical processing and saccadic control by partial word recognition”. In: *Psychology of Learning and Motivation*. Vol. 68. Elsevier, 2018, pp. 263–298 (cit. on p. 37).
- [128] Adrian Staub. “The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation”. In: *Language and Linguistics Compass* 9.8 (2015), pp. 311–327 (cit. on pp. 37, 87).

- [129] Denis Drieghe et al. “Word skipping in reading: On the interplay of linguistic and visual factors”. In: *European Journal of Cognitive Psychology* 16.1-2 (2004), pp. 79–103 (cit. on pp. 37, 87).
- [130] Benthe Kornrumpf et al. “Neural correlates of word recognition: A systematic comparison of natural reading and rapid serial visual presentation”. In: *Journal of Cognitive Neuroscience* 28.9 (2016), pp. 1374–1391 (cit. on p. 37).
- [131] Gregory A Light et al. “Electroencephalography (EEG) and event-related potentials (ERPs) with human participants”. In: *Current protocols in neuroscience* 52.1 (2010), pp. 6–25 (cit. on p. 38).
- [132] Mike X Cohen. *Analyzing neural time series data: theory and practice*. MIT press, 2014 (cit. on p. 38).
- [133] Aamir Saeed Malik and Hafeez Ullah Amin. *Designing EEG experiments for studying the brain: Design code and example datasets*. Academic Press, 2017 (cit. on p. 38).
- [134] Patrick Ledwidge, Jeremy Foust, and Adam Ramsey. “Recommendations for developing an EEG laboratory at a primarily undergraduate institution”. In: *Journal of Undergraduate Neuroscience Education* 17.1 (2018), A10 (cit. on p. 38).
- [135] Fabio Richlan et al. “A new high-speed visual stimulation method for gaze-contingent eye movement and brain activity studies”. In: *Frontiers in systems neuroscience* 7 (2013), p. 24 (cit. on p. 40).
- [136] Eero Ahtola et al. “Use of eye tracking improves the detection of evoked responses to complex visual stimuli during EEG in infants”. In: *Clinical neurophysiology practice* 2 (2017), pp. 81–90 (cit. on p. 40).
- [137] Charles E Davis, Jacob G Martin, and Simon J Thorpe. “Stimulus Onset Hub: An open-source, low latency, and opto-isolated trigger box for neuroscientific research replicability and beyond”. In: *Frontiers in Neuroinformatics* 14 (2020), p. 2 (cit. on p. 40).

- [138] Andreas Schwarz et al. “Analyzing and decoding natural reach-and-grasp actions using gel, water and dry EEG systems”. In: *Frontiers in neuroscience* 14 (2020), p. 849 (cit. on p. 40).
- [139] Mahsa Soufineyestani, Dale Dowling, and Arshia Khan. “Electroencephalography (EEG) technology applications and available devices”. In: *Applied Sciences* 10.21 (2020), p. 7453 (cit. on p. 41).
- [140] Md Kafiul Islam, Amir Rastegarnia, and Zhi Yang. “Methods for artifact detection and removal from scalp EEG: A review”. In: *Neurophysiologie Clinique/Clinical Neurophysiology* 46.4-5 (2016), pp. 287–305 (cit. on p. 42).
- [141] Xiao Jiang, Gui-Bin Bian, and Zean Tian. “Removal of artifacts from EEG signals: a review”. In: *Sensors* 19.5 (2019), p. 987 (cit. on pp. 43, 50).
- [142] Hong Zeng et al. “EOG artifact correction from EEG recording using stationary subspace analysis and empirical mode decomposition”. In: *Sensors* 13.11 (2013), pp. 14839–14859 (cit. on p. 43).
- [143] Christian J Ochoa and John Polich. “P300 and blink instructions”. In: *Clinical Neurophysiology* 111.1 (2000), pp. 93–98 (cit. on p. 43).
- [144] Irina I Goncharova et al. “EMG contamination of EEG: spectral and topographical characteristics”. In: *Clinical neurophysiology* 114.9 (2003), pp. 1580–1593 (cit. on p. 44).
- [145] Mehrdad Fatourechhi et al. “EMG and EOG artifacts in brain computer interface systems: A survey”. In: *Clinical neurophysiology* 118.3 (2007), pp. 480–494 (cit. on p. 44).
- [146] Christine Carl et al. “The saccadic spike artifact in MEG”. In: *Neuroimage* 59.2 (2012), pp. 1657–1667 (cit. on p. 44).
- [147] Shrawan Kumar, Yogesh Narayan, and Tyler Amell. “Power spectra of sternocleidomastoids, splenius capitis, and upper trapezius in oblique exertions”. In: *The Spine Journal* 3.5 (2003), pp. 339–350 (cit. on p. 44).

- [148] Emma M Whitham et al. “Scalp electrical recording during paralysis: quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG”. In: *Clinical neurophysiology* 118.8 (2007), pp. 1877–1888 (cit. on p. 44).
- [149] Emma M Whitham et al. “Thinking activates EMG in scalp electrical recordings”. In: *Clinical neurophysiology* 119.5 (2008), pp. 1166–1175 (cit. on p. 44).
- [150] Xun Chen et al. “Removing muscle artifacts from EEG data: Multichannel or single-channel techniques?” In: *IEEE Sensors Journal* 16.7 (2015), pp. 1986–1997 (cit. on p. 44).
- [151] Abdul Qayoom Hamal and Abdul Wahab bin Abdul Rehman. “Artifact Processing of Epileptic EEG Signals: An Overview of Different Types of Artifacts”. In: *2013 International Conference on Advanced Computer Science Applications and Technologies*. IEEE. 2013, pp. 358–361 (cit. on p. 45).
- [152] Gabriella Tamburro, David B Stone, and Silvia Comani. “Automatic Removal of Cardiac Interference (ARCI): a new approach for EEG data”. In: *Frontiers in neuroscience* 13 (2019), p. 441 (cit. on p. 45).
- [153] Mehdi Bagheri Hamaneh et al. “Automated removal of EKG artifact from EEG data using independent component analysis and continuous wavelet transformation”. In: *IEEE Transactions on Biomedical Engineering* 61.6 (2013), pp. 1634–1641 (cit. on pp. 45, 50).
- [154] Nadia Mammone and Francesco C Morabito. “Enhanced automatic wavelet independent component analysis for electroencephalographic artifact removal”. In: *Entropy* 16.12 (2014), pp. 6553–6572 (cit. on p. 45).
- [155] Rajesh Patel et al. “Common methodology for cardiac and ocular artifact suppression from EEG recordings by combining ensemble empirical mode decomposition with regression approach”. In: *Journal of Medical and Biological Engineering* 37.2 (2017), pp. 201–208 (cit. on p. 45).

- [156] Chinmayee Dora and Pradyut Kumar Biswal. “Efficient detection and correction of variable strength ECG artifact from single channel EEG”. In: *Biomedical Signal Processing and Control* 50 (2019), pp. 168–177 (cit. on p. 45).
- [157] Malik Muhammad Naeem Mannan, Muhammad Ahmad Kamran, and Myung Yung Jeong. “Identification and removal of physiological artifacts from electroencephalogram signals: A review”. In: *Ieee Access* 6 (2018), pp. 30630–30652 (cit. on p. 45).
- [158] Steven J Luck. *An introduction to the event-related potential technique*. MIT press, 2014 (cit. on pp. 45, 46, 52, 83, 92, 93).
- [159] Alain de Cheveigné and Israel Nelken. “Filters: when, why, and how (not) to use them”. In: *Neuron* 102.2 (2019), pp. 280–293 (cit. on p. 45).
- [160] Leyla Baysal Kirac et al. “Notch filter artefact mimicking high frequency oscillation in epilepsy”. In: *Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology* 127.1 (2016), pp. 979–981 (cit. on p. 45).
- [161] Mohammad Reza Keshtkaran and Zhi Yang. “A fast, robust algorithm for power line interference cancellation in neural recording”. In: *Journal of neural engineering* 11.2 (2014), p. 026017 (cit. on p. 45).
- [162] Sabine Leske and Sarang S Dalal. “Reducing power line noise in EEG and MEG data via spectrum interpolation”. In: *Neuroimage* 189 (2019), pp. 763–776 (cit. on p. 45).
- [163] Alain De Cheveigné and Jonathan Z Simon. “Denoising based on time-shift PCA”. In: *Journal of neuroscience methods* 165.2 (2007), pp. 297–305 (cit. on p. 45).
- [164] Javier Escudero et al. “Artifact removal in magnetoencephalogram background activity with independent component analysis”. In: *IEEE Transactions on Biomedical Engineering* 54.11 (2007), pp. 1965–1973 (cit. on p. 46).

- [165] Alain de Cheveigné and Lucas C Parra. “Joint decorrelation, a versatile tool for multichannel data analysis”. In: *Neuroimage* 98 (2014), pp. 487–505 (cit. on p. 46).
- [166] Libo Zhang et al. “Demystifying signal processing techniques to extract task-related EEG responses for psychologists”. In: *Brain Science Advances* 6.3 (2020), pp. 171–188 (cit. on p. 46).
- [167] Arnaud Delorme and Jeffery A Martin. “Automated data cleaning for the Muse EEG”. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2021, pp. 1–5 (cit. on p. 47).
- [168] Hugh Nolan, Robert Whelan, and Richard B Reilly. “FASTER: fully automated statistical thresholding for EEG artifact rejection”. In: *Journal of neuroscience methods* 192.1 (2010), pp. 152–162 (cit. on pp. 47, 48, 77).
- [169] Alexandre Barachant, Anton Andreev, and Marco Congedo. “The Riemannian Potato: an automatic and adaptive artifact detection method for on-line experiments using Riemannian geometry”. In: *TOBI Workshop IV*. 2013, pp. 19–20 (cit. on pp. 47, 49).
- [170] Mainak Jas et al. “Autoreject: Automated artifact rejection for MEG and EEG data”. In: *NeuroImage* 159 (2017), pp. 417–429 (cit. on pp. 47, 48, 78).
- [171] Luca Pion-Tonachini, Ken Kreutz-Delgado, and Scott Makeig. “ICLabel: An automated electroencephalographic independent component classifier, dataset, and website”. In: *NeuroImage* 198 (2019), pp. 181–197 (cit. on pp. 47, 51, 78).
- [172] Florian Yger, Fabien Lotte, and Masashi Sugiyama. “Averaging covariance matrices for EEG signal classification based on the CSP: an empirical study”. In: *2015 23rd European signal processing conference (EUSIPCO)*. IEEE. 2015, pp. 2721–2725 (cit. on p. 49).
- [173] Inbal Horev, Florian Yger, and Masashi Sugiyama. “Geometry-aware principal component analysis for symmetric positive definite matrices”. In: *Asian Conference on Machine Learning*. PMLR. 2016, pp. 1–16 (cit. on p. 49).

- [174] Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. “Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review”. In: *Brain-Computer Interfaces* 4.3 (2017), pp. 155–174 (cit. on p. 49).
- [175] F Perrin et al. “Mapping of scalp potentials by surface spline interpolation”. In: *Electroencephalography and clinical neurophysiology* 66.1 (1987), pp. 75–81 (cit. on p. 49).
- [176] EM Fletcher, CL Kussmaul, and George R Mangun. “Estimation of interpolation errors in scalp topographic mapping”. In: *Electroencephalography and clinical Neurophysiology* 98.5 (1996), pp. 422–434 (cit. on p. 49).
- [177] H Ashida et al. “Field mapping of EEG by unbiased polynomial interpolation”. In: *Computers and biomedical research* 17.3 (1984), pp. 267–276 (cit. on p. 49).
- [178] Janin Jäger et al. “Reconstruction of electroencephalographic data using radial basis functions”. In: *Clinical Neurophysiology* 127.4 (2016), pp. 1978–1983 (cit. on p. 49).
- [179] Sari Saba-Sadiya et al. “EEG channel interpolation using deep encoder-decoder networks”. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2020, pp. 2432–2439 (cit. on p. 49).
- [180] Aapo Hyvärinen and Erkki Oja. “Independent component analysis: algorithms and applications”. In: *Neural networks* 13.4-5 (2000), pp. 411–430 (cit. on pp. 50, 51, 78).
- [181] Arthur Flexer et al. “Using ICA for removal of ocular artifacts in EEG recorded from blind subjects”. In: *Neural Networks* 18.7 (2005), pp. 998–1005 (cit. on p. 50).
- [182] Olaf Dimigen. “Optimizing the ICA-based removal of ocular EEG artifacts from free viewing experiments”. In: *NeuroImage* 207 (2020), p. 116117 (cit. on p. 50).

- [183] Weidong Zhou and Jean Gotman. “Removal of EMG and ECG artifacts from EEG based on wavelet transform and ICA”. In: *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vol. 1. IEEE. 2004, pp. 392–395 (cit. on p. 50).
- [184] Maite Crespo-Garcia, Mercedes Atienza, and Jose L Cantero. “Muscle artifact removal from human sleep EEG by using independent component analysis”. In: *Annals of biomedical engineering* 36.3 (2008), pp. 467–475 (cit. on p. 50).
- [185] Stéphanie Devuyst et al. “Cancelling ECG artifacts in EEG using a modified independent component analysis approach”. In: *EURASIP Journal on advances in signal processing* 2008 (2008), pp. 1–13 (cit. on p. 50).
- [186] Samuel Akwei-Sekyere. “Powerline noise elimination in biomedical signals via blind source separation and wavelet analysis”. In: *PeerJ* 3 (2015), e1086 (cit. on p. 50).
- [187] Zhaojun Xue et al. “Using ICA to remove eye blink and power line artifacts in EEG”. In: *First International Conference on Innovative Computing, Information and Control-Volume I (ICICIC’06)*. Vol. 3. IEEE. 2006, pp. 107–110 (cit. on p. 50).
- [188] Izabela Rejer and Paweł Górski. “Independent Component Analysis for EEG data preprocessing-algorithms comparison”. In: *IFIP International Conference on Computer Information Systems and Industrial Management*. Springer. 2013, pp. 108–119 (cit. on p. 51).
- [189] Michael Plöchl, José P Ossandón, and Peter König. “Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data”. In: *Frontiers in human neuroscience* 6 (2012), p. 278 (cit. on p. 54).
- [190] Marc Brysbaert. “How many words do we read per minute? A review and meta-analysis of reading rate”. In: *Journal of memory and language* 109 (2019), p. 104047 (cit. on p. 54).

- [191] *PsychoPy: The Open-Source Software for Running Psychology Experiments*. <https://www.psychopy.org/> (cit. on p. 55).
- [192] Brain Products. *ActiCHamp*. <https://www.brainproducts.com/solutions/actichamp/> (cit. on p. 58).
- [193] George H Klem. “The ten-twenty electrode system of the international federation. The international federation of clinical neurophysiology”. In: *Electroencephalogr. Clin. Neurophysiol. Suppl.* 52 (1999), pp. 3–6 (cit. on pp. 58, 94).
- [194] Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. “Faster independent component analysis by preconditioning with Hessian approximations”. In: *IEEE Transactions on Signal Processing* 66.15 (2018), pp. 4040–4049 (cit. on pp. 58, 78).
- [195] Indra Mohan Chakravarti, Radha Govira Laha, and Jogabrata Roy. “Handbook of methods of applied statistics”. In: *Wiley Series in Probability and Mathematical Statistics (USA) eng* (1967) (cit. on p. 59).
- [196] Sendy Caffarra et al. “Second language syntactic processing revealed through event-related potentials: An empirical review”. In: *Neuroscience & Biobehavioral Reviews* 51 (2015), pp. 31–47 (cit. on p. 65).
- [197] Anna M Beres. “Time is of the essence: A review of electroencephalography (EEG) and event-related brain potentials (ERPs) in language research”. In: *Applied psychophysiology and biofeedback* 42 (2017), pp. 247–255 (cit. on p. 65).
- [198] Ina Bornkessel-Schlesewsky and Matthias Schlewsky. “Toward a neurobiologically plausible model of language-related, negative event-related potentials”. In: *Frontiers in psychology* 10 (2019), p. 298 (cit. on p. 65).
- [199] Olaf Dimigen et al. “Coregistration of eye movements and EEG in natural reading: analyses and review.” In: *Journal of experimental psychology: General* 140.4 (2011), p. 552 (cit. on pp. 65, 83).

- [200] Federica Degno, Otto Loberg, and Simon P Liversedge. “Co-Registration of Eye Movements and Fixation—Related Potentials in Natural Reading: Practical Issues of Experimental Design and Data Analysis”. In: *Collabra: Psychology* 7.1 (2021) (cit. on p. 65).
- [201] Stéphane Dufau et al. “A thousand words are worth a picture: Snapshots of printed-word processing in an event-related potential megastudy”. In: *Psychological science* 26.12 (2015), pp. 1887–1897 (cit. on pp. 68, 69, 88).
- [202] C Ervin Davis et al. “Brain function with complex decision making using electroencephalography”. In: *International journal of psychophysiology* 79.2 (2011), pp. 175–183 (cit. on p. 68).
- [203] Christian Pfeiffer et al. “Neural dynamics of sentiment processing during naturalistic sentence reading”. In: *NeuroImage* 218 (2020), p. 116934 (cit. on p. 69).
- [204] Alex Murphy et al. “Decoding part-of-speech from human EEG signals”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022, pp. 2201–2210 (cit. on pp. 69, 93).
- [205] Martín Antúnez et al. “Semantic parafoveal processing in natural reading: Insight from fixation-related potentials & eye movements”. In: *Psychophysiology* 59.4 (2022), e13986 (cit. on p. 69).
- [206] Melissa Troyer et al. “Nuances of knowing: Brain potentials reveal implicit effects of domain knowledge on word processing in the absence of sentence-level knowledge”. In: *Psychophysiology* (2023), e14422 (cit. on p. 69).
- [207] Yiqun Duan et al. “DeWave: Discrete Encoding of EEG Waves for EEG to Text Translation”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023 (cit. on pp. 69, 159).
- [208] Taiqi He et al. “Neural correlates of word representation vectors in natural language processing models: Evidence from representational similarity analy-

- sis of event-related brain potentials”. In: *Psychophysiology* 59.3 (2022), e13976 (cit. on pp. 69, 91, 137).
- [209] Yina Guo et al. “End-to-end translation of human neural activity to speech with a dual–dual generative adversarial network”. In: *Knowledge-based systems* 277 (2023), p. 110837 (cit. on p. 69).
 - [210] Nora Hollenstein et al. “Decoding EEG brain activity for multi-modal natural language processing”. In: *Frontiers in Human Neuroscience* (2021), p. 378 (cit. on p. 69).
 - [211] Dan Schwartz and Tom Mitchell. “Understanding language-elicited EEG data by predicting it from a fine-tuned language model”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 43–57 (cit. on p. 69).
 - [212] Simone Sulpizio et al. “Very early and late form-to-meaning computations during visual word recognition as revealed by electrophysiology”. In: *Cortex* 157 (2022), pp. 167–193 (cit. on p. 69).
 - [213] *MNE Developers: Kiloword Dataset*. <https://mne.tools/stable/documentation/datasets.html> (cit. on p. 69).
 - [214] Nora Hollenstein et al. “ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading”. In: *Scientific data* 5.1 (2018), pp. 1–13 (cit. on pp. 69, 88).
 - [215] Nora Hollenstein et al. “ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation”. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. 2020, pp. 138–146 (cit. on p. 69).
 - [216] Nora Hollenstein et al. “Multilingual Language Models Predict Human Reading Behavior”. In: *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

- NAACL-HLT 2021*. Association for Computational Linguistics (ACL). 2021, pp. 106–123 (cit. on p. 69).
- [217] Zhenhailong Wang and Heng Ji. “Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 2022, pp. 5350–5358 (cit. on pp. 69, 159).
- [218] Steven J Luck and Emily S Kappenman. *The Oxford handbook of event-related potential components*. Oxford university press, 2013 (cit. on p. 70).
- [219] *Amazon Mechanical Turk*. <https://www.mturk.com/> (cit. on pp. 70, 73, 75).
- [220] Bruno Bianchi et al. “Human and computer estimations of Predictability of words in written language”. In: *Scientific reports* 10.1 (2020), p. 4396 (cit. on p. 70).
- [221] Hannah Murray and Sharif El-Leithy. “Behavioural experiments in cognitive therapy for posttraumatic stress disorder: why, when, and how?” In: *Verhaltenstherapie* 31.1 (2021), pp. 50–60 (cit. on p. 70).
- [222] Nathan Gagné and Léon Franzen. “How to run behavioural experiments online: Best practice suggestions for cognitive psychology and neuroscience”. In: *Swiss Psychology Open: the official journal of the Swiss Psychological Society* 3.1 (2023) (cit. on p. 70).
- [223] Shohini Bhattachali et al. “The Alice Datasets: fMRI & EEG observations of natural language comprehension”. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020, pp. 120–125 (cit. on pp. 72, 88).
- [224] Jixing Li et al. “Le Petit Prince multilingual naturalistic fMRI corpus”. In: *Scientific data* 9.1 (2022), p. 530 (cit. on pp. 72, 88).
- [225] Melissa Troyer and Marta Kutas. “Harry Potter and the Chamber of What?: The impact of what individuals know on word processing during reading”. In: *Language, cognition and neuroscience* 35.5 (2020), pp. 641–657 (cit. on pp. 72, 88).

- [226] Shenghao Wu, Aaditya Ramdas, and Leila Wehbe. “Brainprints: identifying individuals from magnetoencephalograms”. In: *Communications Biology* 5.1 (2022), p. 852 (cit. on p. 72).
- [227] Kristijan Armeni et al. “A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension”. In: *Scientific Data* 9.1 (2022), p. 278 (cit. on p. 72).
- [228] Joshua R De Leeuw. “jsPsych: A JavaScript library for creating behavioral experiments in a Web browser”. In: *Behavior research methods* 47 (2015), pp. 1–12 (cit. on p. 73).
- [229] Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. “Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data”. In: *Behavior Research Methods* (2023), pp. 1–24 (cit. on p. 75).
- [230] Stefan L Frank et al. “Reading time data for evaluating broad-coverage models of English sentence processing”. In: *Behavior research methods* 45 (2013), pp. 1182–1190 (cit. on p. 75).
- [231] Amazon Web Services. *Amazon Simple Storage Service (S3)*. <https://aws.amazon.com/s3/> (cit. on p. 76).
- [232] Marta Kutas, Katherine A DeLong, and Nathaniel J Smith. “A look around at what lies ahead: Prediction and predictability in language processing”. In: *Predictions in the brain: Using our past to generate a future* 190207.10.1093 (2011) (cit. on p. 77).
- [233] MNE Developers. *MNE-Python: Software for processing MEG and EEG data*. <https://mne.tools/1.4/index.html> (cit. on p. 77).
- [234] Tzyy-Ping Jung et al. “Removing electroencephalographic artifacts by blind source separation”. In: *Psychophysiology* 37.2 (2000), pp. 163–178 (cit. on p. 78).

- [235] Frank Smith. “The role of prediction in reading”. In: *Elementary English* 52.3 (1975), pp. 305–311 (cit. on p. 79).
- [236] Steven G Luke and Kiel Christianson. “Limits on lexical prediction during reading”. In: *Cognitive psychology* 88 (2016), pp. 22–60 (cit. on p. 81).
- [237] Jakub M Szewczyk and Kara D Federmeier. “Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability”. In: *Journal of memory and language* 123 (2022), p. 104311 (cit. on p. 81).
- [238] Riccardo Orlando. *IPA: Incremental Parsing and Alignment*. <https://github.com/Riccorl/ipa> (cit. on pp. 82, 94).
- [239] Oriano Mecarelli. “Electrode placement systems and montages”. In: *Clinical Electroencephalography* (2019), pp. 35–52 (cit. on p. 83).
- [240] Emily R Dickinson, Jill L Adelson, and Jesse Owen. “Gender balance, representativeness, and statistical power in sexuality research using undergraduate student samples”. In: *Archives of Sexual Behavior* 41 (2012), pp. 325–327 (cit. on p. 88).
- [241] Ellen F Lau, Phillip J Holcomb, and Gina R Kuperberg. “Dissociating N400 effects of prediction from association in single-word contexts”. In: *Journal of cognitive neuroscience* 25.3 (2013), pp. 484–502 (cit. on p. 90).
- [242] Alan Bell et al. “Predictability effects on durations of content and function words in conversational English”. In: *Journal of Memory and Language* 60.1 (2009), pp. 92–111 (cit. on p. 90).
- [243] Sara Dellantonio, Remo Job, and Claudio Mulatti. “Imageability: now you see it again (albeit in a different form)”. In: *Frontiers in psychology* 5 (2014), p. 279 (cit. on p. 90).
- [244] Simone Mätzig et al. “Noun and verb differences in picture naming: Past studies and new evidence”. In: *Cortex* 45.6 (2009), pp. 738–758 (cit. on p. 91).

- [245] Julie L Earles and Alan W Kersten. “Why are verbs so hard to remember? Effects of semantic context on memory for verbs and nouns”. In: *Cognitive science* 41 (2017), pp. 780–807 (cit. on pp. 91, 113).
- [246] Marta Kutas and Steven A Hillyard. “Reading senseless sentences: Brain potentials reflect semantic incongruity”. In: *Science* 207.4427 (1980), pp. 203–205 (cit. on p. 91).
- [247] Kara D Federmeier et al. “Brain responses to nouns, verbs and class-ambiguous words in context”. In: *Brain* 123.12 (2000), pp. 2552–2566 (cit. on pp. 92, 120).
- [248] Chia-lin Lee and Kara D Federmeier. “To mind the mind: An event-related potential study of word class and semantic ambiguity”. In: *Brain Research* 1081.1 (2006), pp. 191–202 (cit. on pp. 92, 120).
- [249] Robert Oostenveld and Peter Praamstra. “The five percent electrode system for high-resolution EEG and ERP measurements”. In: *Clinical neurophysiology* 112.4 (2001), pp. 713–719 (cit. on p. 92).
- [250] Wendy Zhang and Emily S Kappenman. “Maximizing signal-to-noise ratio and statistical power in ERP measurement: Single sites versus multi-site average clusters”. In: *Psychophysiology* 61.2 (2024), e14440 (cit. on p. 92).
- [251] Steven J Luck and Nicholas Gaspelin. “How to get statistically significant effects in any ERP experiment (and why you shouldn’t)”. In: *Psychophysiology* 54.1 (2017), pp. 146–157 (cit. on pp. 92, 96).
- [252] Audrey Bürki, Jaromil Frossard, and Olivier Renaud. “Accounting for stimulus and participant effects in event-related potential analyses to increase the replicability of studies”. In: *Journal of neuroscience methods* 309 (2018), pp. 218–227 (cit. on p. 92).
- [253] David M Groppe, Thomas P Urbach, and Marta Kutas. “Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review”. In: *Psychophysiology* 48.12 (2011), pp. 1711–1725 (cit. on pp. 93, 96).

- [254] Aida Hajizadeh et al. “Why do humans have unique auditory event-related fields? Evidence from computational modeling and MEG experiments”. In: *Psychophysiology* 58.4 (2021), e13769 (cit. on p. 93).
- [255] Kenneth A Norman et al. “Beyond mind-reading: multi-voxel pattern analysis of fMRI data”. In: *Trends in cognitive sciences* 10.9 (2006), pp. 424–430 (cit. on p. 93).
- [256] Timon Merk et al. “Machine learning based brain signal decoding for intelligent adaptive deep brain stimulation”. In: *Experimental Neurology* 351 (2022), p. 113993 (cit. on p. 93).
- [257] Carlos Daniel Carrasco et al. “Using multivariate pattern analysis to increase effect sizes for event-related potential analyses”. In: *Psychophysiology* (2024), e14570 (cit. on pp. 93, 98, 102, 122).
- [258] Timothy Trammel et al. “Decoding semantic relatedness and prediction from EEG: A classification method comparison”. In: *NeuroImage* 277 (2023), p. 120268 (cit. on pp. 93, 98, 123).
- [259] Marty G Woldorff. “Distortion of ERP averages due to overlap from temporally adjacent ERPs: Analysis and correction”. In: *Psychophysiology* 30.1 (1993), pp. 98–119 (cit. on p. 93).
- [260] Katherine A Delong et al. “Overlapping dual ERP responses to low cloze probability sentence continuations”. In: *Psychophysiology* 48.9 (2011), pp. 1203–1207 (cit. on p. 93).
- [261] Michael GH Coles and Michael D Rugg. *Electrophysiology of Mind: Event-related brain potentials and Cognition*. Oxford University Press, 1996 (cit. on p. 93).
- [262] Edvard Heikel, Jona Sassenhagen, and Christian J Fiebach. “Time-generalized multivariate analysis of EEG responses reveals a cascading architecture of semantic mismatch processing”. In: *Brain and language* 184 (2018), pp. 43–53 (cit. on p. 93).

- [263] Jona Sassenhagen and Christian J Fiebach. “Finding the P3 in the P600: Decoding shared neural mechanisms of responses to syntactic violations and oddball targets”. In: *NeuroImage* 200 (2019), pp. 425–436 (cit. on p. 93).
- [264] Alona Fyshe. “Studying language in context using the temporal generalization method”. In: *Philosophical Transactions of the Royal Society B* 375.1791 (2020), p. 20180531 (cit. on p. 93).
- [265] Karl Friston. “Ten ironic rules for non-statistical reviewers”. In: *Neuroimage* 61.4 (2012), pp. 1300–1310 (cit. on p. 95).
- [266] Daniyal Rajput, Wei-Jen Wang, and Chun-Chuan Chen. “Evaluation of a decided sample size in machine learning applications”. In: *BMC bioinformatics* 24.1 (2023), p. 48 (cit. on p. 95).
- [267] Eric Maris and Robert Oostenveld. “Nonparametric statistical testing of EEG-and MEG-data”. In: *Journal of neuroscience methods* 164.1 (2007), pp. 177–190 (cit. on pp. 96, 135).
- [268] Robert H Riffenburgh and Daniel L Gillen. *Statistics in medicine*. Academic press, 2020 (cit. on p. 97).
- [269] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. “A training algorithm for optimal margin classifiers”. In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 144–152 (cit. on p. 98).
- [270] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000 (cit. on p. 98).
- [271] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006 (cit. on pp. 100, 102, 134).
- [272] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20 (1995), pp. 273–297 (cit. on pp. 100, 101).

- [273] Leyla Isik et al. “The dynamics of invariant object recognition in the human visual system”. In: *Journal of neurophysiology* 111.1 (2014), pp. 91–102 (cit. on p. 102).
- [274] Tijl Grootswagers, Susan G Wardle, and Thomas A Carlson. “Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data”. In: *Journal of cognitive neuroscience* 29.4 (2017), pp. 677–697 (cit. on p. 102).
- [275] Gi-Yeul Bae and Steven J Luck. “Dissociable decoding of spatial attention and working memory from EEG oscillations and sustained potentials”. In: *Journal of Neuroscience* 38.2 (2018), pp. 409–422 (cit. on p. 102).
- [276] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013 (cit. on p. 105).
- [277] Daniël Lakens. “Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs”. In: *Frontiers in psychology* 4 (2013), p. 863 (cit. on p. 105).
- [278] John PA Ioannidis. “Why most published research findings are false”. In: *PLoS medicine* 2.8 (2005), e124 (cit. on p. 105).
- [279] Marta Kutas and Kara D Federmeier. “Electrophysiology reveals semantic memory use in language comprehension”. In: *Trends in cognitive sciences* 4.12 (2000), pp. 463–470 (cit. on p. 110).
- [280] Helen Bird, David Howard, and Sue Franklin. “Why is a verb like an inanimate object? Grammatical category and semantic category deficits”. In: *Brain and language* 72.3 (2000), pp. 246–309 (cit. on p. 113).
- [281] Jun Feng et al. “No morphological markers, no problem: ERP study reveals semantic contribution to distinct neural substrates between noun and verb processing in online sentence comprehension”. In: *Frontiers in Neuroscience* 13 (2019), p. 957 (cit. on p. 113).

- [282] Gabriella Vigliocco et al. “Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies”. In: *Neuroscience & Biobehavioral Reviews* 35.3 (2011), pp. 407–426 (cit. on p. 113).
- [283] Silvia P Gennari et al. “Context-dependent interpretation of words: Evidence for interactive neural processes”. In: *Neuroimage* 35.3 (2007), pp. 1278–1286 (cit. on p. 114).
- [284] Pawel Matykiewicz and John Pesian. “Effect of small sample size on text categorization with support vector machines”. In: *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*. 2012, pp. 193–201 (cit. on p. 118).
- [285] Minghui Meng and Chuanfeng Zhao. “Application of support vector machines to a small-sample prediction”. In: *Advances in Petroleum Exploration and Development* 10.2 (2015), pp. 72–75 (cit. on p. 118).
- [286] Lee Osterhout and Phillip J Holcomb. “Event-related brain potentials elicited by syntactic anomaly”. In: *Journal of memory and language* 31.6 (1992), pp. 785–806 (cit. on p. 122).
- [287] Ana C Gouvea et al. “The linguistic processes underlying the P600”. In: *Language and cognitive processes* 25.2 (2010), pp. 149–188 (cit. on p. 122).
- [288] Chunhui Zhang et al. “Working Memory Identifies Reasoning Limits in Language Models”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 16896–16922. DOI: 10.18653/v1/2024.emnlp-main.938. URL: <https://aclanthology.org/2024.emnlp-main.938/> (cit. on p. 124).
- [289] Yifei Wang et al. “Unveiling Factual Recall Behaviors of Large Language Models through Knowledge Neurons”. In: *Proceedings of the 2024 Conference*

- on Empirical Methods in Natural Language Processing*. 2024, pp. 7388–7402 (cit. on p. 124).
- [290] James A Michaelov et al. “Different kinds of cognitive plausibility: why are transformers better than RNNs at predicting N400 amplitude?” In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 43. 43. 2021 (cit. on p. 125).
- [291] James A Michaelov, Seana Coulson, and Benjamin K Bergen. “So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements”. In: *IEEE Transactions on Cognitive and Developmental Systems* 15.3 (2022), pp. 1033–1042 (cit. on pp. 125, 148).
- [292] Thimira Amaratunga. “NLP Through the Ages”. In: *Understanding Large Language Models: Learning Their Underlying Concepts and Technologies*. Springer, 2023, pp. 9–54 (cit. on p. 128).
- [293] Bhavin Desai et al. “Large Language Models: A Comprehensive Exploration of Modern AI’s Potential and Pitfalls”. In: *Journal of Innovative Technologies* 6.1 (2023) (cit. on p. 128).
- [294] Paula Vaz Lobo and David Martins de Matos. “Fairy Tale Corpus Organization Using Latent Semantic Mapping and an Item-to-item Top-n Recommendation Algorithm”. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. 2010 (cit. on p. 128).
- [295] Dan Klein and Christopher D Manning. “Fast exact inference with a factored model for natural language parsing”. In: *Advances in neural information processing systems* 15 (2002) (cit. on p. 128).
- [296] Matthew W Lowder et al. “Lexical predictability during natural reading: Effects of surprisal and entropy reduction”. In: *Cognitive science* 42 (2018), pp. 1166–1183 (cit. on p. 131).

- [297] Byung-Doh Oh, Shisen Yue, and William Schuler. “Frequency Explains the Inverse Correlation of Large Language Models’ Size, Training Data Amount, and Surprisal’s Fit to Reading Times”. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 2644–2663 (cit. on pp. 134, 149).
- [298] Arthur E Hoerl and Robert W Kennard. “Ridge regression: applications to nonorthogonal problems”. In: *Technometrics* 12.1 (1970), pp. 69–82 (cit. on p. 134).
- [299] Tom Dupré La Tour et al. “Feature-space selection with banded ridge regression”. In: *NeuroImage* 264 (2022), p. 119728 (cit. on p. 135).
- [300] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830 (cit. on p. 135).
- [301] Sander Greenland et al. “Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations”. In: *European journal of epidemiology* 31.4 (2016), pp. 337–350 (cit. on p. 135).
- [302] Stephen M Smith and Thomas E Nichols. “Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference”. In: *Neuroimage* 44.1 (2009), pp. 83–98 (cit. on p. 136).
- [303] R Quian Quiroga et al. “Performance of different synchronization measures in real data: a case study on electroencephalographic signals”. In: *Physical Review E* 65.4 (2002), p. 041903 (cit. on p. 137).
- [304] JD Bonita et al. “Time domain measures of inter-channel EEG correlations: a comparison of linear, nonparametric and nonlinear measures”. In: *Cognitive neurodynamics* 8 (2014), pp. 1–15 (cit. on p. 137).
- [305] Hugo Touvron et al. “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023) (cit. on p. 148).

- [306] Ebtesam Almazrouei et al. “The falcon series of open language models”. In: *arXiv preprint arXiv:2311.16867* (2023) (cit. on p. 148).
- [307] Marten van Schijndel and Tal Linzen. “Can Entropy Explain Successor Surprisal Effects in Reading?” In: *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*. 2019, pp. 1–7 (cit. on p. 149).
- [308] Mengzhou Xia et al. “Training Trajectories of Language Models Across Scales”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 13711–13738 (cit. on p. 149).
- [309] Peng Qian and Roger P Levy. “Neural language models as psycholinguistic subjects: representations of syntactic state”. In: *Association for Computational Linguistics*. 2019 (cit. on p. 149).
- [310] Ethan Wilcox, Pranali Vani, and Roger Levy. “A Targeted Assessment of Incremental Processing in Neural Language Models and Humans”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 939–952 (cit. on p. 149).
- [311] Suhas Arehalli, Brian W Dillon, and Tal Linzen. “Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities”. In: *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*. 2022, pp. 301–313 (cit. on p. 149).
- [312] Simon Suster, Stéphan Tulkens, and Walter Daelemans. “A Short Review of Ethical Challenges in Clinical Natural Language Processing”. In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. 2017, pp. 80–87 (cit. on p. 151).

- [313] Elisabeth Krauel and Andreas Frewer. “Dual use concerns in artificial intelligence and the neurosciences: How medical research can end up in war”. In: *Research Ethics* (2024), p. 17470161241262149 (cit. on p. 151).
- [314] Joseph Henrich, Steven J Heine, and Ara Norenzayan. “The weirdest people in the world?”. In: *Behavioral and brain sciences* 33.2-3 (2010), pp. 61–83 (cit. on p. 151).
- [315] Robert M Kaplan, David A Chambers, and Russell E Glasgow. “Big data and large sample size: a cautionary note on the potential for bias”. In: *Clinical and translational science* 7.4 (2014), pp. 342–346 (cit. on p. 151).
- [316] Shannon Spruit and Dirk Hovy. “The social impact of natural language processing”. In: *Proceedings of the Association of Computational Linguistics*. 2016, pp. 591–598 (cit. on p. 151).
- [317] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. “Semantics derived automatically from language corpora contain human-like biases”. In: *Science* 356.6334 (2017), pp. 183–186 (cit. on p. 151).
- [318] Sarah Fabi and Hartmut Leuthold. “Racial bias in empathy: Do we process dark-and fair-colored hands in pain differently? An EEG study”. In: *Neuropsychologia* 114 (2018), pp. 143–157 (cit. on p. 151).
- [319] Francisco Pereira et al. “Toward a universal decoder of linguistic meaning from brain activation”. In: *Nature communications* 9.1 (2018), p. 963 (cit. on p. 157).
- [320] Jon Gauthier and Roger Levy. “Linking artificial and human neural representations of language”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 529–539 (cit. on p. 157).

- [321] Huizhen Tang, Stephen Crain, and Blake W Johnson. “Dual temporal encoding mechanisms in human auditory cortex: Evidence from MEG and EEG”. In: *Neuroimage* 128 (2016), pp. 32–43 (cit. on p. 157).
- [322] Sophie K Herbst, Lorenz Fiedler, and Jonas Obleser. “Tracking temporal hazard in the human electroencephalogram using a forward encoding model”. In: *eneuro* 5.2 (2018) (cit. on p. 157).
- [323] Cory Shain. “A large-scale study of the effects of word frequency and predictability in naturalistic reading”. In: *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long and Short papers)*. 2019, pp. 4086–4094 (cit. on p. 158).
- [324] Jacob Louis Hoover et al. “The plausibility of sampling as an algorithmic theory of sentence processing”. In: *Open Mind* 7 (2023), pp. 350–391 (cit. on p. 158).
- [325] Ethan G Wilcox et al. “Testing the predictions of surprisal theory in 11 languages”. In: *Transactions of the Association for Computational Linguistics* 11 (2023), pp. 1451–1470 (cit. on p. 158).
- [326] Blake A Richards et al. “A deep learning framework for neuroscience”. In: *Nature neuroscience* 22.11 (2019), pp. 1761–1770 (cit. on p. 158).
- [327] Uri Hasson, Samuel A Nastase, and Ariel Goldstein. “Direct fit to nature: an evolutionary perspective on biological and artificial neural networks”. In: *Neuron* 105.3 (2020), pp. 416–434 (cit. on p. 158).
- [328] Mariya Toneva and Leila Wehbe. “Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)”. In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 158).
- [329] Ariel Goldstein et al. “Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns”. In: *Nature communications* 15.1 (2024), p. 2768 (cit. on p. 158).

- [330] Francis R Willett et al. “High-performance brain-to-text communication via handwriting”. In: *Nature* 593.7858 (2021), pp. 249–254 (cit. on p. 159).
- [331] Xiang Zhang et al. “Converting your thoughts to texts: Enabling brain typing via deep feature learning of eeg signals”. In: *2018 IEEE international conference on pervasive computing and communications (PerCom)*. IEEE. 2018, pp. 1–10 (cit. on p. 159).
- [332] Jing Yang et al. “Thoughts of brain EEG signal-to-text conversion using weighted feature fusion-based multiscale dilated adaptive DenseNet with attention mechanism”. In: *Biomedical Signal Processing and Control* 86 (2023), p. 105120 (cit. on p. 159).
- [333] Alexander B Silva et al. “A bilingual speech neuroprosthesis driven by cortical articulatory representations shared between languages”. In: *Nature Biomedical Engineering* (2024), pp. 1–15 (cit. on p. 160).
- [334] Ying Yang et al. “Commonalities and differences in the neural representations of English, Portuguese, and Mandarin sentences: When knowledge of the brain-language mappings for two languages is better than one”. In: *Brain and language* 175 (2017), pp. 77–85 (cit. on p. 160).
- [335] Morteza Dehghani et al. “Decoding the neural representation of story meanings across languages”. In: *Human brain mapping* 38.12 (2017), pp. 6096–6106 (cit. on p. 160).
- [336] Douglas G Bonett and Thomas A Wright. “Sample size requirements for estimating Pearson, Kendall and Spearman correlations”. In: *Psychometrika* 65 (2000), pp. 23–28 (cit. on p. 162).
- [337] Thomas D Gauthier. “Detecting trends using Spearman’s rank correlation coefficient”. In: *Environmental forensics* 2.4 (2001), pp. 359–362 (cit. on p. 163).
- [338] Jiaqi Ye et al. “Time series similarity evaluation based on spearman’s correlation coefficients and distance measures”. In: *Cloud Computing and Big Data: Second International Conference, CloudCom-Asia 2015, Huangshan, China,*

June 17-19, 2015, Revised Selected Papers 2. Springer. 2015, pp. 319–331
(cit. on p. 163).