DECODING LINGUISTIC INFORMATION FROM EEG SIGNALS

by

ALEX GRAEME MURPHY

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

School of Psychology

College of Life and Environmental Sciences

University of Birmingham

January 2022

Abstract

For many years, the fields of the cognitive neuroscience of language and natural language processing (NLP) have been relatively distinct and non-overlapping. Recent breakthrough research is starting to show that these two fields, in their common goal towards understanding and modelling language, have a lot to offer each other. As developments in machine learning continue to break into new ground, due largely in part to the successful development of novel classifiers that can be efficiently trained to model highly nonlinear dynamic systems, such as language, the open question is how well these models perform on human neural signals during language processing. Recent results are beginning to show that various types of human signals (eye-tracking, fMRI, MEG) can successfully model various linguistic aspects of what is being concurrently processed by the brain. EEG is a cheap and relatively accessible way to access neural signals and this thesis explores the extent to which decoding of EEG data, using state-of-the-art models common in NLP, to carry out this task. Critically, an important foundation needs to be in place that can fully explore the types of linguistic signal that is decodable with EEG. This thesis attempts to answer this question, setting the stage for joint modelling of text and neural signals to advance the field of NLP. This research is also of interest to cognitive neuroscientists as the data collected for this thesis will be openly accessible to all, with accompanying linguistic annotation, which can help to answer various questions about the spatiotemporal dynamics during the reading of naturalistic texts. In *Chapter 1*, I provide an overview of the major literature that has investigated the status of linguistic processing from neural signals, setting the research question in the correct historical context. This literature review serves as the basis for the two experimental chapters which follow and is thus subdivided into

two main sections. *Chapter 2* explores the various aspects of linguistic processing which are decodable from the novel EEG dataset collected for this thesis, with a strong emphasis on controlling for potential confounds as much as possible. Using a novel machine learning classifier, I show that with specialised training methods, generalisation to novel data relating to part-of-speech decoding is possible. In *Chapter 3*, the preprocessing steps involved in preparing the data are examined, in which I show that depending on the modelling goal, some steps are particularly useful to boost performance of linguistic decoding of EEG stimuli. Finally, in *Chapter 4*, a broad review of the results, their implications and limitations are considered.

Dedication

Dedicated to the memory of my beautiful mother, Jillyan (1958-2020)

Acknowledgements

I would first and foremost like to thank my supervisors, Uta Noppeney & Bernd Bohnet, for all their support during these years. It has been quite an experience to witness what it's like for two experts in their respective fields to come together and focus their attention on a project combining both the worlds of cognitive neuroscience and machine learning. Together, they are truly a force to be reckoned with. I was extremely privileged to work alongside Bernd for a few months during my internship in London at Google and I will never forget the kindness, support and encouragement he showed during that time. Their high standard of scientific rigour is infectious, and it's been a pleasure to work with them.

I would also like to thank the other members of the Computational Cognitive Neuroimaging Group at the University of Birmingham, for welcoming me into their group. I would like to thank:

- **Arianna Zuanazzi,** for all the stimulating discussions on linguistics, even though our time together was only short
- **Ambra Ferrari,** for making it such a welcoming lab environment and of course the occasional Italian practice
- **Giulio Degano**, for always being on hand and happy to help with technical difficulties
- **Steffen Bürgers,** for so many interesting discussions and coding help
- **Máté Aller,** for being so helpful in so many different ways, never too busy to take time to explain a concept, discuss code issues, help fix equipment or video chat for general queries

- **Sam Jones,** for demonstrating what is required to be a superb teaching assistant, being the chief GPU fitter of the lab and a source of plenty of fun facts and interesting information that won’t soon be forgotten
- **David Meijer,** for making the working environment so pleasant, offering endless useful tips and advice on a whole range of different topics
- **Patrycja Delong**, for letting me observe your EEG experiments when I first joined and introducing me to best practices, as well as the Christmas parties you hosted for the entire lab
- **Michael Joannou:** it would have been a much more daunting prospect if we had not started our PhDs at the time. It was great to share those initial years settling in with someone else in the same position, who also had an AI-focused project. As the saying goes, a problem shared is a problem halved (unless you're trying to install CUDA on a Windows machine, then it's definitely a problem *doubled!*)

This PhD was funded by the Biotechnology and Biological Sciences Research Council (BBSRC).

Table of Contents

List of Figures

Chapter 1 – Literature Review & Background Information

Chapter 2 – Decoding PoS from EEG Signals

Chapter 3 – EEG Preprocessing for Decoding of Linguistic Stimuli

List of Tables & Formulae

Chapter 1 – Literature Review & Background Information

Chapter 2 – Decoding PoS from EEG Signals

Chapter 3 – EEG Preprocessing for Decoding of Linguistic Stimuli

List of Abbreviations

EEG	Electroencephalography
ERP	Event-Related Potential
FDR	False Discovery Rate
ICA	Independent Components Analysis
LSTM	Long Short Term Memory
MEG	Magnetoencephalography
MNN	Multivariate Noise Normalisation
NLP	Natural Language Processing
PoS	Part of Speech
ROI	Region of Interest
RNN	Recurrent Neural Networks
RSVP	Rapid Serial Visual Presentation
SNR	Signal-to-noise Ratio
SOA	Stimulus Onset Asynchrony
SVM	Support Vector Machine
VWFA	Visual Word Form Area

CHAPTER 1: LITERATURE REVIEW & BACKGROUND INFORMATION

When decoding word class or part-of-speech information from EEG data, the two primary confounds present are related to word *length* and word *frequency*. Section 1.1 will present a short review of the available evidence that is known about the electrophysiological responses related to length and frequency processing in the brain. An understanding of both is important due to the fact that one needs to be acutely aware of how responses to these confounds manifest themselves in the data that could otherwise explain target results of interest. Section 1.2 will present a review on the status of word class in the brain. Section 1.2.1 will focus on the evidence presented (and contested) relating to the status of word class and grammatical category distinctions in the human brain. The experiments carried out in this thesis have a bearing on this long-standing debate in the literature, with the novel adoption of using the latest state of the art techniques in machine learning. I am not aware of any published work that aims to decode part-of-speech directly from EEG in such a multi-class setting, using such models. The proposed analyses aim to add to the literature on this topic by decoding from multiple parts-of-speech, particularly in the cases where different classes are correlated along confounding dimensions, where successful decoding implies classifier sensitivity to something deeper. The proposed use of analysing trial-averaging techniques across a wide range of semantic classes, highlighting response commonalities more closely connected to PoS-status than anything semantic, are also informative of this question. The theoretical grounding of experimental results is an important point to consider and future developments in neural decoding can benefit greatly from strong theoretical grounding in (neuro)linguistic theory, though issues relating to divisions along semantic lines is not within the scope of this thesis.

1.1 – Length and frequency responses in EEG data

In order to explore linguistic properties relating to word class in EEG data, it is important to review the observed effects that arise in neural data during reading in order to better establish the response profiles, i.e. in terms of spatial and temporal coding, of processes such as the brain's response to the length of words (in terms of number of characters) and also responses tuned to lexical frequency, as measured by corpus-level statistics. These features are termed *confounds* when the explicit goal is to analyse neural features that characterise the profile of e.g. lexical class. In other applications, the correlation between word length and frequency can be exploited as a helpful hint at the lexical class of a word being read, either in single word decoding or while reading full sentences. This idea underlies experimental approaches presented in later chapters. This section characterises some of the major work that has gone into understanding the brain's EEG responses to word length and frequency so that these effects can be understood as best as possible when interpreting such contributions to the observed EEG data of the experimental chapters.

Other important observations relating to the ERP effects of length and frequency come from Osterhout et al., (1997), who showed that, when trying to model open vs closed class differences, important properties relating to word length and word frequency were observable in terms of peak latency and scalp distribution, depending on the length of the word and how familiar it is (i.e. frequently-used). One major finding in this work is a frontal symmetric scalp distribution associated with word length (short vs long words, categorised in a binary fashion) strongly present at 200 ms post-stimulus. Dufau et al. (2015) reported effects of word length, 100-150 ms post-stimulus as well as early frequency effects starting around 120 ms, citing similar results in Amsel (2011). This paper is worthy of highlighting because it contains EEG responses to 1,000 words over 75 participants. The observed effects are therefore over a wide range of subjects and therefore likely indicate a very robust effect. However, the stimuli set were word lists and did not represent language in use, therefore missing a lot of the potential neural signals arising from sustained mental structure building, semantic integration, syntactic effects which could have affected the observed results.

Dambacher et al. (2012) explores the effects of word frequency and predictability effects in EEG responses. The interesting analysis explored in this paper is centred on varying the experimental stimulus-onset-asynchrony (SOA) between visual presentations of words in a visual reading experiment, showing that typical long SOAs (~ 700 ms) typically used in experiments resulted in larger N400 responses, which also occur much earlier, than when SOAs are used that are more in line with the normal reading speed of humans (~ 450 ms or shorter). This is relevant in designing an EEG acquisition experiment that is as naturalistic as possible and is important to keep in mind should any comparisons be made between my results and those of long-SOA experiments with respect to e.g. word frequency. The very early word frequency responses reported in Dufau et al. (2015) were from a word list experiment with a 600 ms SOA. This long SOA perhaps facilitates an earlier effect when compared with experiments that use shorter SOAs that are more in line with the natural pace of human reading speed. Considering that responses to confound-corrected data are planned in this experiment, if word frequency responses are to be compared with the literature, then it is important to keep in mind that different experimental SOAs can facilitate responses with different latencies.

Important work in other recording modalities also gives important clues as to the location and timing of word length and frequency effects. An example of this is in Schuster et al., (2016), who use fMRI and eye-tracking to look at the effects of these variables in sentence-reading. The authors found that higher word length was correlated with larger activations seen in occipital cortex, that a U-shaped response in areas of the brain strongly linked to eye-movements was present for medium-length words, showing perhaps binary divisions between short vs long are not suited to certain localised regions. Furthermore, with regard to frequency, they found that increased word frequency was correlated with a reduction in activation across many language regions of the brain, such as the left-inferior gyrus, the superior and occipitotemporal gyri, the site of the VWFA that is purported to model more sublexical components of word reading, including word frequency (ibid.; Dehaene et al., 2001; Dehaene et al., 2004).

Sereno et al. (2020) report an early effect of word frequency, which manifests itself as positive posterior activity between 80 - 120 ms and a larger effect between 200 - 300 ms over midline and anterior ROIs for low-frequency over high-frequency words. A complementary negative-going potential response, also to low-frequency words, was found in the 400 ms range in anterior ROIs, accompanied by a posterior positive activation. Faísca et al. (2019) report that as explicit word retrieval happens, frequency effects can occur from 120 ms in ERP studies, but in cases more focused on visual recognition (and not necessarily lexical access) then frequency effects can be delayed by 100 ms. This shows that task demands interact with frequency effects, but the study itself did not reveal early frequency effects prior to 250 ms, even though the implicit reading task required word meaning retrieval. These results are important to keep in mind in light of proposed data collection techniques as the planned data collection would include a similar task and therefore, we can likely expect similar results.

A recent study (King et al., 2020) demonstrated a novel linear method to disentangle the neural contributions of word length, frequency and class towards single-trial MEG data obtained during sentence-reading. This research complements some of the research goals outlined in later chapters with a different but related neuroimaging modality (MEG). Therefore, it’s useful to outline the basic findings as the results obtained in later sections could support or challenge the findings in MEG data, which would require an explanation should any large-scale differences present themselves. Using a novel method to disentangle correlated features, the authors find an early strong response to word length beginning around 100 ms, a later strong response to word frequency arising around 200 ms and a sustained response to word class of much lower magnitude spread out over a long post-stimulus period.

1.2 - Word class responses in EEG data

The primary objective is to review the evidence for open- and closed-class word effects, with a careful eye on the potential complications that arise through confounding with other linguistic features described in the previous section. Open-class words are those which belong to systematic groups that are open-ended, i.e.

which readily admit new entries, such as nouns, verbs, adjectives etc. Conversely, closed-class words are those which belong to groups that are functional in nature, which are systematically core to a language, and which do not admit new entries, such as determiners, prepositions, pronouns etc. (Akmajian et al., 2001).

1.2.1 Background on decoding word class information in the brain

There is a long and rich history relating to the linguistic information content of EEG signals. A prime example of this is the discovery of the N400 event-related potential (ERP), discovered in Kutas & Hillyard (1980). The N400 is defined as a negative-going waveform, peaking between 300-500 ms, which can be modulated by changing the linguistic properties of a stimulus to which a subject is attending, time-locked to the exact moment of presentation. Throughout the 40 years that cognitive neuroscience has known about this effect, it has been shown to be remarkably robust, with further nuances discovered after decades of research, such as the sensitivity of the N400 to thematic role assignment (Frisch & Schlesewsky, 2001; Kutas & Federmeier, 2011). The role of the N400 in the history of neurolinguistics set the stage for further research to come.

At the same time that research into the N400 began, researchers in the related field of aphasiology were starting to unpack some curious observations that had been reported with regard to the processing of open- versus closed-class linguistic stimuli in Broca’s aphasics **(**Swinney et al. 1980**)**. It appeared that damage to Broca’s area in the left prefrontal cortex disproportionately affected the processing of closed-class words, which are essential for syntactic structure building. While non-aphasic control subjects did not exhibit any difference to word class effects in reaction time studies, this distinction clearly presented itself when assessing Broca’s aphasics’ processing of open- vs closed-class words. This research laid the foundation for a strong research programme that created a prominent role for the word class distinction in psycholinguistics. When EEG methods became available to study specific questions related to language processing in the brain, researchers soon returned to the word-class distinction, with a new analysis toolkit at their disposal.

The first major study to address the idea of word class information in EEG data was Neville et al. (1992)**.** This study primarily concentrated on addressing the theoretical issue of a biological distinction in neural processing between semantics and grammatical structure. The central thesis was that cerebral processing of language was fundamentally different if the processed word conveyed semantic (open-class) information or grammatical (closed-class) information. As is common in these early studies on word class information in EEG, there is little appreciation for the fundamentally confounded nature of linguistic variables with other intimately shared linguistic dimensions. In this specific case, the dichotomy between semantics and grammar is highly confounded with the linguistic content of the stimuli, e.g. semantic processing will activate many aspects of the linguistic system such as idiosyncratic memories, colours, associated actions and mental visualisations that are absent during processing of grammatical / closed-class words. This is a point that will be revisited in greater detail later (particularly when discussing Vigliocco et al. (2011) and Kemmerer (2014)). These ideas are important to set up in advance of the later experiments of this thesis as one core theoretical underpinning is that recordable brain activity exists that can aid decoding of word class from single-trial stimuli. As more research was done in this area, an awareness of the potential for confounded responses associated with word length and word frequency arose. Closed-class words are more likely to be short and high-frequency, with respect to open-class words. Attempts were made to limit such confounds, but to varying degrees of success. For example, the attempts in Neville et al. (1992) at confound-minimisation were widely criticised in follow-up research, since only semantic items were coded for frequency and length, while grammatical items were not coded for frequency and length, due to their nature of being primarily high-frequency and short.

Specifically, the authors contrasted the processing of open-class vs closed-class stimuli between hearing and deaf subjects and found that the processing of semantic stimuli were virtually identical in both the hearing subjects and deaf subjects, but markedly different for grammatical stimuli. The main conclusion drawn was that there are variables that affect the linguistic development of deaf language users that differentially affect the cerebral systems employed when processing closed-class stimuli. The extraneous factors that affect deaf subjects appeared to have a strong impact on how they processed closed-class stimuli, an effect which wasn't present in normal hearing subjects. For this to be possible, the two subsystems need to be in part distinct. The authors categorise this effect as a left anterior negative-going peak in the time-locked EEG signal, peaking approximately 280 ms post stimulus onset. The N280, as it was introduced, was the first demonstration of a purported EEG response specifically to the open/closed class status of a word and an important stepping-stone to later experiments, which explored neural processing distinctions along the boundary of word class.

A later study by King & Kutas (1995) explored the effects of closed-class, word length and word frequency on the ERP responses to linguistic stimuli. The authors were acutely aware that effects relating to length and frequency must be formally taken into account as they were known to be very strong modulators of EEG responses. They review evidence that previous reaction time and eye-tracking studies are modulated by word length and frequency. Specifically, they review evidence claiming that each character added to a word increases the expected gaze duration by 30 ms and each per-unit increase in log-frequency results in a similar reduction in gaze-duration, with many closed-class words not explicitly fixated at all. Later chapters will discuss preprocessing methods that aim to module the effects of eye movements in EEG signal prior to decoding, so it's useful to specify some of the specific details from earlier work here.

Taking aim at the N280, the proposed ERP index proposed by Neville et al. (1992), the authors note that another left anterior ERP component arises in open-class words but at approximately 410 ms post stimulus-onset. According to the authors, this ERP is characteristically different from the classic N400 response (which has a more posterior-central scalp topography) yet such a proximity with a classic language-modulated response might have resulted in less attention given to the N410 response. Essentially, the spatial distribution of the response is similar between the N280, reported only for closed-class words, to the N410, a response seen for open-class words. The authors then hypothesise that if there is a left anterior effect, which is at least partially modulated by lexical factors, it could be that this response is the same between open- and closed-classes, but shifted in its peak latency. In this case, the N280 and N410 would be the same type of response, but other factors (word length and frequency) affect the response latency when time-locked to stimulus-presentation.

Closed-class words are often high-frequency and consist of few characters (i.e. they are short words) and given that these manifest in clear temporal differences in reaction time and gaze-duration in eye-tracking studies (approximately 30 ms latency changes, as reported earlier), the combination of short-word and high-frequency might mean earlier responses to closed-class words (N280), while open-class words take longer to process, and this was interpreted as the N410. If these two ERP components are related, then it doesn't make sense to talk about them as two separate components and thus the authors propose the term Lexical Processing Negativity (LPN). These results are important as they underline the importance for careful matching of stimuli, specifically showing that if other confounding factors are not controlled for, one cannot fully interpret contrasts that involve further correlated variables. A lot of care will be taken in validating the acquired EEG dataset collected for this thesis in terms of variables that are confounded in natural language statistics, largely motivated by this early work that often did not attempt to take steps towards controlling for such confounds.

The results from King & Kutas (1995) are consistent with their proposal that closed-class words do not exhibit a categorically different EEG response to closed-class words, but rather that this effect is also seen with open-class words but at a later time. Importantly, the modulation of the LPN latency is attributed to the confounded linguistic factors of word length and frequency. Therefore, according to their claim, one cannot use this ERP as a measure to fundamentally distinguish between open- and closed-class processing as open-class words that are short and high-frequency would exhibit typical profiles of closed-class words. Previous research had failed to adequately take into account this confounding and the natural language statistics and profiles of open- vs closed-class words was primarily driven by factors not related to status of word class, but other related properties such as word length and word frequency. One important result from the authors' analysis is that the variance of the latency is much smaller than expected, considering the eye-tracking results mentioned earlier. For example, per-unit log change in word frequency, one sees only a 5 ms shift in latency (compared to a 30 ms change in gaze-duration in eye-tracking research). A lot of questions are still left unanswered at this point and while the analyses presented in the previous studies provide the guideposts for further refined analyses, there is still a lack of theoretical-grounded insight behind the observations.

Pulvermuller et al. (1995) also investigated the status of open- vs closed-class words in the brain using an EEG-based lexical decision task. The authors found no hemispheric peak differences for open-class words but did find closed-class words were more strongly associated with the left hemisphere (similar to results in Neville et al. 1992). The following hypothesis is proposed: the brain processes open-class words by relying on distributional assemblies equally present across both hemispheres in equal measure (equal in the sense that global averaging does not reveal a dominant polarity when contrasting both hemispheres) but closed-class words rely much more on left-hemisphere processing around the perisylvian cortices. Pulvermuller et al. claim that from 160 ms post-stimulus, different signals emerge between the two classes. The report in the paper does claim that 12 out of the 17 electrodes are centred around the perisylvian cortex so a fair criticism is how they could have measured equal contributions of both hemispheres in the open-class case and perhaps this distribution of EEG montage has played a role in eliciting the observed results.

A later study, Osterhout et al. (1997) investigated this claim further by examining the extent that word-class (open vs closed) were driven by word length and word frequency effects and ultimately found high correlation with such confounding features. This result is an expected result yet showing high correlation doesn’t directly preclude the absence of effects at the word-class level. An explicit goal of the paper was to address Neville et al. (1992)’s claim of an independent closed-class marker (N280) and the authors did not find supporting evidence for this claim in their attempt to replicate that result, but do state when looking at the grammatical class of articles (determiners), something akin to the N280 effect was visible, but this only held for this specific part of speech and not for closed-class elements as a whole. Furthermore, they did claim to find N400 effects for closed-class stimuli, which complicates the picture that had been established from multiple earlier studies. Numerous features in the many results found in this paper are interesting and relevant, but a key take home message is that the impact of word class or grammatical category needs to be considered *within sentences*, given that responses in the 400-700 ms window seemed to be consistent across grammatical categories. This is taken to be evidence of differential use of sentential-level features of word-level class / category features, but one which disappears in the averaged ERP results typical of that time. This study further investigated the scalp topographies of ERP responses and did find that they were well explained as a function of grammatical categories (parts of speech), which is early evidence of discriminability of these classes within EEG data. The suggestion arises, considering these analyses, as to whether the dichotomy between open- and closed-class elements is really a function of linguistic part-of-speech and in the aggregate accumulations of open- and closed-class data sets, different distributions of these classes give rise to the different results seen in the numerous studies that have investigated neural processing of word class.

This leaves open the possibility that there could be features of grammatical processing that are shared among the grammatical categories of closed-class words, perhaps, differentially, to collections of categories that are designated as open-class. Slightly earlier work on the noun-verb distinction in aphasics (Caramazza & Hillis. 1991; Miceli et al., 1988), most notably the double-dissociation between the impaired verb processing in Broca's aphasics and impaired noun processing in Wernicke's aphasics, had raised the profile of partially-shared and partially-distinct neural processing architectures of different parts-of-speech. This prior finding and general acceptance thereof meant that the results that Osterhout et al. found in differential part-of-speech processing fit in well with an emerging consensus that it was possible to subdivide the word class distinction into the various subcomponents (parts-of-speech) that linguists were already familiar with, but for which there was very limited evidence in the brain in terms of a precise neural homologue. Osterhout and colleagues also report interesting results on the temporal dynamics of parts-of-speech, in which negative-going waveforms peaked at 280 ms for articles / determiners, followed by prepositions at 320 ms, then pronouns at 350 ms. Slightly later, peaks were found to be associated with auxiliary verbs, and nouns and verbs peaked collectively at just over 400 ms, with a general trend for left hemisphere distributions to be larger over the left hemisphere over the temporal cortex. For the grammatical category of article (determiner), this also extended into anterior regions. It was this left anterior negative peak at 280 ms that the authors likened to the original description of the N280 ERP component for all closed-class elements. The key element in this research is that temporal dynamics are indicative of word class, and highlights the importance of data acquisition methods that have a high temporal resolution, which are able to make use of such latency-based information to help in decoding linguistic features.

The documentation of these latency-based results led some researchers to closely consider the potential mechanistic underpinnings that could explain time-varying responses to different word classes, particularly by considering both early and late responses to the processing signatures of word-class effects. An influential attempt to clarify this was proposed by Friederici (2002), building on other earlier publications on the same topic, resulting in what is known as the *syntax-first* model. The model incorporates both serial and parallel processes (Heim, 2005) across three main stages. The model is primarily focused on auditory comprehension and first proposes the processing of phonetic features and early access to word-class information during the first stage, to select candidate syntactic structures which will later be fleshed out with full semantic detail. The second stage assigns thematic roles to each element in the syntactic structure and the process ends with the final stage that checks whether the initial candidate syntactic structure matches and is compatible to be incorporated into ongoing linguistic processing. If re-analysis is needed, this process is then claimed to drive the P600 ERP, which was widely believed to be an effect of structural reanalysis of incompatible candidate syntactic structures.

Similarly, early left anterior negativity (ELAN) responses were tied to the early access of word-class information and the N400 connected to the semantic and thematic-role assignment from the incoming stream of phonetic features. The predictions of this model state that early retrieval of morphosyntactic information occurs alongside other processes and is in theory available for detection early on after onset of the stimulus, as well as in later post-stimulus temporal ranges. This suggests that information relating to morphosyntactic features, as this thesis is centred on, is available for long time periods. This is particularly relevant when considering the modelling procedure that takes into account complex interactions of context (Transformer-based models) and suggests that analyses over longer windows might be more beneficial to get a more informative temporal context.

Brown et al. (1999) attempted to provide some novel insight into the disputed claim of differential neural architectures subserving the processing of open- vs closed-class words, described as the "*lexical-categorical distinction*" in their paper. The previous studies have largely followed the pattern of concluding differential effects without fully taking into account the confounding nature of word length and word frequency effects. Follow-up research correctly points out the instability of the original conclusions in light of this, by showing the previous responses can also be explained with careful attention to the confounds. Brown et al. seek to model the confounds in a more systematic way, such that any word-class distinctions found cannot be attributable to inattention to confounds. They find two such markers of interest to the discussion of differential processing across open- vs closed-class words. The first is a bilateral anterior negativity in the 230-350 ms window, where closed-class words peak earlier than open-class words. A second negative-going trend occurs in the 350-500 ms window predominantly in the left-hemisphere, also only for closed-class words. The earlier response is not attributed to the confounding factors of length or frequency and therefore seems to be an ERP that specifically indexes categorical information relating to linguistic status of words. Although this effect does not serve as a differentiating feature that one could use to decode processing of open- or closed-class words, a class-specific latency effect was found that is modulated by word class status. This paper broadly supports Neville et al.'s 1992 claim of the closed-class specific ERP (N280) and finds support for a lexical N350 effect with a slightly different spatial topography. In this regard they agree with King & Kutas in that it is a matter of seeing the same process at different temporal latencies, but the attention paid to confounds shows that after taking length and frequency information into account, a broadly similar result to King & Kutas' was indeed seen, providing a more convincing argument as to the word-class signal being detectable after confounds had been taken into account.

The important role that Brown et al. (1999) has in the ongoing discussion about differential neural processing of word class is that it takes results from earlier papers which came to opposite conclusions and accounted for the observed effects in a clear and coherent way. Concretely, they agreed with the hypothesis put forward by King & Kutas that the earlier observed N280 and N410 (N350) are the same component, just at different latencies, yet showed these effects are still present after taking into account the confounds that King & Kutas previously said were the main modulators of such latency variations. Furthermore, the finding of the later (350-500 ms) window in the left-hemisphere is offered as a candidate for closed-class word processing (peak at 420 ms), while slightly later (480 ms) peaks in the right hemisphere occur only for open-class words. This paper reverses a trend in trying to account for word class distinctions only in terms of supposed confounds. However, it is not just word length and frequency that has claimed to be behind the observed word class effects reported in the literature. This topic will be addressed in the following subsection.

The aforementioned studies focused exclusively on English and Dutch stimuli, which raises questions as to the generalisability of findings to languages in other language families (i.e. non-Germanic). Yudes et al. (2016) studied the ERPs of open and closed-class words, with a subsequent emphasis the noun-verb distinction. Their analysis was carefully controlled for varying linguistic confounds and found that closed-class words were processed significantly earlier than open-class in left-anterior areas, supporting (E)LAN-based hypotheses, with a semantic division emerging later that defined the open-class words. More recent research on time-frequency analyses of EEG signals, relating to the open vs closed-class distinction, suggests that the mechanism underlying neural processing of open vs closed-class elements, namely a strong theta-band effect (discovered in Bastiaansen et al. (2005) in a cohort of younger subjects) is not robustly detected in older populations (Mellem et al., 2012) and thus might reflect a developmental aspect of language processing across the lifespan.

1.2.2 - Arguments against word-class as an organisational principle

A few references in the preceding section hint at the fact that observed effects from experimental research on the nature of open vs closed class processing in the brain is more associated with brain responses to confounded variables, such as length and frequency. A related claim on the lack of an organisational principle on the basis of word class comes from Vigliocco et al. 2011, similarly echoed in Kemmerer (2014). These claims state that the principles upon which the brain organises linguistic knowledge and carries out language processing are divided along the correlated dimensions of how the brain processes the semantic and conceptual structure of the elements, i.e. Kemmerer (2014) points out that nouns involve recruitment of the ventral temporal lobe to access features such as shape, while verbs recruit the posterior middle temporal and frontoparietal regions.

The analysis and presented evidence is primarily related to the noun-verb dissociation widely popularised by the work of Caramazza and colleagues in the 1900s. Higher processing demands are reported to be recruited when processing verbs, as these often contain multiple participants and require effort to integrate all the associated verbal meanings within the context. This is something not present (as much) in nouns and goes some way to explaining why some aphasics struggle with verb processing but not noun processing. The suggestion is that the limitation is at the level of effort required for successful linguistic processing and integration into the ongoing mental context.

The claims that the brain does not respond differentially to morphosyntactic classes, with reported effects explained away once one takes into account word length and frequency (Münte et al., 2001) or along semantic dimensions that are tightly correlated to various parts-of-speech (Vigliocco et al., 2011), present theoretical challenges towards the idea that machine learning systems can explicitly model these classes from neural data, either for direct decoding or as part of a larger system in which neural data accompanies standard text-based NLP techniques. The reason this is important to review is that successful development of machine learning techniques to work with NLP relies on the ability to detect aspects of neural signals that are more than just correlations of word length, frequency and other correlated linguistic variables. If neural responses to purported part-of-speech categories can be explained via such correlations, then this presents a theoretical problem that would inhibit a learning system from discriminating between, e.g. determiners and short high-frequency adjectives, since the distribution of word length and frequency in these two classes would overlap to a great degree. In such a case, we must accept that the scope for successful decoding of part-of-speech is limited only to what confounds and correlations can reveal. However, from an engineering perspective, which will be adopted in later chapters, the goal is primarily driven by successful generalisation and to this end, if the recorded neural responses aid PoS decoding, but are associated with other cognitive phenomena, this does not preclude successful development of systems that can be of great use to NLP systems in the future.

1.2.3 - Arguments for word-class as an organisational principle

Boye & Harder (2012) gave rise to an entire research programme that examined a proposed fundamental dichotomy between neural processes that subserve (i) lexical and (ii) grammatical processing. Their work ties together a look of unexplained results from linguistic theory and reframes them into a set of results that are neatly explained under their proposed assumptions of differential processing mechanisms. The core idea is that lexical items carry information and are primary, while grammatical elements are secondary and rely on lexical hosts. The underlying principle involves a series of operations that interact with lexical access, conceptual structure for information-carrying (lexical / open-class) elements while the secondary (closed-class) operations do not require such interactions with the wider linguistic system and are supported by differential neural processes. Supporting evidence for this idea has been found in aphasia studies (Garraffa & Fyndanis, 2020; Boye & Bastiaanse, 2018; Ishkhanyan et al., 2017; Nielsen et al., 2019) and more recently in TMS studies (Ishkhanyan et al., 2020).

While this information supports a dichotomy along the dimension of open vs closed-class, the status of discernible sub-classes in terms of the part-of-speech classes that are well-defined in linguistic theory and key elements in NLP remains unspecified. Taken together, there are both arguments for and against organisational principles that allow for PoS-specific neural decoding and the issue remains an open scientific question.

1.3 - Natural language processing applications of neuroimaging data

A recent paper (Hollenstein et al., 2020) has reviewed the numerous ways so far in which neural signals derived from human subjects via neuroimaging and eye-tracking are being leveraged towards NLP goals. This work places a strong emphasis on the use case of such applications and summarises promising techniques and strategies that have led to successful applications, while highlighting the many issues that research faces when dealing with the complexity of the experimental choices that are possible when both collecting data and training models for specific tasks.

An early paper that deserves attention is Bingel et al. (2016), in which the authors used fMRI data from text reading in a part-of-speech induction paradigm, specifically by convolving voxel-level BOLD values with a hemodynamic response function, deriving token-level fMRI vectors for part-of-speech classification. The authors reported a 4% reduction in error rate when using fMRI data over and above models trained on text alone. This is an important point to establish that cognitive signals from humans contain information that can be leveraged towards computational NLP algorithms.

Frank et al. (2015) employed an information-theoretic approach to combining EEG and NLP, using the metric of *surprisal*, calculated over text corpora, to track the correlation between brain signals and semantic processing. They showed that covariation among NLP metrics and corresponding human processing from EEG existed, highlighting once again that there is utility in using such signals as a window into the human processing of language. This is an important point that is built upon throughout this thesis. Hale et al. (2018) demonstrated that by injecting phrase structure building processes into the model architecture, that EEG data could be predicted during auditory sentence comprehension. This paper again highlighted how human brain signals and NLP methods can work together, but it showed that we can use representative models in order to make claims about human language processing, not just advancing NLP via the addition of neural data. A final example worth highlighting is in Schwarz & Mitchell (2019), where the authors are able to predict, from large language models, multiple language-connected ERPs (i.e. N400, ELAN, P600) by mapping LSTM output vectors to textual statistics such as word length and log probabilities. These works, taken together, provide an important fundamental basis that establishes a strong connection between human language processing and language processing tasks that can be solved by machines.

Hollenstein et al. (2018) showed specifically that information extraction, named entity recognition and semantic analyses are just some of the NLP tasks that can benefit from eye-tracking and EEG data being jointly modelled along with textual inputs. This shows a first foray into some of the more common tasks, but as of yet direct modelling of PoS information from single-trial EEG data has not been demonstrated, which is a gap that this thesis aims to bridge.

A promising strategy for future work is one that can utilise a smaller neural signal dataset along with a corresponding gold-standard linguistic parse tree and full linguistic annotation to develop a model that jointly learns from both streams of input, but which can be applied on text inputs after training, thereby gaining benefits from human neural signals but not being restricted to requiring neural signals in mass-application. A recent example of this is in Ren & Xiong (2021), who showed that attention vectors can be used to extract relevant information from saved neural signals, which can then be applied to novel input data, requiring only cognitive signals during training. A further benefit of this approach is that it is claimed that, unlike other approaches that concatenate high-dimensional vectors of brain activity, the human-derived signals have been trained so that only task-relevant information is extracted, meaning that all extraneous information not relevant to the task at hand is not processed by the model as, for its purposes, this extra data is functionally equivalent to pure noise.

1.3.1 - Transformer-based Neural Network Models

Advances in machine learning in the past five years have been considerable, with the adoption of new models that are continually breaking records in many tasks over many diverse datasets across a broad spectrum of domains. Recurrent-based models were largely in favour in the NLP community until 2017 and had often been applied to decoding approaches on EEG data. A new class of models was introduced by Vaswani et al. (2017) that involved a decomposable mechanism to take contextual information into account via a parallelisable mechanism that was more significantly more efficient than the iterative training methods inherent to recurrent architectures used in RNNs. See Figure 1.1 for a visualisation of the standard Transformer architecture.

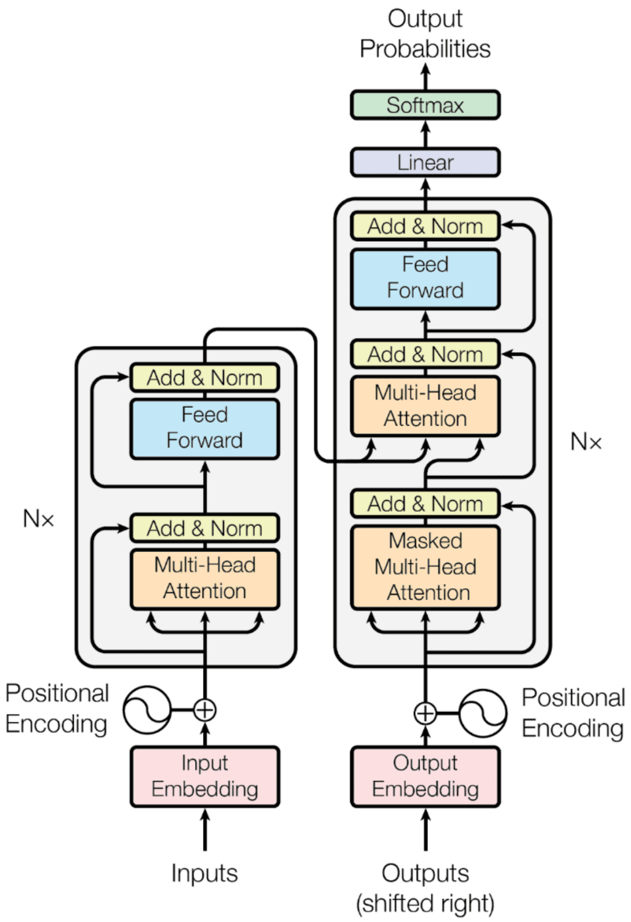

Figure 1.1. The standard Transformer model of Vaswani et al. (2017)

The original implementation of the Transformer model was composed of an encoder and a decoder, each symbolised by stacked layers called the encoding stack, decoding stack, respectively. The role of the encoder blocks is to progressively *transform* the input sequence by iteratively mapping token-level layer representations to new forms that are weighted linear combinations of the other token representations. The more encoder layers there are, the more arbitrarily and complex derived contextual representations can be. The end goal of the encoder is to create a sufficiently expressive representation where each token-element is mapped to a representation derived from all other tokens in the input sequence. In this sense the representations can be called *contextualised embeddings* of the input, namely numerical vector-space representations of tokens, where tokens can be words, sub-words or characters. Outside of natural language processing applications, there is no restriction on what can count as a token, as long as it represents a sequence of data one wishes to model (e.g. EEG vectors). On the other hand, the role of the decoder is to provide a sequential output as a function of the encoder representation. In a language translation task, the encoder could represent a sentence in one language and, using this representation, the role of the decoder would be to emit the translation of the sentence which was given to the encoder. It does this by predicting the most likely next word in the decoder's vocabulary, or more generally, the distribution of classes encoded in the final layer of the decoder. Masking is a technique that is used so that the decoder at time point t does not use information from the encoder at later time steps (since representations are derived on sentences in discrete time-steps). This allows the decoder to work step-by-step on new data, after sentences are passed to the trained encoder. The notion of *attention* in models of this type is to derive a weight-vector that, for each input token in a sentence, gives a value which delineates its contribution to the token-level representation one layer higher in the processing hierarchy. Each token in each layer (except the input) is a linear weighted contribution of the tokens from the previous layer and *attention* is the mechanism which calculates these weights.

The leftmost section of Figure 1.1 shows the encoding stack, and the rightmost section shows the decoding stack. As these models developed in the years since, derived Transformer models typically use either the encoder part exclusively (i.e.

BERT-style models, as introduced in Devlin et al., 2019) or the decoder part exclusively (i.e. autoregressive models such as GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020)). A core component of both styles of models, as well as of the original encoder-version, is the *self-attention* mechanism. This method allows for vector-based input representations to be shaped by the contexts in which they sit, ultimately conveying nuanced representations heavily shaped by each input's relation to each other input in the context.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Formula 1.1. The self-attention mechanism from the Transformer model

At the core of this method are three important matrices, which are altered during training. The input representations (in matrix form) are matrix-multiplied with each of the three matrices in order to transform the model's input into a form that is expected by the model. The Q (query) matrix treats each input as a "query" which is then matrix-multiplied by the K (key) matrix, which can be conceptualised as a "key". These together are known as the attention scores which determines how much of each vector in the context should help shape each individual representation itself. This is then scaled by the square root of the dimensionality of the vectors (which was found to give better performance) and multiplied by the V (value) matrix. Each triplet of Q, K, V matrices consists of a single "head" and by using multiple sets of triplets, concatenating the results, and projecting the expanded results down to a smaller cardinality, this leads to "multi-head" self-attention, where each triplet of matrices can be trained to detect various aspects of the input, such as syntax or morphology.

This mechanism puts the notion of the contextual environment of the inputs in centre stage and is therefore a good choice of model to capture the high-dimensional complex time-varying dynamics that are observed with neural signals typically emanating out of EEG / MEG experiments. This work is reviewed in detail here because the mechanisms of modelling are important and part of the expected advantages I hope to see during the experimental chapters. It will be important to consider the mechanisms by which Transformers perform their calculations when interpreting results in contrast to more basic machine learning classifiers. Transformers are only just beginning to be used with EEG data and this thesis aims to explore this in greater detail, with a particular emphasis on EEG-specific training steps to even further boost modelling capability and derive better models.

1.4 - Effects of EEG Preprocessing on Language Tasks

Preprocessing choices relating to specific domains of research ought to be thoroughly understood in terms of how these choices affect the downstream output of the neural representations. For example, given that many modern paradigms attempt to link machine learning and cognitive neuroscience via methods that map the outputs of artificial neural networks to vectors of brain data (encoding models), if preprocessing choices will change the vectors of brain data input to these correlational tasks, results could be significantly different had a different preprocessing scheme been applied to the data during the preprocessing stage. This is also important for decoding models as different data might contain representations associated with the target classes that allow for better metric scores, depending on the methods used to preprocess this data. This chapter covers the main steps of preprocessing commonly applied to EEG data in light of the potential downstream effects on decoding linguistic tasks, with particular emphasis on the steps which might affect the linguistic information in the signal.

1.4.1 - Ocular artefacts during reading

The main sources of noise found in EEG recordings, which is the primary target of standard preprocessing strategies in cognitive neuroscience, is considered here. This puts the following sections, relating to specific preprocessing techniques, how they work and what is achieved with them, into context. Two are standard across most instances of EEG data acquisition (electrical and muscular), but special attention must be paid to another (ocular) when dealing with acquisition of linguistic data, due to the potential informative nature of eye movements that is tied to language processing. This section highlights important information on the role of eye movements during reading in order to understand the expected effects on the recorded EEG data and how this information might be relevant for some types of linguistic decoding tasks

The electrical gradient across the eye, being positive at the cornea (front) and negative at the retina (back) gives rise to the corneoretinal dipole, which is very easily detected in EEG, particularly around the fronto-lateral electrodes (Dimigen, 2020). During left-to-right reading as is typical in Western cultures, saccadic eye movements are to the right, resulting in positive visible topographies over the right hemisphere and more negative over the left, the effects of which are linearly related to the saccadic movement (Keren et al., 2010). Corneoretinal dipole artefacts occur during blinking, too, where upward rotation of the eyeballs (Bell's phenomenon) lifts the positive end of the cornea towards and drops it back down when the eye opens, causing a transient deflection observable to the naked eye (Dimigen, 2020). It is, however, the effect of eyelid closures that has an even greater disruptive effect in the EEG recording (Iwasaki et al., 2005) though in many ways resembles effects due to blinking yet is caused by the eyelid allowing current to flow to the forehead (Dimigen, 2020). The effects of these artefacts are symmetric frontal positivity due to the raised cornea that happens during the palpebral-oculogyric reflex.

Another ocular artefact commonly observed is the myogenic spike potential which precedes saccadic eye movement, peaking at saccade onset (Keren et al., 2010). This artefact is believed to be due to the extraocular muscles and results in negative topographies around the facial electrodes, yet exhibits a less spatially-defined detectable in posterior electrodes and consequently quite difficult to remove. With this knowledge, it's beneficial to design reading-based EEG acquisition experiments with an eye to presenting stimuli in a way where saccades and their accompanying artefacts are minimised.

1.4.2 – Filtering

The presence of noise can mask a signal of interest in myriad ways. In cases where noise occupies a spectral region different to that of the signal of interest, the regions containing the noise can be attenuated (or removed completely) by filtering (de Cheveigne & Nelken, 2019). In the field of neuroimaging, such noise / nuisance signals abound, emanating both from exogenic and endogenic sources, such as power-line noise and electrical interference as instances of the former, and ocular artefacts (particularly saccades and eye-blinks) as well as muscle and cardiac signals as instances of the latter (ibid.)

Depending on the analysis, different signals from inside the body might be differentially treated as signal or noise. For example, often one treats ocular artefacts from the corneoretinal dipole during eye-blinking as noise that should be removed from EEG data, while in other experimental conditions, this could be the signal of interest while neuronal modulations of cognitive processing might be considered to be *noise*. It's important to be aware of the various filtering techniques and what problems they explicitly address, as well as any potential inadvertent implicit problems that might arise from their (mis)application to neural signals on further downstream analyses.

1.4.3 - Low-pass filters

Most aspects of cognitive processing of interest to cognitive neuroscience have been previously linked to slow neural dynamics, meaning that such signals typically have higher power at the lower frequency ranges (de Cheveigné & Nelken, 2019). On the other hand, most noise signals, ones which typically interfere with the signals of interest, belong to the higher ends of the spectral continuum, resulting in a situation where a conceptual boundary between signal and noise can be placed, where we attenuate any signal of a higher frequency and allow all lower frequencies to pass through the filter unaffected. This is typically done to remove any electrical line noise at 50 Hz or 60 Hz (depending on the recording location). Low-pass filters are also applied on the analog signal prior to digital conversion in order to avoid anti-aliasing effects. This is largely constrained by the sampling rate of the recording equipment, which in turn defines the Nyquist frequency, under which digital samples can be unambiguously reconstructed. There have been numerous claims in the literature that, through low-pass filtering, many neural signals belonging to higher frequency bands are inadvertently removed (i.e. gamma) that contain useful information for neural decoding (de Cheveigne & Nelken, 2019). A recent example of this, as it pertains to decoding of language stimuli using EEG, is given in Synigal et al. (2020). They found that including higher-frequency information from the gamma band increased the decoding accuracy of a system predicting natural speech from EEG signals. Such examples are important to keep in mind when building encoding / decoding models of linguistic brain data.

1.4.4 - High-pass filters

High-pass filters are used to remove the DC component of a signal, which can vary quite drastically over recording sessions, especially if the duration of the recording is long (Luck, 2005). Given that the data collected in this thesis are subsequently used for the decoding of linguistic features via RSVP reading, it is important to consider the preprocessing pipeline carefully, i.e. the choice of an appropriate value of the high-pass filter to apply during signal preprocessing. One such consideration is the effect of eye movements during reading, which requires correction, often implemented via ICA. Since ICA is sensitive to low-frequency drifts, the toolbox used in this paper (MNE-Python) recommends a cut-off frequency of 1 Hz if applying ICA to correct for ocular artefacts. To accept more standard values (i.e. 0.1 Hz) for the high-pass threshold, would potentially cause problems if trying to obtain a good ICA decomposition from an experiment that involved a reading task, due to this ICA sensitivity to the threshold value. A final important aspect to consider with regard to high-pass filtering is that, if many blinks contaminate the signal and are in any way condition-dependent, such as might be the case in different levels of difficulty during text reading, as will be assessed in the following data collection procedure, care must be taken if high-pass filtering is applied prior to epoch generation, since causal filter responses to blinks can extend into epochs immediately following blink activity, and acausal filter responses can be affected by blinks that occur after the epoch ends, both of which can introduce condition-dependent differences that might be picked up during later classification tasks but which are not connected with neural processing (de Cheveigne & Nelken, 2019).

1.4.5 - Band-pass / Notch filters

The electrical noise frequencies pervasive around any non-shielded recording equipment pose a problem for preprocessing and analysis of EEG data (as well as many other recording modalities). In the UK, power-line noise occurs at 50 Hz, while in other countries such as the USA, this occurs at 60 Hz. This means that preprocessing pipelines for similar experiments will attenuate different frequency bands and any signal of interest in those bands, for the same experimental task, could contain task-relevant neural information differentially-depending on where in the world the data were recorded. This also ties into the application of low-pass filters since, if the low-pass frequency is close to the powerline frequency (or in cases with more distance between them, if the powerline distance is particularly strong) then notch filtering becomes more important as simple low-pass filtering in these boundary cases will not be sufficient to attenuate the noise introduced by the electrical disturbances in the recording environment. Many analyses from the oscillatory world of neural signals focus on the properties that filtered signals contain, with reference to internal / external stimuli, in different filter bands, such as delta (0.5

- 4 Hz), theta (4 - 7.5 Hz), alpha (8-13 Hz), beta (14 - 26 Hz) and gamma (30+ Hz) (Sanei et al., 2017). In cases where the frequency band of interest is sufficiently far (in frequency space) away from the boundaries that typically cause a lot of noise, which an experimenter would like to remove from the data, it is sufficient to only band-pass the filter as this can also by default remove the effects of low-frequency drifts and electrical interference from powerline noise.

1.4.6 - ICA correction

Independent Components Analysis aims to find a weight matrix that linearly transforms observed data into a series of source vectors that exhibit the property of being maximally statistically independent (Sun et al., 2005; Luck, 2005). By decomposing neural data into such components, it is possible to plot corresponding topographies and observe the time series of each source vector and assess whether it conforms to the typical behaviour of a neural component or a noise component. Noise can arise from many different scenarios, from electrical interference to muscle activity. The sources that are not deemed to be neural in origin are then zeroed out and the inverse of the weight matrix (the mixing matrix) then rebuilds the EEG data matrix minus any defined noise source vectors. One particular collection of interferences recorded in EEG data is ocular interference in the form of blinks saccades. Horizontal movement of the ocular dipole (saccades) and upward eye movements during Bell’s phenomenon (blinking), as well as eyelid closure effects are sources that are strongly represented in EEG (Iwasaki et al., 2005), resulting in a strong positive frontal topography (Croft & Barry, 2000).

Many methods have been proposed to remove ocular effects, which are in most cases treated as noise and their presence, often orders of magnitude higher than the effect of interest (Dimigen, 2020), is undesirable. Regression-based approaches (Gratton, 1998) are commonly applied, as well as explicitly recording eye movements with electrodes during EEG acquisition via an electrooculogram. Some recent proposals involve concurrent eye-tracking during EEG acquisition, which can detect blinks and saccades and attempts to cleverly remove the effects in the data via an optimised-ICA procedure (Dimigen, 2020). A common strategy is to apply ICA to high-pass filtered EEG data and inspect topography for broad frontal positivity alongside a time series containing regular high-magnitude deflections characteristic of blinks. Some implementations of the ICA algorithm first apply Principal Components Analysis (PCA) in order to obtain a cleaner version of the data, but questions have been raised about this stage when applying correcting EEG data via ICA (Fiorenzo et al., 2018). The ocular information typically removed from the signal might, in fact, be useful when decoding linguistic information from a reading experiment (as expanded on later), therefore a careful understanding of preprocessing methods that aim to remove traces of eye movement activity is warranted.

1.4.7 - Baseline correction

Baseline correction is a standard technique in ERP research in which each channel in a window of data, containing a time-locked response to an experimentally-driven stimulus (in the case of evoked responses) other experimental response of interest, has subtracted from it a channel-wise average from a different window. It is a necessary technique to correct for the tendency for electrophysiological signals to drift over time, with a non-zero mean and an offset that is not associated with any experimental manipulation (Luck, 2005). Signal drift is still a major issue after high-pass filtering, even if the duration of recording is not comparatively long (Tanner et al., 2016).

The major assumption that underlines the most typical applications of baseline correction is that there is nothing relevant in the baseline interval that is connected to the stimulus. In many cases, this is not an issue as randomised stimuli presentations are accompanied by large stimulus onset asynchronies (SOAs) or relatively large interstimulus intervals (ISIs), wider than the effect of interest, thereby allowing recorded pre-stimulus data that is not correlated with any experimental target. Recent advances in cognitive neuroscience are many and multifaceted, with one particular development in the use and application of naturalistic stimuli. When these stimuli are of a temporal and sequential nature, such as listening to music, speech or reading continuous texts, a trade-off presents itself between the ability to record data while a subject is exposed to a more naturalistic environment, which is more ecologically valid and likely to result in more naturalistic brain responses. In such cases, where it becomes unnatural to punctuate each successive stimulus item with a sufficient gap to allow for effective baseline periods to be calculated and regularly subtracted, the immediately preceding time window will not contain responses that are not completely uncorrelated. This poses an issue when dealing with responses that are analysed collectively in which the recorded data from one epoch is contrasted with another epoch, where the former contains the time window used to baseline-correct the other. A particular example of this is component overlap, where large-scale deflections in the EEG signal systematically affect the surrounding data in systematic ways both in terms of the temporal responses, but also with respect to the topography, i.e. certain channels might be systematically affected.

This gives rise to an alternative option in which baseline correction periods can be drawn from local temporal windows that are consistent with the type of stimuli being recorded. For example, if listening to speech then any sustained pause that separates utterances is a suitable option. In reading, the interval between sentential units also provides a suitable candidate. As sentences are presented, the SOA can be increased in this interval to allow for a period of recording in which channel averaging can occur from data that will not be used during any experimental analysis. This means that a certain level of signal drift is accepted, which correlates then with sentence length. Beyond the benefit of applying baseline correction on data truly external to any experimental analysis, sentence-length dependent drift is also potentially useful for downstream linguistic decoding.

1.4.8 - Feature Scaling

Feature scaling is an important step for most classifiers, with the notable exception of tree-based methods (Bishop, 2006). A requirement is that different features are of comparable magnitudes, which allows the learning mechanisms to become sensitive to all the input features and not focus only on features that are of higher orders of magnitudes than other features. This is especially important in cognitive neuroscience, particularly with methods like EEG, which suffer from issues such as electrodes becoming loose and generating lots of noise that can’t always be filtered out with prior preprocessing steps. Noise from the environment, alongside issues with electrode connectivity, often results in noise that is orders of magnitude larger than the neural responses of interest. This means that the correct scaling of EEG data is important in general, particularly when measuring responses to linguistic stimuli, which are not as large in magnitude as responses to visual or aural stimuli.

Two implementations of feature scaling are considered in this chapter: (i) (univariate) standardisation and (ii) multivariate noise normalisation, in which the covariance of the data is taken into account in order to account for potential noisy electrodes. Due to the fact that the magnitude of the neural signal in language experiments is much smaller than noise that is typically also recorded, methods that aim to enhance signals of interest are of critical importance for optimal decoding.

1.4.8.1 – Univariate Standardisation

Univariate standardisation is a simple method by which each feature in a dataset (i.e. electrode in the case of EEG) is scaled independently with respect to all other electrodes (c.f. magnetometers in MEG; voxels in fMRI) by subtracting the mean of the channel and dividing by the standard deviation such that the resulting channel of data exhibits the properties of being zero-mean and unit variance.

$$x_{scaled} = \frac{x - mean}{sd}$$

Formula 2.1. The formula used to standardise a vector of data.

The benefit of this method is that it is computationally cheap to apply and requires minimal assumptions about the underlying data. Due to the univariate nature of this scaling procedure, no contextual information relating to the signal-to-noise ratio (SNR) of other channels is taken into consideration. Therefore, while the time series of each channel is equal in terms of variance and average value, there is no attempt to discern true neural signals of interest from those which exhibit high noise. This is exactly the rationale behind multivariate methods to provide adequate feature scaling in the context of abundant sources of potential noise interference during EEG acquisition.

1.4.8.2 - Multivariate Noise Normalisation

Multivariate Noise Normalisation is a mechanism that considers the error covariance between sensors / channels and scales the data according to the formula given below. Univariate standardisation is the case when the covariance matrix is diagonal, i.e. all off-diagonal covariances between signals is 0.

$$x^* = \Sigma^{-\frac{1}{2}} x$$

Formula 2.2. The formula for multivariate noise normalisation. Various implementations are specified by the precise form of the covariance matrix

As introduced in Guggenmos et al. (2018), it comes in three main forms: (i) baseline (ii) epoch and (iii) time point. Each version is differentiated by the data used to calculate the covariance matrix of the data, i.e. if it is calculated across the (averaged) baseline period, we get (i); when considering the average over epochs, this gives rise to (ii). The final version is more computationally costly and derives a separate covariance matrix for each time point in the epoch. Although there are doubts about its effectiveness in paradigms that make extensive use of stimuli-based pairwise differences, such as Representational Similarity Analysis (Ritchie et al., 2021), the mechanism in itself can be useful to help under-weight noisy channels and thereby allow true neural sources to be upscaled during data preprocessing.

CHAPTER 2: DECODING PART OF SPEECH FROM EEG SIGNALS

CONTRIBUTIONS

The text and all figures were produced by Alex Murphy and edited by Uta Noppeney. The experimental design was a joint effort among Alex Murphy, Uta Noppeney and Bernd Bohnet. Alex Murphy designed the EEG data acquisition process, the pilot sessions, produced all the code and collected all EEG data, including all preprocessing and SVM analyses. The training of the Transformer models was performed by Bernd Bohnet.

ACKNOWLEDGEMENTS

A paper based on this chapter has been submitted to the Association for Computational Linguistics (ACL) 2022 conference and is currently under review.

Abstract

This chapter explores techniques to predict Part-of-Speech tags from neural signals measured with millisecond temporal resolution with electroencephalography during text reading. We first show that information about word length, frequency and word class is encoded by the brain at different post-stimulus latencies and that averaging trials across these linguistic dimensions boosts classifier performance. We then demonstrate that pre-training on averaged EEG data and data augmentation techniques boosts PoS decoding accuracy of single-trial EEG data. Finally, we show that by applying optimised temporally-resolved decoding techniques, Transformer models substantially outperform linear SVMs on PoS tagging of unigram and bigram data.

Background

Recent research has shown that morphosyntactic information extracted from human functional magnetic resonance imaging (fMRI) signals during sentence-reading tasks can substantially improve the induction of part-of-speech tagging (Bingel et al., 2016). Due to the sluggish nature of the hemodynamic response function, which typically peaks between 4-6s after stimulus onset, this extracted information reflects only the associated blood oxygenation level in a relatively slow manner and is not a measure of contemporaneous neural activity. This renders fMRI as a non-ideal candidate method to model and characterise the rapid neural dynamics that underlie natural language processing in the brain. EEG, on the other hand, measures neural activity at the millisecond resolution level, which allows for the online characterisation of neural and cognitive activity as it unfolds during sentence reading.

Early event-related potential studies demonstrated that the magnitude and topography of EEG responses during text reading tasks are dependent on various aspects of the linguistic stimulus, most notably on word length, word frequency as well as word class, i.e. whether a word belongs to an open or closed class (also termed lexical / grammatical class). Word length effects arise around 150 ms, frequency effects slightly later around 200 ms and word class effects variably from around 400-700 ms (Osterhout et al., 1997; Segalowitz & Lane, 2000; Münte et al., 2001; Dufau et al., 2015). These early studies relied on averaging a vast number of trials into event-related potentials.

In this chapter we combine EEG with linear SVMs and Transformer models to investigate whether morphosyntactic information, such as information relating to part-of-speech, can be extracted not only from trial-averaged data, but also from single-trial data with the goal in mind to later develop a system that can process and categorise linguistic information from data collected in a live setting relating to specific novel input, which therefore would only exist in single-trial form in naturalistic applications. Combining EEG recordings of text corpora with PoS tagging and / or dependency tree annotations would also allow for more reliable morphosyntactic modelling than is currently available in methods that are strictly text-based. Such a development would be particularly useful for the creation of resources for under-resourced languages with limited resources.

2.1 - Decoding word length, frequency and class

2.1.1 - Introduction

Using EEG decoding with linear SVM models, we temporally resolved the linguistic variables of word length, word frequency and word class, during single-word reading of continuous naturalistic texts. To achieve this, we recorded an EEG dataset of word-level trials, from which the data used in this experiment were extracted. This allows a direct comparison with previous research assessing the effects of length, frequency and class in event-related potential research on linguistic processing during sentence and word list reading (ter Keurs et al., 1999; Hauk & Pulvermüller, 2004; Hauk et al., 2006; Osterhout et al., 1997; Osterhout and Holcomb, 1992; Sereno et al., 2020; Faísca et al., 2019; Münte et al.,2001; Dufau et al., 2015). The results of this analysis show that the EEG dataset and decoding methods are sufficient to uncover the typical temporal process underlying the cognitive activity during word reading, as it relates to the processing of word length, frequency and class. Furthermore, the use of pseudotrials (trial-averaged data) at various levels increases the decoding accuracy as the number of averaged trials increases. This confirms the EEG data contains the signal of interest that later sections depend upon and will take advantage of, as well the fact that the signal can be boosted by trial-averaging.

2.1.2 - Methods

2.1.2.1 - Data Selection

Our experimental stimuli derive from a subset of the English Web Treebank corpus (Bies et al., 2012), a collection of English texts across multiple stylistic genres: *weblogs*, *newsgroups*, *reviews* and *Yahoo answers*. These data are already annotated for various linguistic structures and information, such as part-of-speech tags and dependency parse trees. One notable feature about this data set is that it has been released as part of the 2017 CoNLL (Computational Natural Language Learning) challenge (Zeman et al., 2017), meaning that many teams have submitted NLP models on various derived tasks from this data set, providing a range of high-scoring benchmarks with which we could compare our models that also incorporate neural data from the EEG recordings. Pre-annotated data are easy to incorporate into the data structures that contain EEG data of word / sentence reading from our experiment, meaning that it is simple to attach desired metadata to the EEG data structures, which facilitates standard comparisons among any desired linguistic features (i.e. difference waves, topographies) alongside easy ways to manipulate data to be exported efficiently into formats that fit more of a standard machine learning paradigm (e.g. NumPy arrays).

The subsets from the EWT corpus that we chose to use in our experiment were `weblogs`, `newsgroups` and `reviews` and `Yahoo-answers`. The primary reason for excluding `email` was that many text files contained email formatting, URLs, email subject lines and others were of an overly casual nature with many non-standard spellings and use of online linguistic features such as emojis, which would not be used in any analysis we were considering. On the other hand, longer, well-prepared and more formal texts such as news reports (`newsgroups`) provided a range of sentence lengths that incorporate a greater diversity of vocabulary, dependency structure and sentence complexity.

2.1.2.2 - Stimuli set

The stimulus set (outlined in the previous section) contains 4,479 sentences (74,953 tokens) across four text genres. To facilitate future experiments where we can average over multiple EEG recordings of the same syntactic environments, this corpus was acquired approx. 5 and a half times before data acquisition ceased due to external factors. In total we collected EEG data for 24,323 sentences over 404,205 tokens. The mean sentence length was 16.7 words (standard deviation: 12.23 words). Table 2.1 below outlines the distribution of the Universal Dependencies PoS tagset across each partition of data.

Tag	train	dev	test	total
ADJ	24,029	3,489	2,913	30,431
ADP	33,969	5,049	4,235	43,253
ADV	17,492	2,593	2,218	22,303
AUX	19,351	2,833	2,485	24,669
CCONJ	11,758	1,731	1,546	15,035
DET	31,429	4,589	3,962	39,980
INTJ	656	76	90	822
NOUN	59,991	8,691	7,501	76,183
NUM	5,062	712	677	6,451
PART	6,955	970	908	8,833
PRON	27,623	3,973	3,677	35,273
PROPN	27,867	3,737	3,641	35,245
PUNCT	3,716	485	501	4,702
SCONJ	7,116	1,046	943	9,105
VERB	39,710	5,723	5,186	50,619
X	1,029	125	147	1,301
Total	317,753	45,822	40,630	404,205

Table 2.1. Number of samples for each PoS tag across the train, development and test sets along with the total values across the entire dataset

2.1.2.3 - Participant Selection

We decided that we would like to collect multiple repetitions of our corpus from a single subject in order that any trial-averaging is done within-subject in order to quantify the applicability of this method in terms of SNR increases due to repetitions of the stimuli. If we had multiple subjects and trial-averaging did not work, we would struggle to disassociate myriad factors such as exact cap placement, different subject-specific phenomena. We can be more certain of the effects of trial-averaging by acquiring multiple versions of a dataset from the same subject.

The single subject was selected out of a small group who had been previously invited to help trial some potential settings of our data acquisition procedure. We invited 5 subjects to participate in an experimental reading trial in order to determine the most effective word presentation rate. We tested three different levels of reading speed: slow (500 ms SOA), medium (240 ms SOA) and fast (120 ms SOA). Afterwards, we performed a 10-question post hoc memory test as well as a small discussion relating to the participants' subjective experience with regard to the ease of reading at those different word presentation rates, which determined that the medium reading speed was most comfortable and resulted in the highest recall scores on post-reading tests. Two subjects were selected to take part in a trial using EEG recordings, in order to test the experimental setup. Only one of these two subjects was later able to commit to the long-term time frame required in order to acquire the amount of data planned as part of the experiment. We performed a sanity check on the trial EEG data and determined the correctness of the experimental setup. The subject agreed to take part in the full experiment and gave informed consent according to the ethics procedure at the University of Birmingham. All participants invited into the lab received monetary compensation for their time.

2.1.3 - Experimental Procedure

In order to reduce extraneous signal contamination from saccadic eye movements, we presented sentences one word at a time in the centre of a screen using Rapid Serial Visual Presentation (RSVP) with an approximate Stimulus Onset Asynchrony (SOA) of 240 ms. Stimuli were presented in a white monospace font (Courier) on a light grey background of an LCD monitor (1920 x 1080) using the Python package PsychoPy (Peirce et al., 2019). Relative to a white fixation cross at the centre of the screen, we presented each individual word in accordance with its optimal viewing position (Rayner et al., 2016), which has been shown to reduce microsaccades during active reading in RSVP, allowing for increased ease of reading. Relative to the central fixation-cross, each word subtended a horizontal angle of 0.76 degrees to the left and 11.81 degrees to the right. Between sentence boundaries the SOA was increased to 500 ms while a white fixation cross was presented in the centre of the screen. Figure 2.1 shows a visual depiction of the experimental setup.

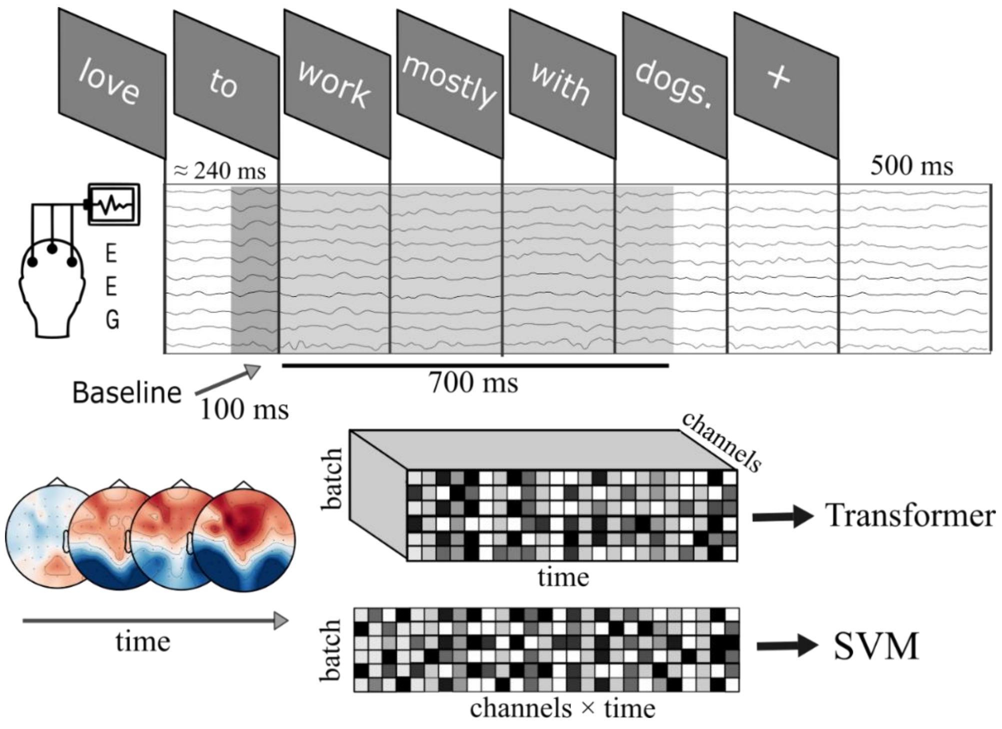

Figure 2.1. An example trial and associated EEG recording. Sentence words were presented on average approximately every 240 ms. EEG signals were extracted from -100 to 700 ms relative to the onset of when the word appeared on screen.

In order to ensure that the subject was actively processing the text on the screen, on approximately 20% of the sentences in each recording session, the subject was given an on-screen prompt to verbalise back to the experimenter as much of the previous sentence as can be remembered. A mean accuracy of 93% across all sessions demonstrated that the subject was actively processing the text as it was being presented on screen. To promote the comfort of the subject during the reading experiment, as well as reducing any potential movement artefacts in the EEG signal, we stabilised the subject in a chinrest that was aligned with the centre of the presentation screen.

2.1.3.1 - EEG data acquisition

Continuous EEG signals were recorded via BrainVision's PyCorder software using reference-free mode at a sampling rate of 1,000 Hz. A 64-electrode cap in 10-20 layout with Ag/AgCl active actiCAP slim electrodes (ActiCAP, Brain Products, GmbH, Gilching, Germany). Prior to each recording session, channel impedances were verified to be below 15 kΩ.

2.1.3.2 - EEG preprocessing

The number of individual EEG recording sessions was 77. The collected EEG data were preprocessed using MNE-Python (Gramfort et al., 2014). The data were first band-pass filtered between 1-40 Hz and then downsampled to a new sampling rate of 250 Hz and re-referenced to the common average reference. Noisy channels were identified by a session-wise power spectral density plot and were then interpolated. 3 recording sessions were discarded due to excessive noise / interference that could not be corrected, resulting in 74 pre-processed recording sessions. Each recording session’s EEG data were then decomposed via Independent Component Analysis (ICA), where an average of 4 non-neuronal components relating to ocular, muscular and electrical artefacts were removed.

Relative to stimulus onset-aligned word-level triggers, EEG data were extracted from -100 ms to 700 ms, resulting in 200 time points, each time point representing 4 ms of data. The pre-stimulus period (100 ms) was then used for baseline correction, such that the channel-wise mean from this period was subtracted from the rest of the epoch. The remaining 700 ms was extracted to ensure that we could capture late syntactic effects, such as the P600. The prestimulus period was then discarded, resulting in word-level EEG trials that represent the 700 ms post-stimulus window (176 time points in total). For all the recording sessions that were not discarded (74 out of 77), we did not perform any epoch-level quality checks that would have resulted in the dropping of noisy trials. The motivation for this was that we wanted to be able to examine sentence-level effects and to get an estimate of expected results when using all trials in a recording session. In live applications, which we envisage for the future, one is not afforded the luxury of being able to drop excessively noisy trials if those trials make up the constituent parts of phrasal structures such as sentences or subject-predicate sentential units. By removing noisy trials in this way, any such accuracy estimates would be an overoptimistic expectation of what might be possible in live applications.

The EEG data were then spatially multivariate noise normalised using a noise covariance matrix, estimated separately for each target class (Guggenmos et al., 2018). The associated metadata for each word-level trial were then annotated with the gold part-of-speech tags of the current and following words, along with their word lengths and the Zipf-logarithmic frequency scores from the *WordFreq* Python package (Speer et al., 2018).

2.1.4 - Data splits

The entirety of the EEG corpus consists of 74 recording sessions (each approximately 20-25 minutes in duration). These sessions were recorded over a period of 20 individual days, in which multiple recording sessions were obtained on each day. In cases where multiple recording sessions occur on the same day, the exact EEG cap placement and thus electrode location might unfairly aid training and generalisation scores so in order to create data splits that were equally balanced in this regard, a specific procedure was implemented to try to carefully match the development and test sets so that expected results on the development set (i.e. to assess early stopping) would also apply to expected performance on the test set.

In order to create a development and test set that were fairly matched yet completely independent of the training data, segmentation of the entire dataset happened at the text-file level. 10% of the data was assigned to the development set (and 10% to the test set) in the following way. First, a recording date was selected, and all the filenames read by the subject on that day were extracted. A random text file from this list was sampled to be part of the development set and its size (measured in terms of number of individual words) was recorded. This text file was then removed from the sample and the text file with the most similar size (in terms of number of individual words) was extracted and became part of the test set. This process was repeated until approximately 9% of the dataset had been each assigned to the development and test sets (whereby the training dataset consisted of 82% of the original data). This process is depicted in Figure 2.2.

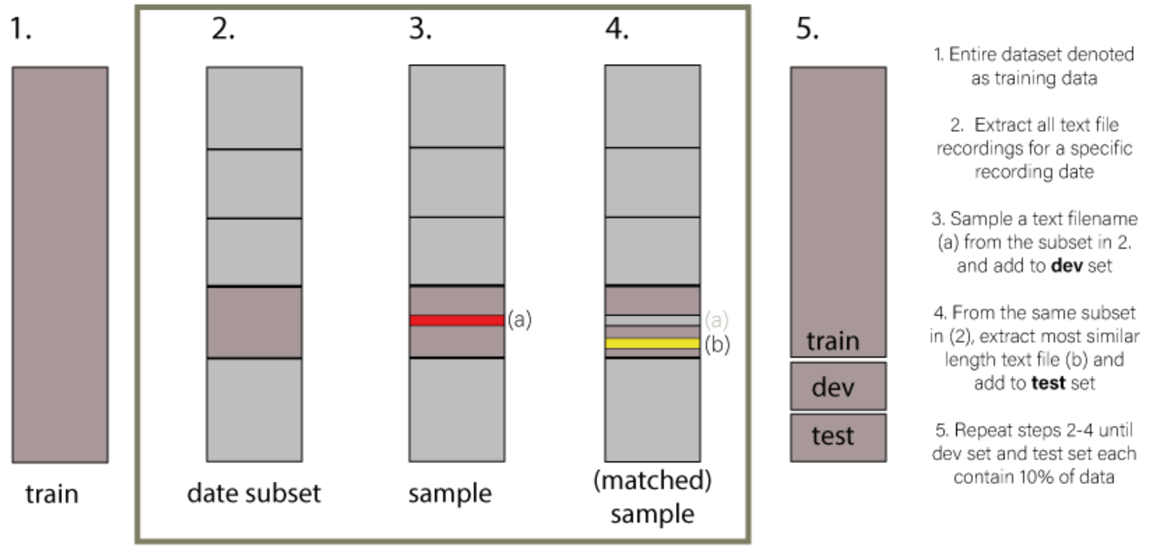
1.
train
2.
date subset
3.
(a)
sample
4.
(a)
(b)
(matched) sample
5.
train
dev
test
1. Entire dataset denoted as training data
2. Extract all text file recordings for a specific recording date
3. Sample a text filename (a) from the subset in 2. and add to **dev** set
4. From the same subset in (2), extract most similar length text file (b) and add to **test** set
5. Repeat steps 2-4 until dev set and test set each contain 10% of data

Figure 2.2 The data splitting procedure. Initially the corpus of text files is designated entirely as training data (1). Dates were EEG data were acquired were then looped through until the end. At each intermediate step, all text files acquired on a specific day were extracted (2) and from that subset a single file was sampled (3). This file is added to the dev set and the closest-matched (in terms of text length) from the same subset is added to the test set (4). This ensures a balance between both data splits. This process is repeated until each of the dev and test set consist of 10% of the text files and associated EEG data.

In Figure 2.3 below, three histograms of the number of word-level EEG trial occurrences are given. On the left (blue) is the training set distribution. The rightmost two (red: development set; green: test set) have the same y-axis and show that across recording dates (x-axis) the distributions of EEG trials is roughly equivalent and shows no major confound that could be explained by the electrode locations due to recordings on the same day. Using the text files that are matched for length and recording dates, the original EEG dataset is then partitioned into training, development and test sets.

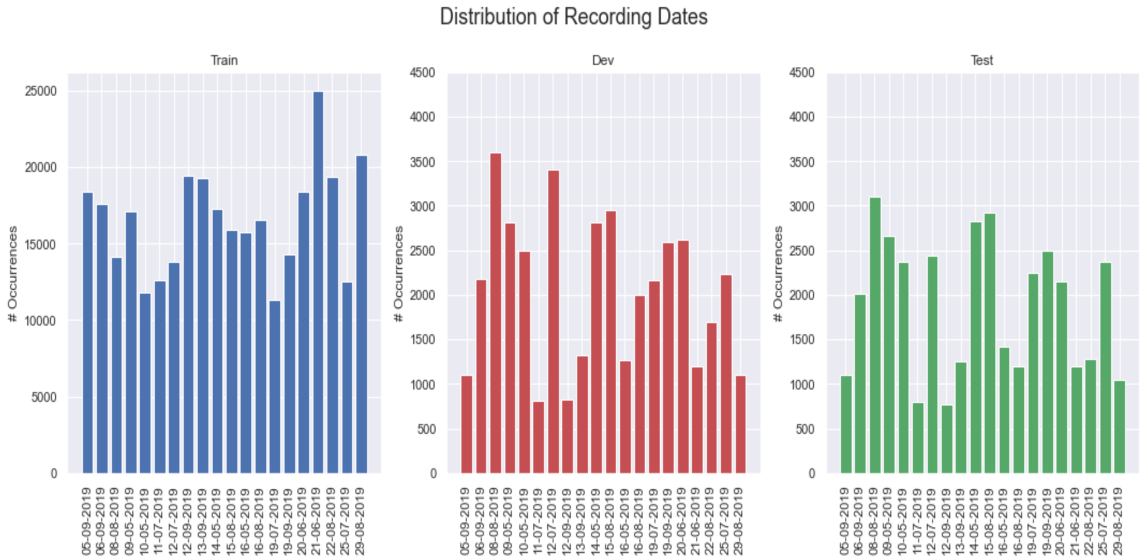

Figure 2.3. Histogram of individual recording dates for the training data set (blue, leftmost), development data set (red, middle) and test data set (green, rightmost). There is a broad balance of recording days across the data splits in order to avoid any systemic imbalance that could affect model training or generalisation performance. The development and test sets were explicitly chosen so as to reflect a similar profile across recording dates, while keeping all data between splits completely independent.

Furthermore, we assessed the distribution of text genre among the different data splits in order to verify there were no significant differences that could affect model training and generalisation. A bar chart showing the number of word-level EEG trials per dataset split is given below in Figure 2.4.

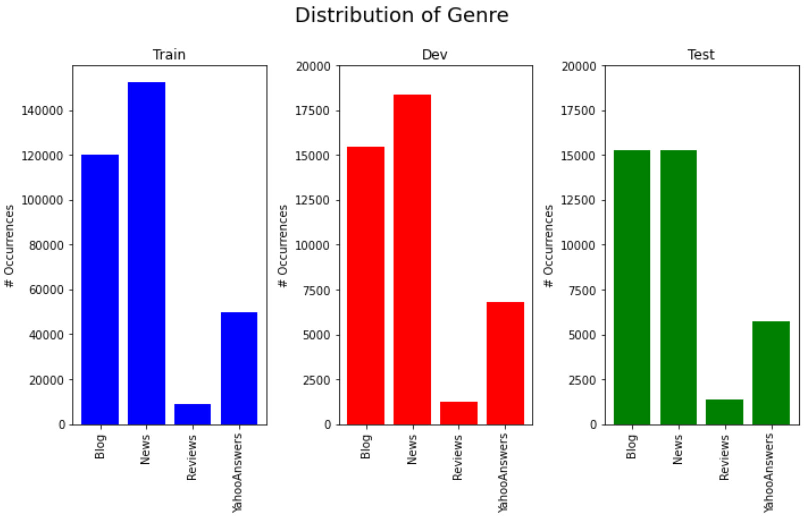

Figure 2.4. Histogram of text genre representation in each of train (left), development (middle) and test (right) data splits

2.1.5 - Decoding

We explored two classifiers in our experiments, (i) linear SVMs and (ii) encoder-based Transformer models. The reason why we chose to contrast these specific algorithms is due in part to the linear vs non-linear nature of each, in which we are able to examine the extent that information can be linearly separated in the broadband space of EEG signals. This sets an important baseline as using only a non-linear classifier, it's unclear how far above more simple approaches that these results will be with a simpler (often linear) algorithm. Namely, if results are given only for the latest state-of-the-art neural networks but a simple linear classifier achieves roughly the same performance, it's important to have such a comparison in order to quantify the difference, as otherwise the implicit biases conveyed by complex model architectures can often be (mis)credited with successful high-level decoding, when in fact comparable results on simpler linear classifiers can show this cannot be the case.

Prior to the neural network revolution, linear SVMs were among the top-performing classifiers used on neuroimaging data in the multivariate pattern analysis paradigm. This also played a role in the selection of the linear SVM as our baseline comparison with which we would compare the highly non-linear mechanisms of the Transformer. There is an on-going debate in which linear vs non-linear classifiers are arguably purported to be better models of neuroimaging data (see section in lit. review on this) and we also felt our results might add useful information to this issue.

2.1.5.1 - Classifier implementation

In later experiments, we intend to contrast linear SVMs (Chang & Lin, 2011) with Transformer models, specifically the style of models that arose out of taking the encoding part of the original encoder-decoder Transformer of Vaswani et al. (2017), similar to the BERT architecture (Devlin et al. 2019) which has gained widespread adoption due to its impressive results on many state-of-the-art datasets. A problem arises in that the training procedure for both classifiers is quite different, whereby the Transformer is trained by using Stochastic Gradient Descent (SGD) iteratively, with convergence assessed on a held-out development set and early-stopping methods which stop training when there is no consistent improvement on the development set. The standard method to train linear SVMs is to solve the constrained optimisation problem via Lagrangian multipliers (Burges, 1998). To train an SVM in the standard paradigm would require that the training data sizes be different between the linear SVM and the Transformer, in that the training data would be smaller for the SVM, while extra data is used for early-stopping during neural network training. We could combine the training and development datasets, but this then means the SVM is trained on more data than the Transformer would see.

One solution to this is to implement an online-learning implementation of a linear SVM via a stochastic gradient descent classifier, which uses the *hinge* loss function (Formula 2.3) in an iterative manner, mirroring the way neural networks are trained.

$$\mathcal{L}(Y, \dot{Y}) = \frac{1}{N} \sum_{n=1}^{N} \max(0, 1 - \dot{y}_n \cdot y_n)$$

Formula 2.3 The hinge loss function, which allows SVMs to be trained via gradient descent. The two inputs to the loss function are (a) the predicted and (b) true class (in +1/-1 coding so that a correct prediction equals 1, which results in no loss for that prediction-label pair.

The formulation of a linear SVM in this way allows us to implement the same training regime on exactly the same data, assessing the development data periodically in order to implement an early-stopping mechanism. We used a Scikit-learn (Pedregosa et al., 2011; Zhang, 2004) implementation, which is based on LIBSVM (Chang & Lin, 2011). This implementation procedure requires the specification of a hyperparameter *alpha*, which is inversely proportional to the *C* parameter in the more common implementation of SVMs. The loss function for the standard SVM is given in Formula 2.4. Higher values of C result in upweighting the sum of the slack-variables (in non-perfect classification scenarios) thereby allowing for looser margins that can increase model complexity. This means higher values of C act in the opposite direction to regularisation coefficients (Bishop, 2006). A value of 0.75 was chosen for *alpha*, based on earlier (independent) experimental analyses using pilot data.

$$C\sum_{n=1}^{N}\xi_n + \frac{1}{2}||\mathbf{w}||^2$$

Formula 2.4. SVM loss function that allows for misclassification of some data points in order to support the maximum-margin classification in non-linearly separable data spaces.

2.1.5.2 - Training procedure

Each EEG trial, containing 176 time points across 64 electrodes, was flattened into a single array, meaning each batch of data consists of a 2D matrix (batch_size x number of features). The batch size used to train all SVM models in this paper is 256, which was preselected and not a tuned hyperparameter. A full epoch is defined as a complete pass through the entire training data. The models were all trained for 8 epochs, or until there was no improvement for over 10 batches of data. During model training, the best model according to the development data was stored. The model that attained the highest accuracy on the development set was then applied to the test data in order to yield a vector of predictions. These predictions are then compared to the true labels and an accuracy value is calculated. The accuracy values are then added to an array that iterates over the 160 windows, which allows for the examination of the temporal dynamics of varying above-chance classification accuracy, namely the ability to visualise the time-varying effects of the neural data. This is performed over all 10 random seed values.

2.1.6 – Class definitions

Linguistic data are highly correlated among a large number of dimensions, meaning that in order to examine one aspect of the data, one must deal with the confounding influence of the other (correlated) dimensions. We attempted to validate the EEG data by assessing the extent to which we can decode the linguistic dimensions of (i) word length, (ii) word frequency and (iii) word class. We performed this using a series of binary experiments across each dimension (word length, word frequency and word class). We split the word length data into two classes: (a) short and (b) long. We furthermore split the word frequency data into: (a) high-frequency and (b) low-frequency. For word class data, we split the data into (a) open-class and (b) closed-class groups. Short words were defined as words that were written with up to four individual letters; anything above this was classed as a long word. We examined the Zipf-frequency scores of all words and found the median value (5.91) in the dataset. This value was used to categorise all words as either low-frequency (any word with a frequency score < 5.91) otherwise the word was categorised as a high-frequency word. Out of the part-of-speech tags used in the Universal Dependencies paradigm, we grouped `NOUN, VERB, ADJ, PROPN` & `ADV` into the open-class group and `ADP, DET, PRON, AUX, SCONJ` & `CCONJ` into the closed-class group. A representation of these binary splits is given below in Table 2.2.

Linguistic variable	**Word length**	**Word frequency**	**Word class**
Group 1	4 characters or fewer (SHORT)	Zipf frequency under 5.91 (LOW-FREQ)	ADJ, NOUN, VERB, ADV, PROPN (OPEN)
Group 2	More than 4 characters (LONG)	Zipf frequency equal or over 5.91 (HIGH-FREQ)	ADP, CCONJ, SCONJ, DET (CLOSED)

Table 2.2. Description of how each linguistic variable was binarised.

2.1.6 - Confound correction

It should be noted here that correcting for confounds in linguistic data is a near-impossible task and that the employed terminology used here (‘correction’) should be interpreted only in the sense that a concerted effort was taken to remove the most serious confounds. For example, it’s very hard to remove the correlation between closed-class words that are short and high frequency, since the class of adpositions consists mostly of short, high-frequency words, such as prepositions and articles in the case of English. We do not claim that we can completely remove such confounds, but we can define a procedure that corrects for them such that between the two binary classes (in all the cases of word length, word frequency and word class) that the distribution of the confounding variables is matched. We also take care to match across various other features such as sentence position (sentence-initial, mid-sentence and sentence-final), as change of contexts and sentence wrap-up effects are known to also have a confounding effect (e.g. that an open-class word is typically the final word in a sentence).

The principal aim of this current experiment was to determine whether we could recover the known linguistic time courses of the expected neural behaviour via classification decoding. The expected neural time courses for word length, frequency and class processing is outlined in greater detail in Chapter 1. By analysing the linguistic properties of the EEG signals in this way, we are able to (a) validate the dataset to show that the experimental and preprocessing paradigm resulted in sensible results that are in accordance with the wider literature on the topic and (b) to determine the range of accuracies we can expect from different levels of pseudotrial averaging. The reason why this latter point is important is when we later explore ways in which we can pre-train deep neural network models to learn from data that has a higher level of SNR. The necessary prior step that is required before this can be achieved is to show that averaging of this data does indeed result in a higher signal quality that is reflected by an improvement in a machine learning classifier to decode the relevant linguistic class (in a binary class setup) via a specific assessment metric (i.e. accuracy).

2.1.6.1 - Implementation of the confound-correction procedure

The following procedure was implemented independently for each of the training, development and test sets. We first discretised the frequency score for each word in the EEG dataset, by rounding to the nearest 0.25 and selecting all frequency levels between [1.0, 8.0] i.e. 28 different frequency score bins. We purposefully discarded any rare words (frequency score < 1.0) due to the fact that they are unlikely to appear in all of the train, development and test sets. With the two levels of word class, we selected all words of increasing length (based on the number of characters in the word) from 1 to 9. Finally, we left the two levels of the word class group as previously defined.

For each of: (i) word length, (ii) word frequency and (iii) word class, we constructed a joint histogram over the other two variables, as well as across sentence position (sentence-start, mid-sentence and sentence-end) and extracted equal number of samples across the main binary partition, until for word length (short / long), word frequency (high / low) and word class (open / closed), both levels of the binary partition were equally balanced with respect to all the confounds. Figure 2.5 below shows a visualisation for both levels of word class (open vs closed class), separately for each of the train, development and test sets, where the marginal distribution of each axis represents a confound w.r.t. word class and the central plot contains their joint distribution. As can be verified, for both levels of the word class variable, the distribution of word length and frequency is equally matched. The corresponding figures for word length and word frequency, across train, development, test sets are given in Appendices A and B.

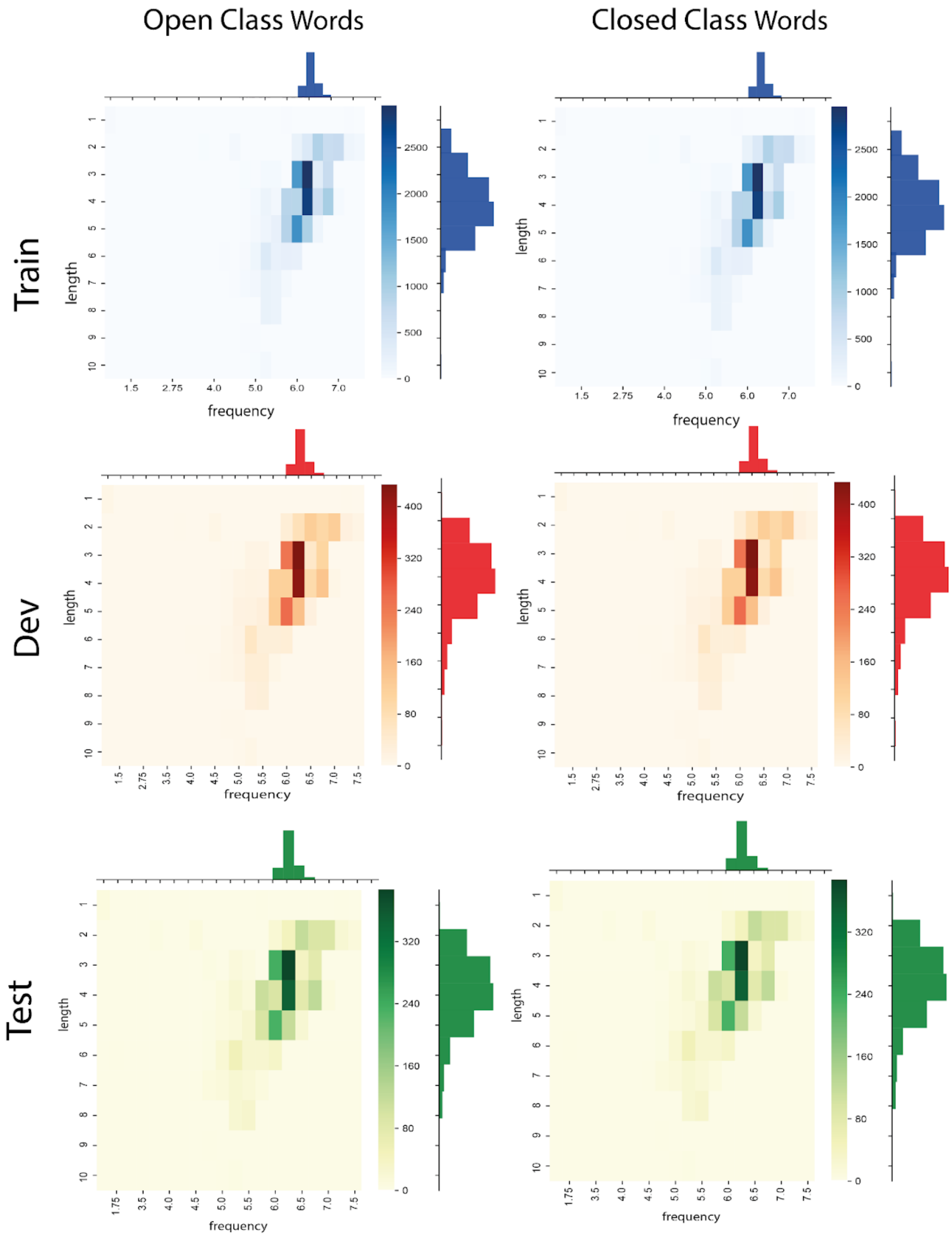

Figure 2.5. Joint histogram of confounding variables of frequency and length with respect to word class as the central variable, i.e. word class is equally balanced across open and closed-class words, both equal with respect to the distributions of the confounds. This process is done internally to each data split, which is given by each row. This process is also applied separately for each of three sentential positions (sentence-start, sentence-middle and sentence-end). Marginal distributions are given along the axes.

2.1.6.2 - Pseudotrial averaging

Trial averaging increases the signal-to-noise ratio of neural signals (Grootswagers et al., 2017; Guggenmos et al., 2018; Roy et al., 2019; Tuckute et al., 2019), while ignoring true variability of EEG data from different words in the same category and class differences in within-class variability (Münte et al., 2001). In order to explore the effect of classification accuracy with respect to the signal-to-noise ratio of the EEG trials, we examined three different levels of trial-averaging: (i) no trial averaging (using just single trials), (ii) averages of three and (iii) averages of 10. This baseline is important to determine because in later sections, we plan on using pre-training methods on higher quality data to tune the weights of neural networks prior to fine-tuning on single trials. We therefore explored how trial-averaging affects the windowed decoding traces specified above (general strategy section).

To create a pseudotrial, we always selected single-trial data from within each data partition so as to avoid any leakage of data across splits that are differentially used to train, evaluate and finally test candidate models. The procedure involved resampling with replacement (i.e. bootstrapping) either averages of 3 or 10, representing a form of low-averaging versus high-averaging. When generating pseudotrial datasets, we can theoretically pick any value to represent the number of pseudotrials to be generated, yet in order to be comparable with the single-trial dataset, we generated pseudotrials until these pseudotrial datasets were the same size as the corresponding single-trial dataset. In order not to violate the independence assumptions of the statistical tests we planned to use to assess the use of trial-averaging on classification accuracy (repeated Binomial tests corrected for multiple comparisons), resampling in the test set was always performed without replacement, meaning that the test sets for each level of averaging is necessarily different, whereby the highest-level of averaging results in the smallest test set. The number of samples in each data set split, for each linguistic variable of interest (word length, frequency and class) are given in Table 2.3.

	length	**frequency**	**class**
train	82,424	51,364	45,502
development	12,402	7,632	6,658
test (single-trial)	10,810	6,590	5,670
test (avg. 3)	3,603	2,196	1,890
test (avg. 10)	1,081	659	567

Table 2.3. Distribution of word length, frequency and class over data partitions

If random samples of single trials, when averaged together, result in higher metrics being measured, then we can assume that the features of the data that are averaged together are driving this effect. Namely, if we are averaging together low-frequency words, then these averages will be across a range of different word lengths and word classes, thereby averaging out many inconsistencies and leaving behind information in the signal that is common to all words that entered into the averaging procedure.

2.1.7 - General Analysis Strategy

The principal aim of this section is to temporally resolve how word length, frequency and class are encoded in the EEG signals recorded between 0 and 700 ms with respect to the on-screen onset of the stimulus during continuous single-word reading in an RSVP paradigm. This allows us to compare obtained results with those of previous research which have assessed the effects of word length, frequency and class in event-related potential (ERP) research during word-list and sentence reading (ter Keurs et al., 1999; Hauk & Pulvermuller, 2004; Hauk et al., 2006;

Osterhout et al., 1997; Osterhout & Holcomb, 1992; Sereno et al., 2020; Faísca et al., 2019; Münte et al., 2001; Dufau et al., 2015).

We can furthermore quantify the effect of trial-averaging on the temporal resolution across the 700 ms epoch (176 time points, where each time point represents 4 ms). We implement a sliding-window approach in order to obtain a temporally-resolved estimate of linguistic processing across the three primary linguistic variables under consideration in this experiment. Each window consists of 16 time points (64 ms) and is shifted along by 1 time point (4 ms) over the entire epoch. The resulting vector of accuracy scores contains 160 values. We repeat this over all levels of averaging for the three linguistic variables of interest (length, frequency & class) as well as over 10 *a priori* randomly-selected seed values. The same 10 seed values are used when running all experiments.

2.1.7.1 - Statistical Analysis

In order to calculate the statistical significance of the decoding traces, we take the 10 seeds for each experiment and extract the one that had the highest score on the dev set. We then extract the corresponding results on the test set, for this seed value that scored highest on the development set. We apply False Discovery Rate correction (Rouam, 2013) at $p < 0.05$ in order to correct for multiple comparisons. We then indicate the temporal windows which are significantly different from chance with respect to an alpha value of 0.05. For plotting, we take the test results obtained across all 10 seeds and calculate the mean as well as the 68% confidence interval.

2.1.8 - Results

The decoding results for the various pseudotrial levels (single trial, averages of 3 and averages of 10) for the linguistic variables of word length, frequency and class are given in Figure 2.6. The top plot associated with each linguistic variable is a butterfly plot of the EEG difference wave for the respective condition stated after the associated letter (**A**: length; **B**: frequency; **C**: class). For example, the butterfly plot for B - Frequency is High > Low, therefore the plot reflects the topography of all electrodes for the average of all high-frequency trials after having subtracted the average of all the low-frequency trials, over the 700 ms epoch window for each word-level epoch in the dataset. The bottom plots contain the three levels of pseudotrial averaging (black: averages of 10, dark grey: averages of 3 and light grey: single-trials). The topographies of two time points are also plotted to the right of each temporal depiction of the EEG data. The timing of these topographies is indicated by grey vertical lines in the butterfly plot. Statistically significant windows are marked in the respective trial-averaging colour above the decoding plot. Each accuracy point is plotted at the end of its 16 time point (64 ms) window. This means that an accuracy time point at 200 ms contains the result of applying the linear SVM on the window of data from 136 - 200 ms, therefore the time points are right-aligned with their window and as such represent results with a slight rightward shift in time.

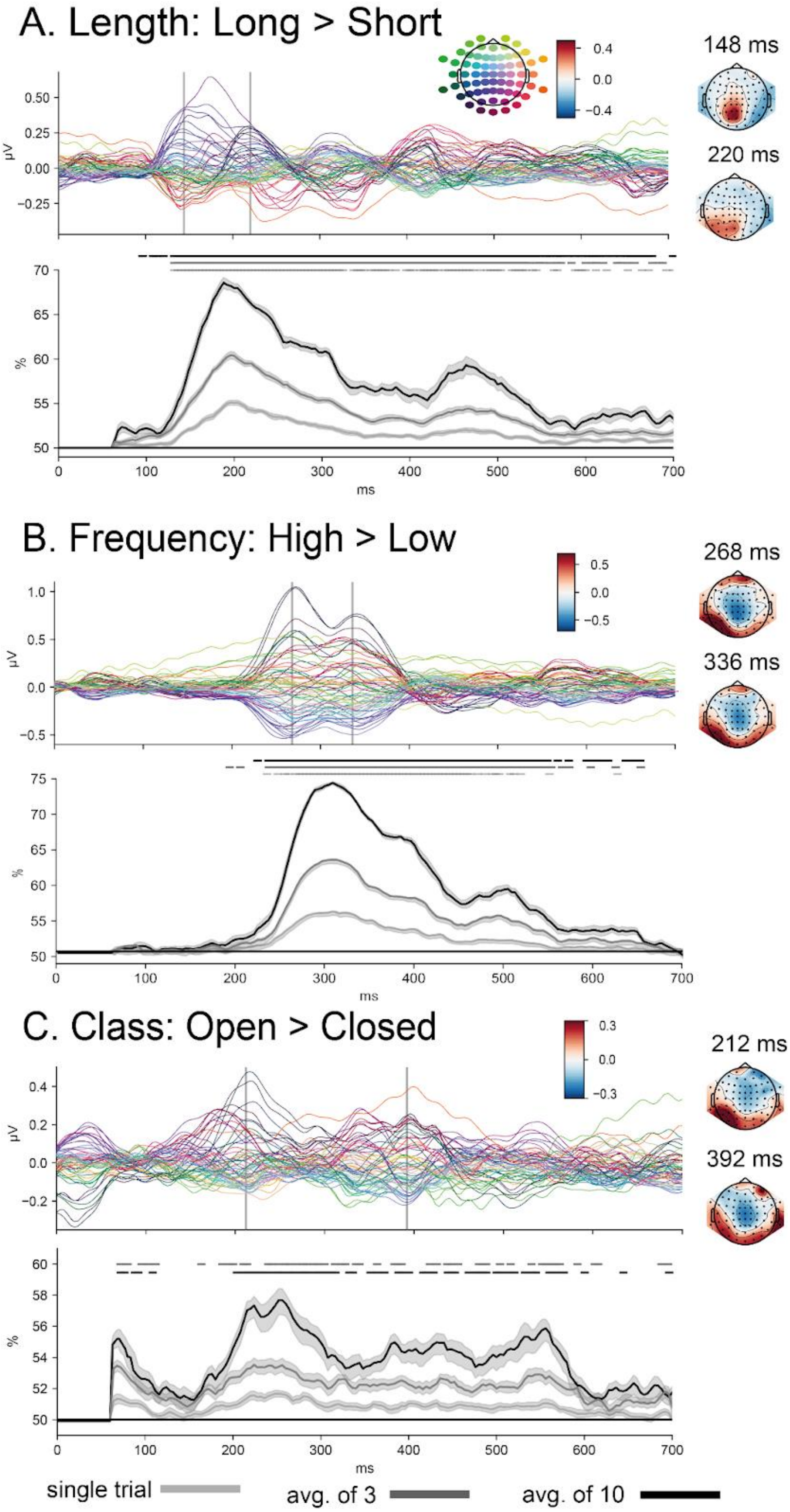

Figure 2.6. Butterfly plots for difference waves across all 64 channels, scalp topographies and time courses of decoding accuracy. (A) Length: LONG > SHORT, (B) Frequency: HIGH > LOW, (C) Class: OPEN > CLOSED. Each line in the butterfly plot represents a channel, colour-coded by its position. EEG topographies are given for the indicated points in the butterfly plots (vertical grey lines). The 68% confidence interval over the 10 runs for each of single trials, pseudotrials (averages of 3) and pseudotrials (averages of 10).

The decoding trace of the linear SVM models replicates the temporal cascade of word length, frequency and class effects previously reported for ERP responses averaged across a large number of trials. The word length effect arises earliest, around 120 ms post-stimulus onset, previously associated with visual word processing in occipitotemporal cortices (Hauk & Pulvermuller, 2004; Pulvermuller et al., 2009; Schuster et al., 2016). The topography at the onset of the word length (~ 148 ms) is centred over the visual cortex and then proceeds forwards along the left hemisphere (~ 200 ms). The effects of word frequency are seen to influence neural processing in a slightly later window, around 220 ms onwards, with a strong left-hemisphere predominance, as observed in earlier studies which examined the effects of word frequency (Griffiths et al., 2012). The word class effect is more sustained than the word length and frequency effect, with an early peak around 250 ms and a later one around 55 ms. The latter peak is in line with the well-known P600 ERP, an index of syntactic processing (Osterhout & Holcomb, 1992; Hagoort et al., 1993; ter Keurs et al., 1999). After 600 ms, the decoding traces drop off back to around chance performance (50%). The selected topographies show an earlier predominance in the left-hemisphere around 212 ms, while later windows around 400 ms show a more balanced distribution of positive ERP activity with respect to hemisphere, highlighting where open-class words elicit stronger responses than closed-class words.

Consistent with a recent report (King et al., 2020), word length and frequency were stronger than the word class effect and the unfolding of the temporal events matched well. This can be taken as converging evidence to support the magnitude and temporal ordering of the neural processes that occur during text reading across multiple neuroimaging modalities. As expected, decoding accuracy increased with the level of trial-averaging. Thus, carefully controlling each comparison of interest (e.g. word class) for the confounding effects of no interest (e.g. word length and word frequency in the case of focusing on word class effects) enabled us to dissociate word length, frequency and class effects, despite their high correlation in natural language.

2.1.9 - Discussion

The experimental results presented thus far show that an EEG dataset, collected on a corpus of natural language where common aspects of language such as word length, frequency and class are all highly confounded, can be split such that we can model these effects individually by correcting for the confounds. This allows us to uncover the known effects of cognitive processing which have been studied before in ERP research, both in terms of the temporal cascade but also of expected topographies that have been reported. We also show that the use of pseudotrials of varying levels can boost decoding accuracy, which is an important point to establish in current data because this point underlies methods in the following sections, namely the pre-training on pseudotrials to expose classifier models to higher SNR data. Furthermore, we show in our word class decoding that effects relating to part-of-speech exist in the data that are separate from the contributing effects of word length and frequency, which is an issue that has been raised in the literature on the status of such groupings of syntactic categories.

2.2 - Improving Training Methods for PoS decoding

2.2.1 - Introduction

After establishing that information relating to word class can be successfully decoded by a linear classifier, in which the binary task (open vs closed class decoding) was set up by collating equal proportions of different part-of-speech tags, we next explored two primary methods to help boost model decoding. These methods are commonly applied in the latest machine learning models, such as context-sensitive representation learning methods such as the Transformer model. We explore how decoding of part-of-speech in EEG signals can be enhanced in Transformer models by comparing their generalisation performance with training paradigms that do not employ these techniques. We furthermore contrast the effects of how these training tweaks differentially affect high-capacity Transformer models by also applying the same process to the linear SVM models of the previous section. Our implementation of the linear SVM model is tuned to mirror the training steps of the Transformer in order to facilitate such comparisons. We can therefore explore how these training methods interact with the model architectures and representational capacity.

The two techniques we look at in this section are (i) data augmentation methods and (ii) pre-training methods. For data augmentation, we bootstrap pseudotrials as described previously, but to much larger dataset sizes so that the classifier has access to a broader range of synthetic data derived from the training data. For pre-training methods, we focus on modelling single-trial data after having pre-trained on trial-averaged data, which has a higher signal-to-noise ratio. The rationale for this methodology is that by accessing higher quality data, when we fine-tune a model on the noisier single-trial EEG data, the model is able to generalise quicker as it has previously had access to less noisy data.

2.2.2 - Data set

The primary goal of our experiments in this chapter is to address the extent that part-of-speech can be decoded from EEG signals, so we therefore use the confound-corrected word class data from the previous section, namely the data that was sampled such that there was no systematic difference between word length and frequency between two binary classes (open and closed-class words). From this dataset, we extracted three PoS categories from the open-class dataset (NOUN, VERB, PROPN) and three PoS categories from the closed-class dataset (ADP, DET, PROP). Equal amounts of each PoS class were extracted, such that the dataset also contains an equal amount of open and closed class words overall. This resulted in a training set where each of the six PoS classes appeared in the training set 3,470 times, 335 in the development set and 335 times in the test set (total dataset size is approximately 20,000 data points). In order to ensure a more balanced distribution of words across the data splits, we selected only higher-frequency words, i.e. words which had a higher Zipf-score than the median value in the dataset (5.91). We also matched the development and test sets such that they had equal distributions of word lengths.

2.2.3 - Transformer Implementation

For the Transformer (Vaswani et al., 2017), we conducted a model architecture and hyperparameter search over model layers, learning rate, multilayer perceptron dimensionality, dropout rate, encoder vs encoder-decoder) on the development set. The selected model consists of four encoder-blocks and a final dense layer that projects the output of the last encoder-block onto the PoS classes via a softmax function. The Adam optimiser was used alongside an early-stopping procedure. The implementation of the model is based on the WMT example of Google's novel ML frameworks Flax / Jax. Table 2.4 lists the selected parameters. The Transformer received EEG channels x time points as inputs and provided a classification response per time point. We aggregated the classification responses across all time points into a single prediction via a majority-voting scheme.

parameter	value	parameter	value
encoder layers	4	MLP size	1024
learning rate	0.04	MLP dropout rate	0.1
batch size	16	QKV size	512
warm-up steps	50k	attention heads	8
training steps	400k	attentional dropout	0.1
Adam B1	0.9	Adam B2	0.98
Adan epsilon	10^(-9)	Adam weight decay	0.0

Table 2.4. Hyperparameters of the Transformer model

2.2.4 - Data augmentation

2.2.4.1 Method

Using this unigram part-of-speech dataset consisting of six classes, we assessed whether data augmentation via bootstrapping and re-averaging increases decoding performance for trial-averaged data (pseudotrials of 3-averaged single trials or 10-averaged single trials, as explored in the previous section). Concretely, we sampled either `{3,10}` single-trials from the same part-of-speech class in the training set and averaged them iteratively until we arrived at a specific overall dataset size. We generated 4 different training set sizes: `N_size = {20k, 100k, 250k, 500k}`, resulting in two (3 vs 10 trial-averages) by four (dataset sizes) = eight training sets. The lower value of 20k was chosen because it is the same size of the single-trial data, and this allows us to contrast single-trial performance and pseudotrial performance while keeping the data set sizes consistent.

A development set containing corresponding trial-averages for 3-averaged and 10-averaged data was also generated, which matched the number of our initial 335 x 6 = 2,010 sample single-trial dataset. Linear SVMs and Transformers were trained on the 8 training sets using 20 random seeds and the mean accuracy (and 68% confidence interval) for the development dataset, across these 20 seeds is shown in Figure 2.7.

2.2.4.2 - Results

Data augmentation systematically influenced the decoding accuracy of the Transformer but not in the case of the linear SVM. The Transformer's additional benefit from data boosting may result from its greater model complexity. In addition, the nonlinear activation functions between successive layers allow for the modelling of a superset of problems compared with that of the linear SVM. For both instances of trial-averaging, the Transformer's decoding accuracy on the development set increased for some levels of averaging (100k, 250k) with respect to the cardinality of the dataset size of the single-trial data (20k), which we define as our baseline result. For the latest augmented training set size (500k) we see a slight decline in decoding performance for the Transformer. This decay in performance may be explained by the increasing dependency of the training samples via continued bootstrapping.

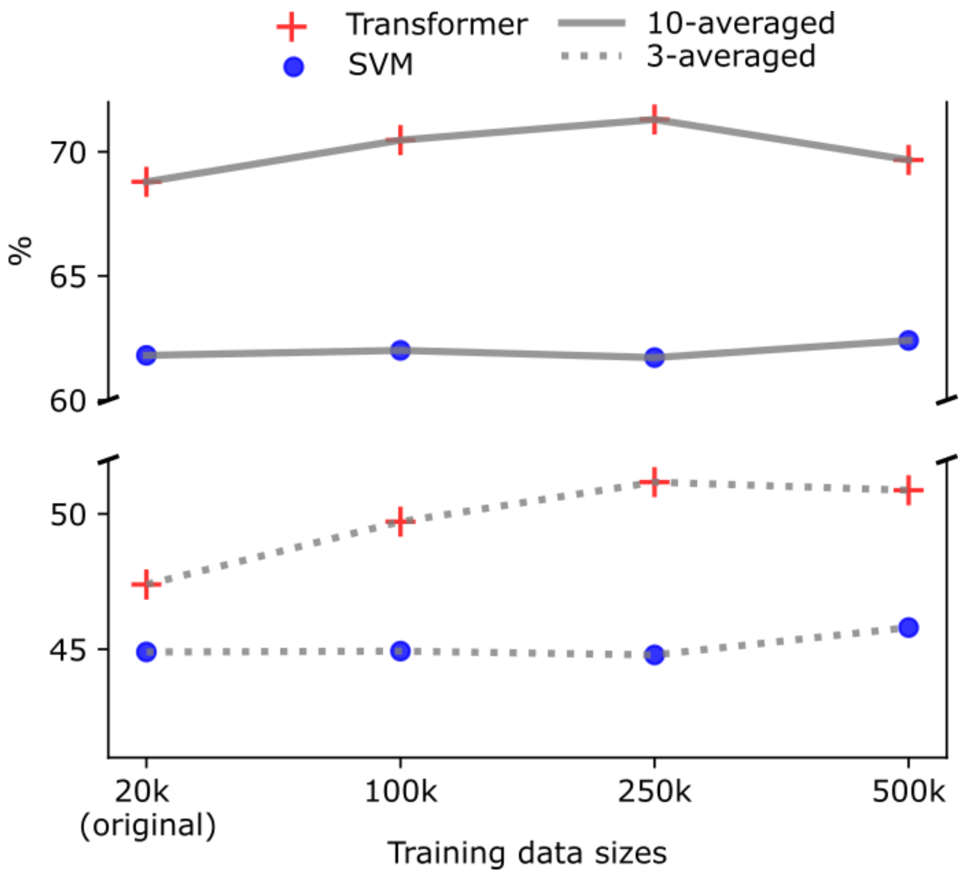

Figure 2.7. Data augmentation results on the development set. The Transformer has higher decoding performance when trained on augmented data, but up to a limit (250k) until performance drops again.

We formally assessed whether the Transformer's decoding accuracy was better using a training set with 250k augmented pseudotrials in comparison to using our baseline, i.e. 20k dataset which is the same size as the original single-trial dataset. To do this, we selected the model that performed best on the 250k development set, as well as the model that performed best on the 20k baseline development set, applying these models to test data that were trial-averaged via sampling without replacement. In the corresponding development sets, we sampled and averaged with replacement, such that the samples are no longer independent, therefore in order to respect the statistical assumptions of the tests we used, we applied the best performing models on pseudotrials of 3 and 10-averaged single trials. In both of these cases, the Transformer's decoding was significantly better for the larger 250k case, but not for the linear SVM. This was assessed using the Wilcoxon signed-rank test ($Z = 2.66$, $p < 0.01$).

2.2.5 - Pretraining

2.2.5.1 - Method

The ultimate goal of our work is to apply successful decoding techniques to single-trial data such that it can be potentially used in a live application, where we do not have access to averaged data in the same manner as assumed earlier in this chapter. We therefore assessed whether pretraining the SVM and / or Transformer models on trial-averaged data, with subsequent fine-tuning on single-trial data, affords an increase in generalisation performance and decoding accuracy over and above training purely on single-trial data. Pretraining may be beneficial because trial-averages have a greater signal-to-noise ratio, however, they attenuate true EEG variability across different words from the same PoS class.

Specifically, we assessed the impact of pretraining in a 2 x 2 factorial design manipulating (i) pretraining scheme: (a) training in three steps (10-3-1) from 10-

averaged single-trials to 3-averaged single trials, ending with single-trial data versus (b) training on two steps (3-1) from 3-averaged trials to single trials and (ii) data augmentation: training only on the baseline dataset (approximately 20k trials) versus augmented dataset (250k trials). We selected the 250k data augmentation as this option resulted in the highest development set accuracy in the previous section. We train both linear SVMs and Transformers on the 2 x 2 training conditions over 20 random seeds and report the mean development set accuracy and 68% confidence interval across those 20 seeds in Table 2.5.

2.2.5.2 - Results

For the linear SVM, the 3-1 pretraining scheme without data augmentation resulted in the highest accuracy on the development set (32.03%), which was marginally better than for training on single-trials directly (31.93%), showing that data augmentation for SVM models does not convey a boost in accuracy by virtue of having access to higher signal-to-noise data. In contrast to the linear SVM, the Transformer is conveyed with a benefit of having access to trial-averaged data. The highest decoding accuracy obtained for the Transformer was with the 3-1 pretraining scheme with 250k data augmentation, which obtained a score of 39.41%. Using the Transformer model with 3-1 pretraining & 250k data augmentation as well as the one trained on single trials alone, we ran both models on the test set and compared both accuracies using the Wilcoxon signed-rank test (Pereira et al., 2009), which confirmed that the Transformer performed significantly better on the test set after 3-1 (250k) pretraining with respect to just training on single-trials ($Z = 2.13$, $p < 0.05$).

	SVM	Transformer
single-trials	31.93 (+/- 0.62)	37.15 (+/- 0.32)
10-3-1	31.74 (+/- 0.51)	38.50 (+/- 0.28)
10-3-1 (250k)	31.89 (+/- 0.67)	39.17 (+/- 0.33)
3-1	32.03 (+/- 0.52)	37.83 (+/- 0.24)
3-1 (250k)	31.79 (+/- 0.58)	39.41 (+/- 0.41)

Table 2.5. Single trial decoding accuracies (%, mean across seeds +/- 68% confidence interval) on the development set for the linear SVM and Transformer models: without pretraining, with 10-3-1 pretraining, with 10-3-1 pretraining and 250k data augmentation, with 3-1 pretraining, with 3-1 pretraining and 250k data augmentation

2.2.6 - Discussion

The reported results show that ideas taken from mainstream machine learning research, such as pretraining and data augmentation, also convey advantages when applied to single-trial EEG decoding of linguistic information using Transformer models, while not for linear SVMs, which are likely constrained by a smaller class of representational capacity, for which these techniques do not play a large role. While pretraining in the NLP world largely consists of unsupervised learning over large volumes of texts, it remains to be seen whether large-scale pretraining of single-trial EEG data also benefits from such ideas, but in terms of artificially inflating the training set size, combined with first priming the model with cleaner data, we see significant increases in decoding accuracy. An open question remains in terms of why using 500k training samples suffered a worse score than 250k training samples. We speculate that the increased dependence on the training set samples has a slight negative effect and a threshold crossed in which pretraining can be maximised, given a task type and original training set size. This highlights the importance of experimenting with different sized datasets in order to extract maximum benefits from these additional training methods of neural networks on EEG data.

2.3 - Temporally resolved part-of-speech decoding

2.3.1 - Introduction

The primary motivation behind this section is to investigate how part-of-speech information is encoded in the neural EEG signal of unigrams and bigrams. This is interesting because a contrast can be established between single-word processing as well as simple multi-word expressions. Part of speech information in natural language involves combinatorial processes that progressively build up a mental representation and syntactic structure of our linguistic input. It's likely that these processes, which are over and above the processes of single-word processing, contain information that can be recorded in neural signals detected by EEG and therefore used in categorisation of these elements. The previous section of this chapter examined optimised training methods for decoding linguistic information from EEG data, from which we determined that the best combination of steps was to apply data augmentation to trial-averaged data (where each derived pseudotrial consists of a bootstrapped average of 3 single trials), pretrain a Transformer model on this higher signal-to-noise data and then fine-tune on our target distribution of single-trials.

For both unigram and bigram data, we investigated how Transformers, as well as our linear SVM baseline models, decode part-of-speech information as it dynamically evolves across post-stimulus time. This is achieved by using a sliding window of 64 ms. Furthermore, we explore how linguistic information in the form of lexical categories is integrated across the post-stimulus window by training on ever-increasing temporal windows. This allows us to compare our classifiers in terms of available data in terms of decoding from restricted windows as well as how models differ when ever-increasing contexts are available.

Previous sections aimed to match the distributions of examples from different word classes for word length and word frequency confounds in order to dissociate their distinct neural contributions to the EEG signals. This is critical, particularly from the perspective of cognitive neuroscience, in order to separate the different neural processing components involved in sentence reading. In this section, we take a more engineering perspective of natural language processing, whereby the length and frequency contributions of the signal are useful in hinting at part-of-speech identity. For example, short high-frequency words are more likely to be determiners (DET) and prepositions (ADP), while long low-frequency words are more likely to be nouns and verbs. Instead of treating these complementary factors as confounds that need to be controlled for, we extract the part-of-speech classes from the data as they exist in the naturalistic format they were recorded in, i.e. the six chosen PoS tags are taken directly out of the data, where confound correction has not been applied, thus information relating to word length and frequency is distributed according to the naturalistic text corpus.

2.3.2 - Unigram Analysis

2.3.2.1 - Dataset

Using the same dataset splits as outlined in Section 2.1.4, which were matched for text length, text genre and day of EEG recording, an equal amount of the 6 most frequent part-of-speech tags from the data was extracted. These classes were `NOUN, VERB, ADP` [adposition], `DET, PRON` and `PROPN` [proper noun]). Each part-of-speech class included 28,265 samples in the training set, 2,948 samples in the development set and 3,183 samples in the test set. Overall, the training set contained 169,590 samples, the development set contained 17,688 samples and the test set contained 19,098 samples.

2.3.2.2 - Methods

We implemented the 3-1 pretraining regime with 250k data augmentation as described in the previous section relating to the improvement of training methods. For the sliding window analysis, we trained and tested linear SVMs and our Transformer model on EEG signals from 64 ms segments of data (16 time points), which shift by 16 ms, from 0 to 700 ms. This process returns a vector of accuracies which contains 41 different points of evaluation. For the incremental window analysis, we ran an ever-increasing temporal window across the range of EEG data in steps of 16 ms (4 time points), starting with the [0,16] ms time window all the way up to 700 ms. This means the final window ranges from [688,700] ms and is 4 ms shorter than the rest. This process returns a vector of accuracies which contains 44 different points of evaluation.

We computed decoding accuracies from the means of 10 randomly selected seed points, alongside the corresponding 68% confidence interval, on the test set, for purposes of visualisation. Across time windows we compared the decoding accuracies on the test sets of the model which performed highest out of all seeds on the development set. We submitted these values to the Wilcoxon signed-rank test with an alpha level of 0.05, corrected for multiple comparisons using False Discovery Rate (FDR) correction across time, i.e. for 41 / 44 dimensional accuracy vectors. Those windows where the Transformer was statistically better than the linear SVM are indicated, unless all windows were significant, in which case no specific indication is given, but this is specified in the figure legend.

2.3.2.3 - Results

In the sliding window analysis (Figure 2.8 middle) the decoding accuracies of both the linear SVM and the Transformer show two prominent peaks around 200 ms and 400 ms, suggesting that part-of-speech decoding relies on several aspects of information encoded in the EEG. Based on the confound-controlled analysis in Section 2.1, the initial peak reflects word length and word frequency processing, while the later peak is more closely related to semantic and syntactic aspects of single-word / unigram processing.

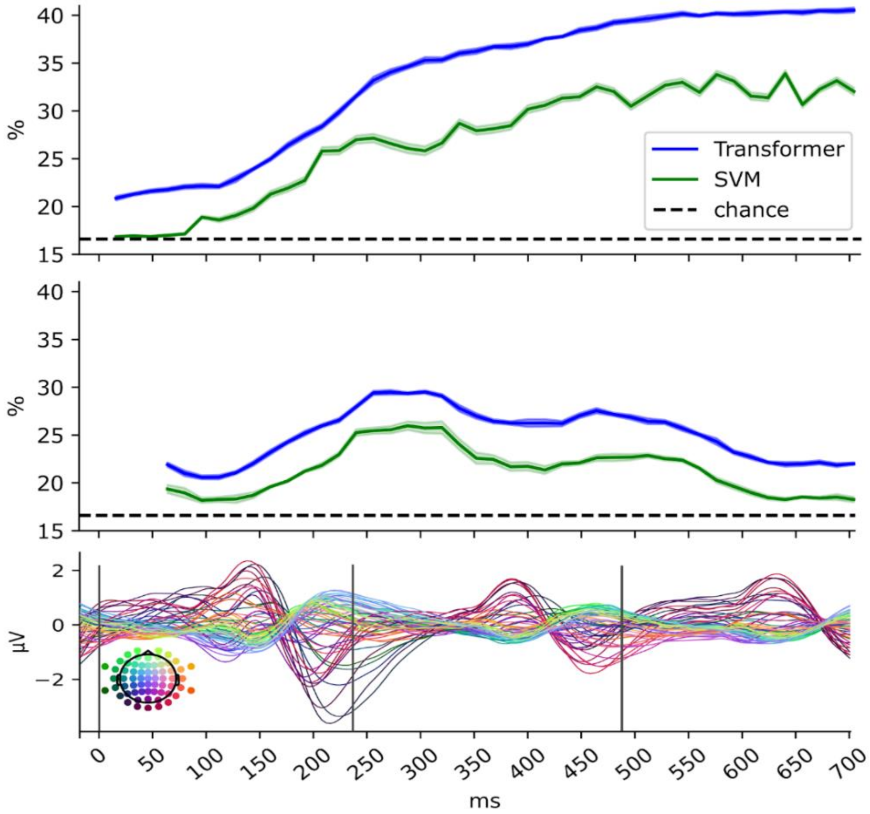

Figure 2.8. Unigram results: Test set decoding accuracies with the mean across 10 seed points, plus the 68% confidence interval), aligned with the last bin of each time window. Top: incremental window analysis. Middle: Sliding window analysis. Bottom: Average ERP of all NOUN trials in the training data to demonstrate a visual example of the unigram data. Vertical lines indicate the timing of on-screen word presentations. All time windows are significant

The Transformer significantly outperformed the SVM for all sliding time windows and all incremental time windows (Wilcoxon signed-rank test, FDR-corrected at $p < 0.05$). The incremental window analysis showed an accuracy benefit of 4.5% for the Transformer, over the linear SVM, starting in the first window of [0,16] ms. This difference in performance between the two classifiers then widened even further, reaching a peak difference where the Transformer scored 11.6% higher than the linear SVM at the corresponding time point. This appeared around 360 ms post-stimulus onset.

Transformers thus benefit from integrating information about word length, frequency and class, which are available in the signal at different post-stimulus latencies, as shown earlier. Moreover, because the part-of-speech of subsequent words is not an independent factor in natural language statistics, the Transformer’s self-attention mechanism may also rely on information about the subsequent words encoded in the EEG data from 240 ms onwards, which is when a new word is presented on screen to the subject.

2.3.3 - Bigram Analysis

2.3.3.1 - Dataset

We designed a bigram dataset that artificially removes the correlations of the part-of-speech of the following word (word 2) with respect to each first word (word 1) in the bigram. This was done in order to assess the distinct contributions of both word 1 and word 2 to the classification of the bigram as a whole. From the same data pool as the unigram analysis of the previous section, we ran an enumeration procedure to extract 3 classes (6 pairs) that could be equally balanced if both word 1 and word 2 could be reversed. Only 3 parts-of-speech appear in all combinations, namely NOUN, PRON and VERB. Concretely, this means our 6 classes are NOUN-PRON, NOUN-VERB, PRON-NOUN, PRON-VERB, VERB-NOUN and VERB-PRON. As a result of this selection, the part-of-speech of word 1 in the bigram is uninformative about the part-of-speech class of word 2 and vice versa. The benefits of this dataset are that we force the models to rely on integrated information across both words, instead of taking advantage of other confounding information as would have otherwise been the case. However, enforcing such a strong condition naturally reduces the size of the dataset available in the EEG corpus. Each bigram class has 3,470 samples in the training set, 322 samples in the development set and 349 samples in the test set. Overall, the training set contained 20,820 samples, the development set contained 1,932 samples and the test set contained 2,094 samples.

2.3.3.2 - Methods

The methods we employ in this section are exactly those described in Section 2.2. from the unigram analysis. Please see Section 2.2 for details.

2.3.3.3 - Results

In accordance with the unigram results, our sliding window analysis (Figure 2.9, middle) revealed two prominent peaks around 200 ms and 400 ms post-stimulus presentation. Yet, unlike the unigram results, the 2nd peak was later and of a higher magnitude than the first peak, compared with the unigram sliding window results. This is because the EEG signal at 500 ms incorporates not only semantic and syntactic aspects of word 1 (as in the case of the unigrams) but also contributes independent information about word 2 of the bigram, which by design cannot be provided by word 1.

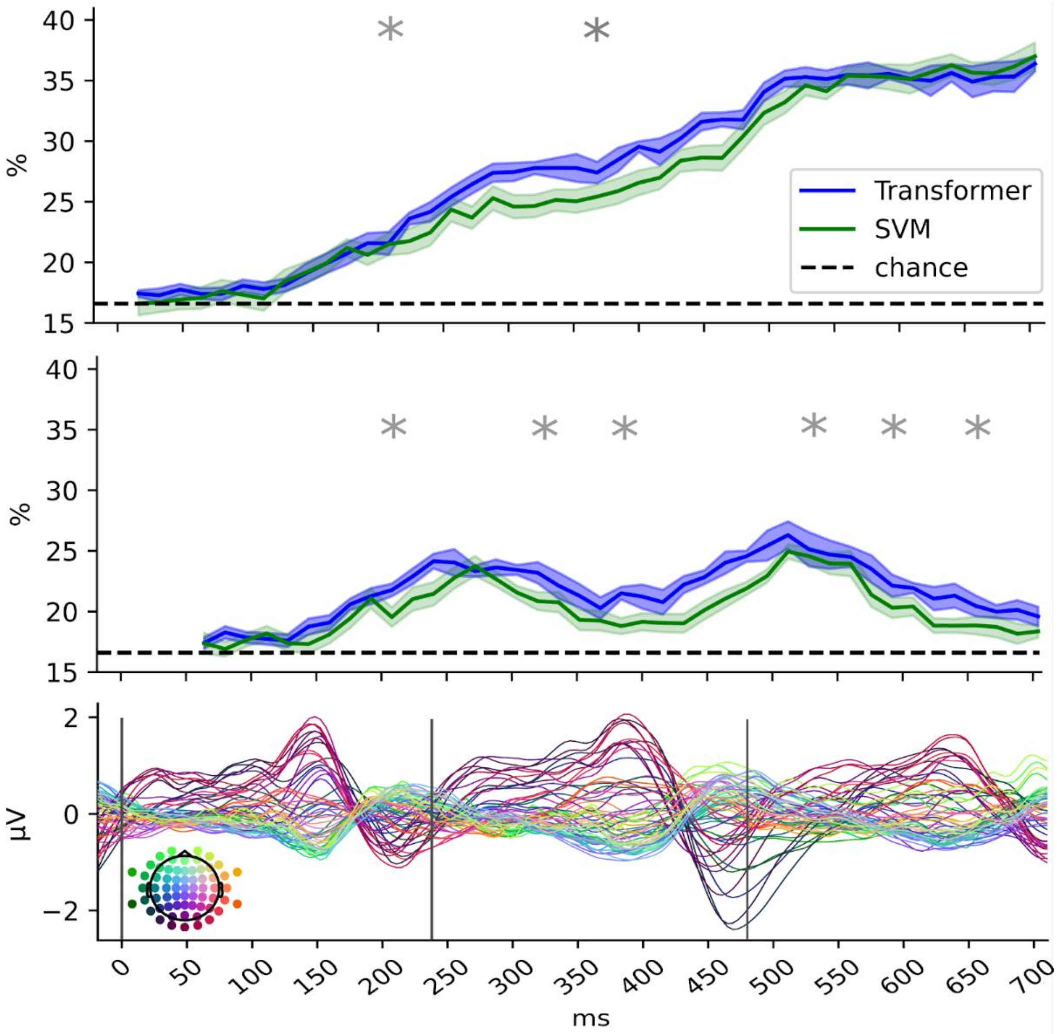

Figure 2.9. Bigram results: Test set decoding accuracies with the mean across 10 seed points, plus the 68% confidence interval), aligned with the last bin of each time window. Top: incremental window analysis. Middle: Sliding window analysis. Bottom: Average ERP of all VERB-PRON trials in the training data to demonstrate a visual example of the unigram data. Vertical lines indicate the timing of on-screen word presentations. Statistically significant time windows are indicated with asterisks

The Transformer significantly outperformed the linear SVM as indicated in Figure 2.9, where asterisks indicate the location of temporal windows where this is the case. This was confirmed by a Wilcoxon signed-rank test, FDR-corrected at $p < 0.05$. In the incremental window analysis (Figure 2.9, top) the Transformer outperforms the SVM at windows 0 - 208 ms and 0 - 336 ms. Yet, unlike for the unigram case, this performance benefit was no longer significant for time windows greater than 336 ms, suggesting that the benefit of the self-attention mechanism is smaller when independent information from both words needs to be additively combined. Our choice of balanced bigrams dramatically reduced the number of samples in the dataset, which may have affected the Transformer's performance more severely.

2.3.4 - Discussion

Both (i) unigrams and (ii) bigrams are decodable from single-trial EEG data, as demonstrated with these results. Both (i) and (ii) represent 6-class problems, for which chance accuracy is approximately 16.6%. For both sliding windows and incremental windows, we see peaks way above chance level. Furthermore, we also see a benefit of using Transformer models that specifically employ a self-attention mechanism that accumulates large contextually-dependent ranges of information into its modelling procedure. Given that the bigram dataset was selected such that it possessed a useful experimental factor, i.e. that of exact balancing of word 1 and word 2, such that identity of one part of the bigram was not informative of the other, necessitating the accumulation of larger windows of data (in the case of unigrams) or specific windows (in the case of bigrams, i.e. shortly after the presentation of word 2), it can be seen that classifier differences are minimal. We suspect that the trade-off with the experimental advantage of exact balancing with data set size has severely limited the Transformer to be able to learn from a wide range of samples. Effective data augmentation for Transformer-based single-trial EEG decoding appears to be linked with the data set size from which augmented data are generated, as demonstrated when 500k samples did not perform as well as 250k in the data augmentation analysis. The reduced pool of examples in the current bigram data, from which 250k samples were created for pretraining, is likely a limiting factor in this case. Future research which focuses on more direct comparisons of unigram and bigram modelling while fully taking advantage of the natural correlations in NLP texts between word length, frequency, previous and following PoS classes, will likely shed more light on this issue.

2.4 - Conclusion

The past three sections have looked at various aspects of modelling linguistic information and part-of-speech class from EEG data, as well as optimised training methods that can be used to boost Transformer-based classifier performance, which do not convey advantages to baseline linear models such as linear support vector machines. Combining neural signals measured at millisecond resolution with EEG and a linguistically annotated corpus, this chapter shows, to the best of our knowledge at the time of writing, the first time that unigram and bigram part-of-speech classes have been decoded directly from single-trial EEG data. Temporally-resolved EEG decoding is a useful tool to unravel how information about linguistic and non-linguistic aspects evolve dynamically over time.

In all experiments, Transformer models with a self-attention mechanism outperformed SVMs, particularly when the former were boosted via data augmentation and pretrained on higher signal-to-noise ratio data. This work provides an important steppingstone for future applications that incorporate human signals into traditional NLP methods. Applications such as part-of-speech induction jointly based on annotated texts and EEG signals could be transformative for corpus generation and tagging of gold-standard data for low-resource languages as well as a multitude of other live processing applications.

CHAPTER 3: EEG PREPROCESSING FOR DECODING OF LINGUISTIC STIMULI

CONTRIBUTIONS

All work in this chapter is the sole work of Alex Murphy, inspired by a research question posed by Uta Noppeney.

3.1 - Introduction

There are many different choices that one must make when taking neuroimaging data from its raw form to the highly-preprocessed form that is later analysed, displayed and interpreted. These choices often represent standard received wisdom in the field and in order to minimise differences between studies, to maximise the probability of successful replication, standard preprocessing pipelines and toolboxes promote such techniques. Differences between preprocessing pipelines from cognitive neuroimaging experiments can be substantive between different research institutions and lead to very different outcomes on the same data if different research labs employ different standardised preprocessing scripts (Botvinik-Nezer et al., 2020).

The principal goal of this chapter is to explore the interaction between specific preprocessing choices during the preprocessing of EEG data, in terms of how they affect decoding accuracy across different types of linguistic stimuli. The motivation for this analysis derives from observations from eye-tracking research that have demonstrated strong associations between aspects of word reading and eye movements. This has been shown for word length (Degno et al., 2019), word frequency (King & Kutas, 1995) and part-of-speech (Barrett et al., 2016). A critical point that needs to be more formally assessed is, given that such associations exist, if one's goal is to decode linguistic information from neuroimaging data such as EEG, standard preprocessing techniques often specifically target features of the data that are perhaps useful for downstream tasks such as classification or regression. If ocular artifacts are removed as a matter of standard preprocessing pipelines, yet this information proves to be useful in tasks of linguistic decoding, perhaps a more carefully-designed preprocessing strategy could be implemented, which is tailored to the end goal more specifically. This chapter addresses how such preprocessing choices affect the decoding of length, frequency and word class information.

This chapter is split up into two principal sections. The first examines the effect on our preprocessing choices on temporal sliding-window decoding analyses for each of the effects of interest (length, frequency and class) in a binary classification paradigm. The second section is an exploratory investigation that aims to quantify how the preprocessing effects lead to the various differences observed in the first section.

3.2 - Effects of EEG preprocessing on temporal decoding

3.2.1 - Introduction

Kornrumpf et al. (2016) studied the various ERP effects that arise during natural reading (with recorded eye movements in the signal) and RSVP (which avoided such responses) and found that small eye movements greatly facilitated lexical processing of language as it was being read. One claim was that lexical load could be modulated in the previous word, leading to facilitatory effects if eye movements were allowed (i.e. naturalistic paradigm instead of RSVP). With such predictive effects from immediately preceding words influencing lexical load in terms of word frequency, it raises the question whether epoch-level baseline correction might remove information in the EEG signal, still present during RSVP, therefore leaving open the question whether different forms of baseline correction (or none at all) might allow the surrounding linguistic environment to better predict linguistic information such as whether a word is high- or low-frequency. Furthermore, it has been reported that eye-tracking information has a positive correlation with decoding linguistic properties of stimuli during reading (Barrett et al., 2016). Taken together, this leads to an interesting research question which examines the effects of various EEG preprocessing steps on the decoding of linguistic information from stimuli during reading.

3.2.2 - Methods

3.2.2.1 - Preprocessing Choices

The preprocessing options examined in this chapter relate to:

1. Independent Component Analysis correction
2. Baseline correction of epochs
3. Feature scaling

3.2.2.2 - ICA

We consider: (i) no ICA correction, (ii) weak ICA correction and (iii) strong ICA correction. ICA correction is defined as 'weak' when a maximum of 4 primary noise components are removed. ICA correction is defined as 'strong' when up to 10 primary noise components are removed. A component is defined as a noise component when, upon examination of the topography and time series, it relates primarily to electrode noise, eye movements / blinks or muscle activity. We use MNE-Python's standard implementation of ICA decomposition, which first whitens the data, applies PCA and then applies the FastICA implementation of the ICA decomposition. A visualisation of the ICA solution for a randomly-selected session of EEG data is given below in Figure 3.1.

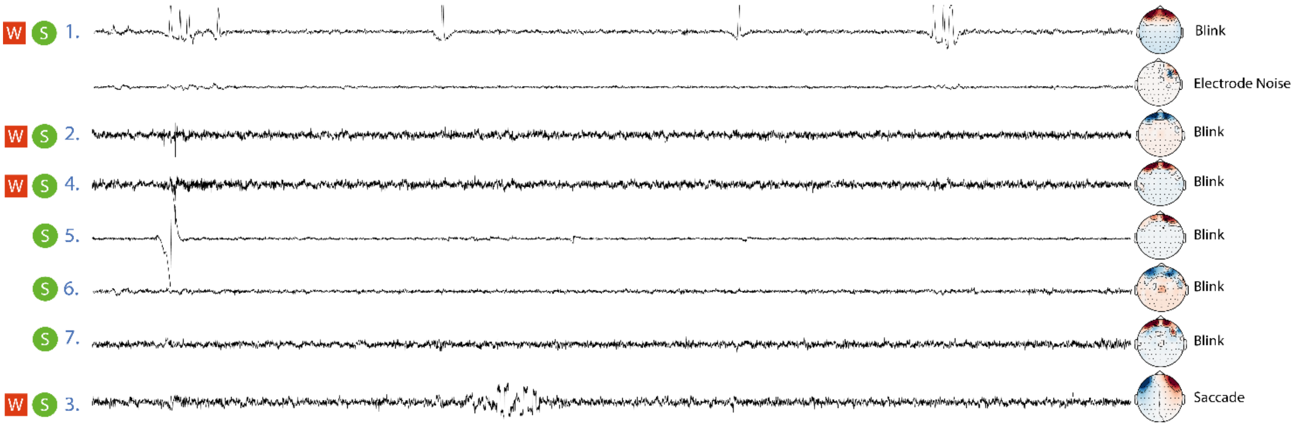

Figure 3.1. An example of a typical session's ICA decomposition. The 8 highest-magnitude independent components are plotted alongside their topographies (to the right) over 20 seconds of EEG recording. Each is accompanied by a label identifying a characterisation of the component. To the left, a removal order is present according to the description in the text. A red square identifies those components removed in the weak-ICA correction condition and a green circle identifies those components removed in the strong-ICA correction condition. Effects of blinks are primarily in the first component (39.60% explained variance), but are also spread among other components that are mostly characterised by inactivity, except for high-magnitude bursts throughout the recording session

In the case of preprocessing the EEG data for *weak ICA correction*, where a maximum of 4 noise components were removed, the first step was to assess the time series and topography of the two largest-magnitude independent components (top two rows in Figure 3.1) and mark these for deletion if they exhibited characteristics related to either electrode noise or oculomotor signals (blinks or saccades). The two highest-magnitude components were almost invariably related to blink artifacts, characterised by a broad symmetric frontal distribution of activity which was consistent in polarity i.e. always positive or always negative for that component, with frequent sharp deflections in the time course caused by the upward rotation of the eyes during blinking. After these were identified, then the next artifact components to be removed were ones that exhibited lateral eye movements (i.e.

saccades), which are characterised by large frontal and distinct patches of activity of opposite polarity, caused by repeated lateral shifts of the eye’s dipole into different visual fields. An example of such a topography is given in Figure 3.1 above (component 11). If there were at least two of these, this would make up the correction procedure for weak ICA correction. If there were only a single instance of a lateral component (as above) then the next highest-magnitude component relating to blinking was removed.

The method for strong ICA correction is comparable to the procedure for weak ICA correction above, except for two minor changes: (i) an allowance for the removal of up to 3 lateral eye-movement components was given, but the decomposition typically never created more than 2 and (ii) after returning to the highest-magnitude independent components, blink or other strong artifacts were removed until the total of 10 were removed. In this case, however, if the eye movements had been modelled relatively successfully, many fewer components than 10 would be removed and this would still be classed as a case of strong ICA correction. The case where oculomotor artifacts are spread over multiple high-magnitude components and then removed is very similar to the situation where these effects were captured in fewer components.

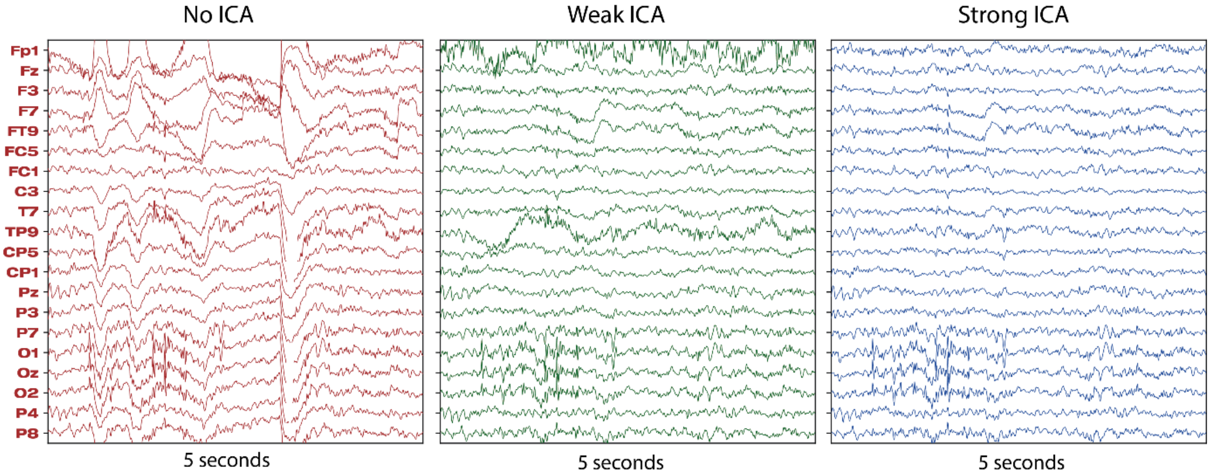

Figure 3.2. A selection of frontal, occipital and temporoparietal electrodes around a large eye movement (seen most clearly in the no ICA condition to the left in red), alongside weak and strong ICA correction of the same 5 seconds of EEG. Strong ICA correction (right; blue) removes major ocular artefacts, while the weaker version (middle; red) removes the strongest of these, with remaining eye movement information present but not as pronounced as in the case without ICA.

Figure 3.2 demonstrates the downstream effects of ICA-correction on the preprocessed EEG data of a single 5 second segment from a randomly-picked (yet representative) recording session with the three types of ICA correction outlined above, surrounding a prototypical typical eye movement seen in the dataset. As more oculomotor components are removed, the reconstructed EEG data suffers less from artefacts rooted in eye movements. As only frontal components were removed, bursts of activity over occipital electrodes remain consistent, as shown in Figure 3.2.

3.2.2.3 - Baseline Correction

We consider: (i) no baseline correction, (ii) sentence-level baseline correction and (iii) epoch-level baseline correction. Sentence-level baseline correction consists of subtracting, from each derived epoch for a sentence, the electrode-wise mean of the immediate 100 ms prior to the sentence start. Epoch-level baseline correction is defined similarly, but the average 100 ms electrode-wise values are immediately prior to the on-screen appearance of a word stimulus on screen. The issues surrounding baseline correction and the potential effects it has on EEG decoding are discussed in Section 1.4.7.

Feature Scaling

We consider: (i) standardisation and (ii) multivariate noise normalisation. By standardisation we refer to the common practice of removing the channel-wise mean from each electrode series and dividing by the standard deviation such that the scaled time series has unit variance. For (ii) we refer to the epoch-level implementation of multivariate noise normalisation outlined in Guggenmos et al. (2018). See section 1.4.8.2 for implementational details.

3.2.2.4 - Dataset

In the previous chapter we introduced the idea of confound-correction as a measure to equally balance two classes in a binary classification problem, each with respect to a balanced joint histogram over confounding variables. For example, we derived a dataset consisting of two classes relating to word-length, namely a class of EEG trials where the word that appeared on screen had 4 or fewer characters (which we call the `SHORT` class) and those with more than 4 characters (which we call the `LONG` class). In both of these classes, we ensured that the exact distributions of word class, a fine-grained discretised frequency score and position in sentence (start, middle, end) were equal among both the `SHORT` and `LONG` classes. Please refer to Section 2.1.6 for the full description of the data set and acquisition procedure, alongside the confound-correction procedure that was implemented to create these binary splits.

We use single-trial data and pseudotrials each derived from an average of 10 single trials. Importantly, for the test dataset in the pseudotrial case, averaging is done by random sampling *without replacement* (as outlined in Section 2.1.6.2), which means that each data point in the test set of the pseudotrials is independent of all other samples, i.e. each test set data point contributes only to a single averaged pseudotrial. Table 3.1 displays the number of data points for each of word length, frequency and class groups.

	length	**frequency**	**class**
train	82,424	51,364	45,502
development	12,402	7,632	6,658
test (single-trial)	10,810	6,590	5,670
test (avg. 10)	1,081	659	567

Table 3.1**.** The number of data points across the linguistic variables of word length, word frequency and word class, across the training set, development set, and test sets. The training and development sets in the pseudotrial case are resampled with replacement to match the same size as the single-trial dataset, but not for the test set. Thus, two rows are required to state the differing number of samples in this case.

3.2.2.5 - Analysis Strategy

We employ the same analysis as in the previous chapter, i.e. with a 64 ms (16 time point) sliding window shifted by 16 ms (4 time points) at every step. We vary the input into the model in terms of the amount of ICA correction in each dataset. We keep the same random seeds in the data generation process when running sliding window classifiers over the data, such that the principal difference in any reported results stems directly from the status of the data with respect to the ICA level of processing we applied to it.

The analysis is performed over both single trials and pseudotrials consisting of averages of 10 single trials. The previous chapter outlined how high signal-to-noise ratio data is useful for model pre-training and therefore the choice of preprocessing of this data is also pertinent to maximising the expected generalisation when using trial-averaged pseudotrial pretraining methods. In order to test whether observed results generalise over to the test set, a 3 x 2 x 3 factorial ANOVA will be performed separately for (i) high-SNR pseudotrial data and (ii) single-trial data, in which temporal windows that contain an effect of interest (as defined by observing the associated development set figures) are calculated over the 10 runs of the test set, the mean of which will serve as the dependent variable of the ANOVA. Tukey's Honest Significant Difference test will be applied where appropriate to examine the pairwise contrasts at an alpha level of 0.05, using the Familywise Error Rate to control for multiple comparisons.

3.2.3 – Results

The effects of manipulating various preprocessing options on the task of temporal decoding (via a sliding window analysis) are presented in this section. The results on the development data are first given for each of: (i) word length, (ii) word frequency and (iii) word class, as outlined above. The subsequent section collects notable results from the observations on the development set and formally assess them via statistical tests using completely independent test data, in order to see which results can be expected to generalise to novel data and are not due to idiosyncrasies during training and / or overfitting on the development set. The highest signal-to-noise ratio trial-averaged data is also included here, since effects can be more pronounced thanks to the boosted signal. Since we have shown that decoding is boosted by optimised training methods, such as pretraining on higher SNR / trial-averaged data (Section 2.2), a focused attention on these results can help to inform machine learning approaches to decoding various linguistic features from EEG signals. For each of our variables of interest (word length, frequency and class) the sliding window decoding traces are given. The top traces refer to pseudotrials (pseudotrials consisting of averages of 10 single trials) and the bottom refer to single-trials. Each figure modulates the effect of baseline correction (blue = epoch baseline correction; red = sentence baseline correction; green = no baseline correction) for a specific setting of feature scaling and ICA correction.

3.2.3.1 - Word length

For the confound-corrected word length data, the temporal decoding traces that result from each setting of the preprocessing pipeline are given below in Figure 3.3.

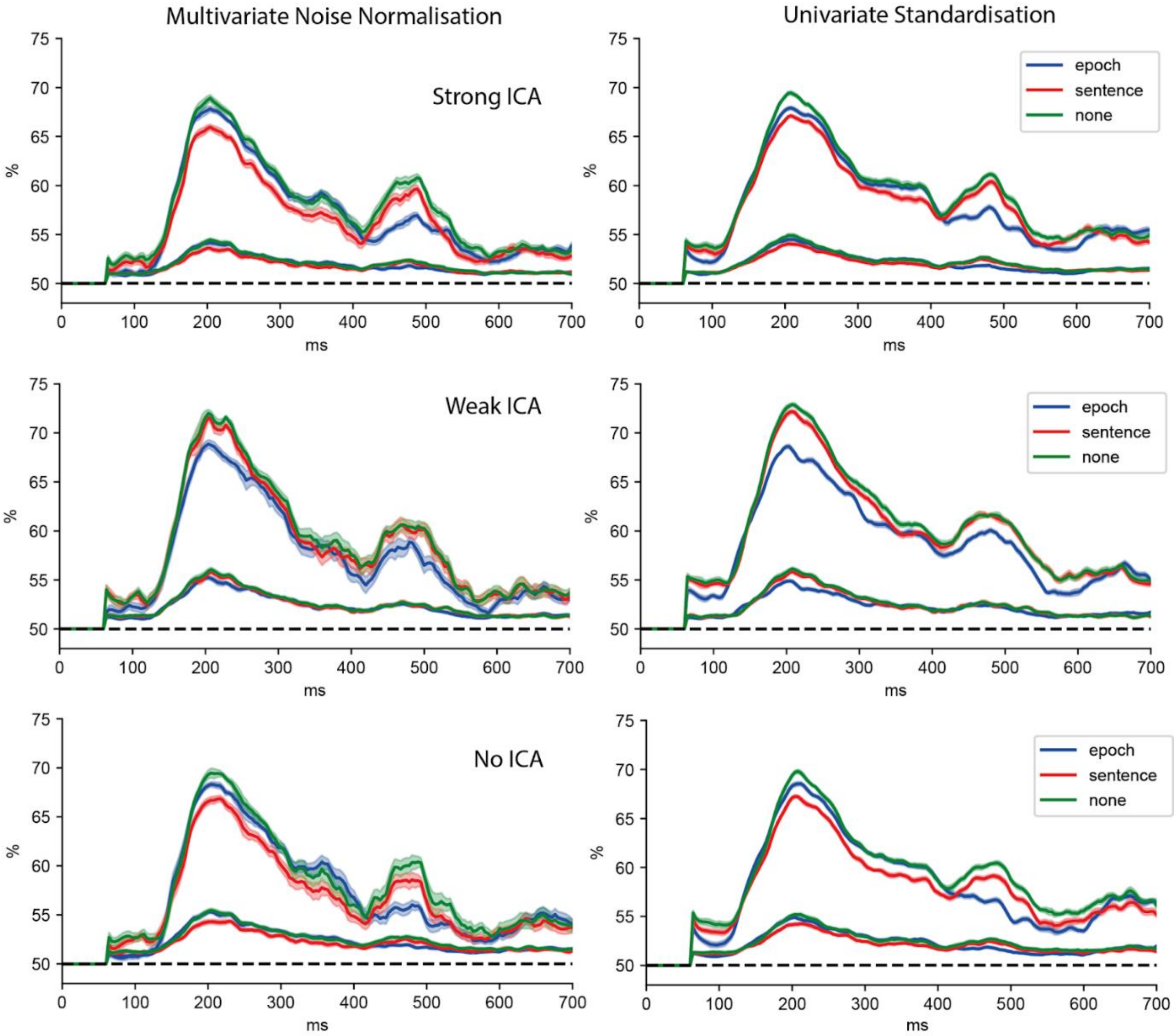

Figure 3.3. Development set sliding window decoding traces for word-length decoding. The top three traces of each sub-figure refer to 10-averaged pseudotrials and the bottom three refer to single trials. Each figure modulates the baseline correction method over a specified setting of ICA correction (rows) and feature scaling (columns), i.e. the first column contains 3 plots which have all been scaled via MNN, and the second column, univariate scaling

The decoding traces across both types of feature scaling (MNN vs univariate standardisation) are largely similar, with a noticeable effect being that with univariate standardisation, the confidence intervals around each mean decoding trace is tighter when compared with the corresponding traces using multivariate noise normalisation. For EEG data that has had a large number of ICs removed during preprocessing, sentence-level baseline correction appears to perform worst in the early peak around 200 ms. When no baseline correction is applied, there is higher sensitivity to the length response of the following word, around 460 ms (stimulus SOA is approximately 240 ms and this response arises 200 ms afterwards). Subtracting the mean of each channel directly preceding stimulus onset appears to reduce the sensitivity of this response. Both decoding traces for single-trial data in this case are very similar.

For weak ICA-corrected EEG data, no discernible difference arises between no baseline correction and sentence baseline correction, which both perform better than epoch baseline correction. The peak accuracy over both forms of feature scaling (middle row) is highest here, over all other combinations of scaling, baseline correction and ICA correction, with the exception of epoch baseline correction. When no ICA is applied to the data, the decoding traces are in line with the other plots, where again we see a sensitivity to the presumed decoding of the word length information of the following word. We also see that using epoch baseline correction reduces the sensitivity to this with respect to the other forms of baseline correction. A final observation is that when univariate standardisation is used, the peaks after 200 ms decay more slowly, which is likely due to the fact that this form of feature scaling isn’t sensitive to noisy channels and might be fitting more to noise in the data as these results are reported on the development set. When channel noise is taken into account, as is the case with multivariate noise normalisation, the peaks drop off more rapidly.

3.2.3.1.1 - Statistical analysis

A number of formal hypotheses can be derived from observing the plots over the development set data, which lead directly to predictions we can make and then assess on our independent test data set. The plots for temporal decoding of word length on the development set suggest the following:

1. Over the main word length peak (180 - 250 ms) in weakly-corrected ICA, epoch baseline correction performs worse than either no baseline correction or sentence baseline correction
2. When no ICA correction is applied, the period between 200-400 ms shows that no baseline correction outperforms sentence baseline correction overall
3. For all levels of ICA correction, not using baseline correction results in the highest sensitivity to word length decoding of the following word between 450-500 ms

The corresponding decoding traces on the test set are given in Appendix B.

3.2.3.1.2 - Pseudotrials

<u>180-250 ms</u>

For the 180-250 ms window, a significant interaction effect was found between ICA correction and baseline correction (F(4,15) = 26.9, $p < 0.05$, $\eta_p^2 = .00082$). Post-hoc tests (Tukey’s HSD, FWE = 0.05) confirmed that for weakly-corrected ICA, that contrasts involving epoch baseline correction with (i) sentence baseline correction and (ii) no baseline correction were significant also in the test data. This supports the hypothesis (1) from the development data.

200-400 ms

A significant interaction between ICA type and baseline correction was found in this time window, as had been suggested by observing the development set data ($F(2,15) = 25.84$, $p < 0.05$, $\eta_p^2 = .00092$). Post-hoc tests verified that it was exactly the case when no ICA correction or baseline correction was applied.

450-500 ms

In the 450-500 ms time window, there were no significant interactions between ICA correction type, baseline correction type or feature scaling type, but the main effects of ICA correction and baseline correction types were statistically significant (ICA correction: $F(2,15) = 148.7$, $p < 0.05$, $\eta_p^2 = .0034$), baseline correction: $F(2,15) = 150.6$, $p < 0.05$, $\eta_p^2 = .0035$). Post-hoc tests (Tukey's HSD, FWE = 0.05) confirmed that the no baseline-correction condition was responsible for the significant difference between baseline correction group means. Similarly, for ICA-correction, the no-ICA condition was responsible for the significant difference between ICA-correction group means. This result supports the observations made on the development set for this temporal window.

3.2.3.1.3 - Single trials

180-250 ms

No significant interactions were found in this time window, but all main effects were significant: ICA ($F(2,15) = 289.41$, $p < 0.05$, $\eta_p^2 = .00082$), baseline correction ($F(2,15) = 37.38$, $p < 0.05$, $\eta_p^2 = .00018$), scaling ($F(1,15) = 70.87$, $p < 0.05$, $\eta_p^2 = .00011$). Post-hoc tests (Tukey's HSD, FWE = 0.05) confirmed that the no-ICA condition was significantly different from both weak ICA correction and strong ICA correction. For baseline correction, the contrast between no baseline correction and sentence baseline correction was the source of the significant effect.

200-400 ms

This time window revealed significant main effects for all three groups (ICA: F(2,15) = 560.14, $p < 0.05$, η_p^2 = .00043), baseline correction (F(2,15) = 67.09, $p < 0.05$, η_p^2 = .000052), scaling (F(1,15) = 59.35, $p < 0.05$, η_p^2 = .000023). During post-hoc testing with Tukey's HSD, the pairwise contrasts for the baseline group did not provide sufficient evidence to reject the null hypothesis. For ICA-correction, it was found that the no ICA vs strong ICA contrast was significant, as well as the strong ICA vs weak ICA contrast.

450-500 ms

In the 450-500 ms time window, only the main effects of baseline correction and ICA correction type were significant: ICA (F(2,15) = 300.33, $p < 0.05$, η_p^2 = .00026), baseline correction (F(2,15) = 132.61, $p < 0.05$, η_p^2 = .00013). Post-hoc tests (Tukey's HSD, FWE = 0.05) confirm that weak ICA correction versus the other types is a significant effect, as well as epoch baseline correction versus no baseline correction.

3.2.3.2 - Word frequency

For the confound-corrected word frequency data, the temporal decoding traces that result from each setting of the preprocessing pipeline are given below in Figure 3.4.

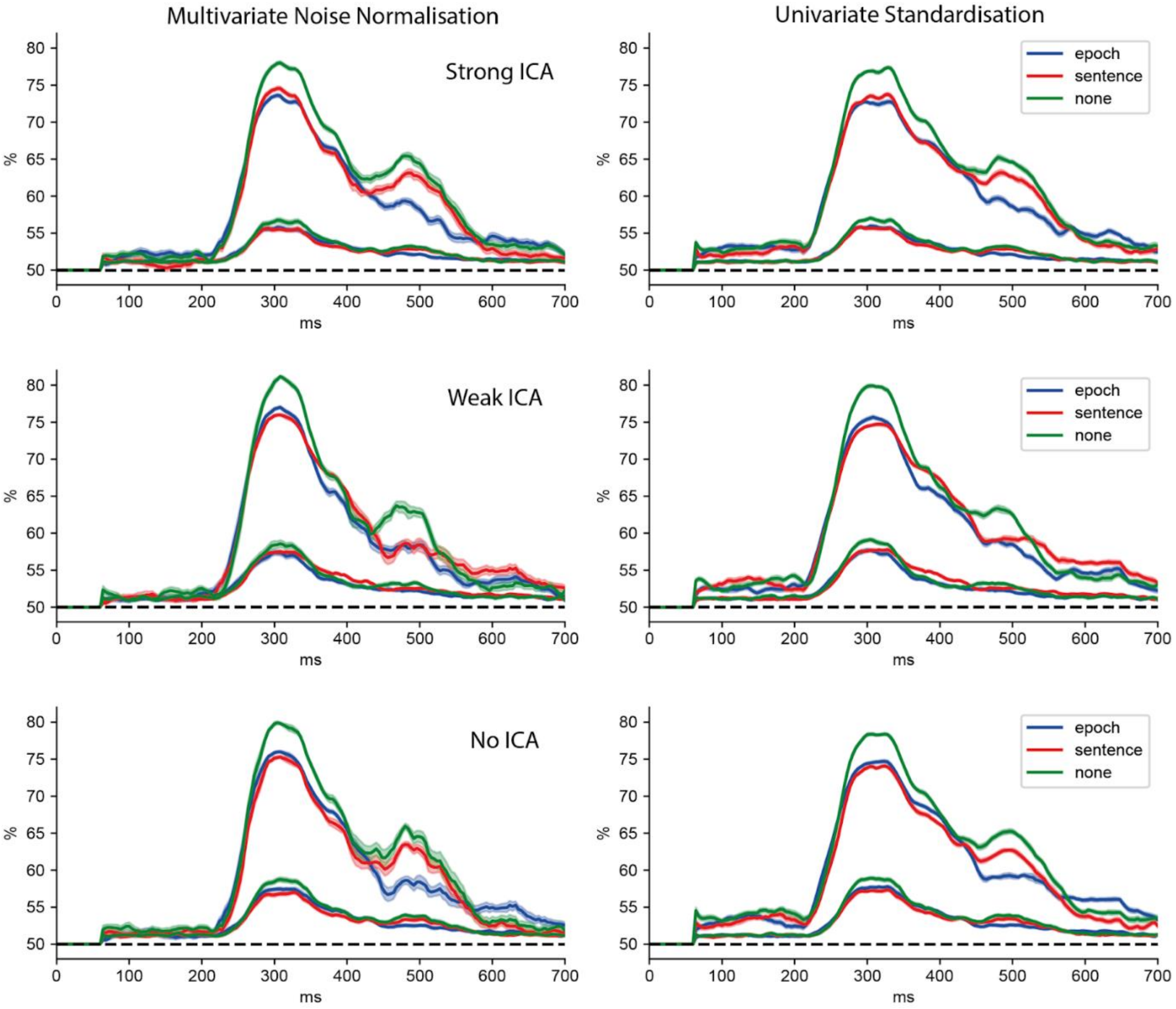

Figure 3.4. Development set sliding window decoding traces for word frequency decoding. The top three traces of each sub-figure refer to 10-averaged pseudotrials and the bottom three refer to single trials. Each figure modulates the baseline correction method over a specified setting of ICA correction (rows) and feature scaling (columns), i.e. the first column contains 3 plots which have all been scaled via MNN, and the second column, univariate scaling

A result that appears across every combination of preprocessing settings is that not applying baseline correction results in the highest overall peak decoding, alongside highest sensitivity to either late effects or responses to the next word (which appear on screen ~ 240 ms). Across the three levels of ICA correction, when strong ICA correction was applied, peak decoding was a few percentage points lower than weak ICA correction or no ICA correction. Around the main decoding peak at 280-400 ms, for all baseline correction methods and across all ICA correction types, sentence baseline correction is very close to epoch baseline correction. For long sentences, where the distance from each word to the pre-sentence baseline correction period is long, drifts can accumulate which give rise to similar patterns as if no baseline correction were applied at all.

3.2.3.2.1 - Statistical analysis

The plots for temporal decoding of word frequency on the development set suggest the following:

1. Using multivariate noise normalisation results in higher peak decoding accuracy when there is no baseline correction over the peak response period to frequency (280-320 ms)
2. In the same period as (1), the no baseline correction achieves the highest peak across all levels of ICA-correction type and feature scaling choice
3. There is a differential sensitivity to decoding the following word in the 450-500 ms time period across all levels of baseline correction when strong / no ICA-correction is applied, but only between no baseline correction vs sentence-based and epoch-based when the data is weakly corrected with ICA

The corresponding decoding traces on the test set are given in Appendix C.

3.2.3.2.2 - Pseudotrials

280-320 ms

All three possible interactions among groups showed that there a statistically significant effect present: ICA-baseline: ($F(4,15) = 57.28$, $p < 0.05$, η_p^2 =.00038), ICA-scaling ($F(2,15) = 24.45$, $p < 0.05$, $\eta_p^2 = .0008$), baseline-scaling ($F(2,15) = 12.44$, $p < 0.05$, $\eta_p^2 = .0004$). Post-hoc tests (Tukey’s HSD, FWE = 0.05) confirmed that 5 out of the 6 pairwise contrasts involving MNN scaling paired with no baseline correction were significant, supporting observation (1) above. Further tests on the interactions determined all pairwise contrasts in which no baseline correction was paired with another baseline correction method were statistically significant.

450-550 ms

A significant interaction effect was found for ICA-correction type and baseline correction type (F(4,15) = 8.99, $p < 0.05$, $\eta_p^2 = .00024$). The main effect of scaling, not present in this interaction, was found to be non-significant in this time window. The significant interaction is expected given observation (3) above. Pairwise post-hoc tests (Tukey's HSD) determined that the exact nature of the effects observed in (2) over the development set did not carry over to the test set, but various pairwise contrasts did show statistical differences with each other.

3.2.3.2.3 - Single trials

280-320 ms

No interactions were significant that concerned feature scaling, which was significant as a main effect (F(1,15) = 33.82, $p < 0.05$, $\eta_p^2 = .000063$), which supports observation (1) above. There was a significant interaction between ICA-correction and baseline correction (F(4,15) = 12.03, $p < 0.05$, $\eta_p^2 = .000089$). Tukey's HSD confirmed that the contrasts involving no ICA-correction and no baseline correction were largely behind the significant interaction.

450-550 ms

The main effects of baseline correction and ICA-correction were found to contain significant differences among their group means: ICA (F(2,15) = 184.32, $p < 0.05$, $\eta_p^2 = .0003$), baseline (F(2,15) = 172.99, $p < 0.05$, $\eta_p^2 = .00028$). Post-hoc tests (Tukey's

HSD) revealed that all mutual pairwise contrasts of baseline correction were significant, while none of the ICA-correction contrasts were.

3.2.3.3 - Word class

For the confound-corrected word frequency data, the temporal decoding traces that result from each setting of the preprocessing pipeline are given below in Figure 3.5.

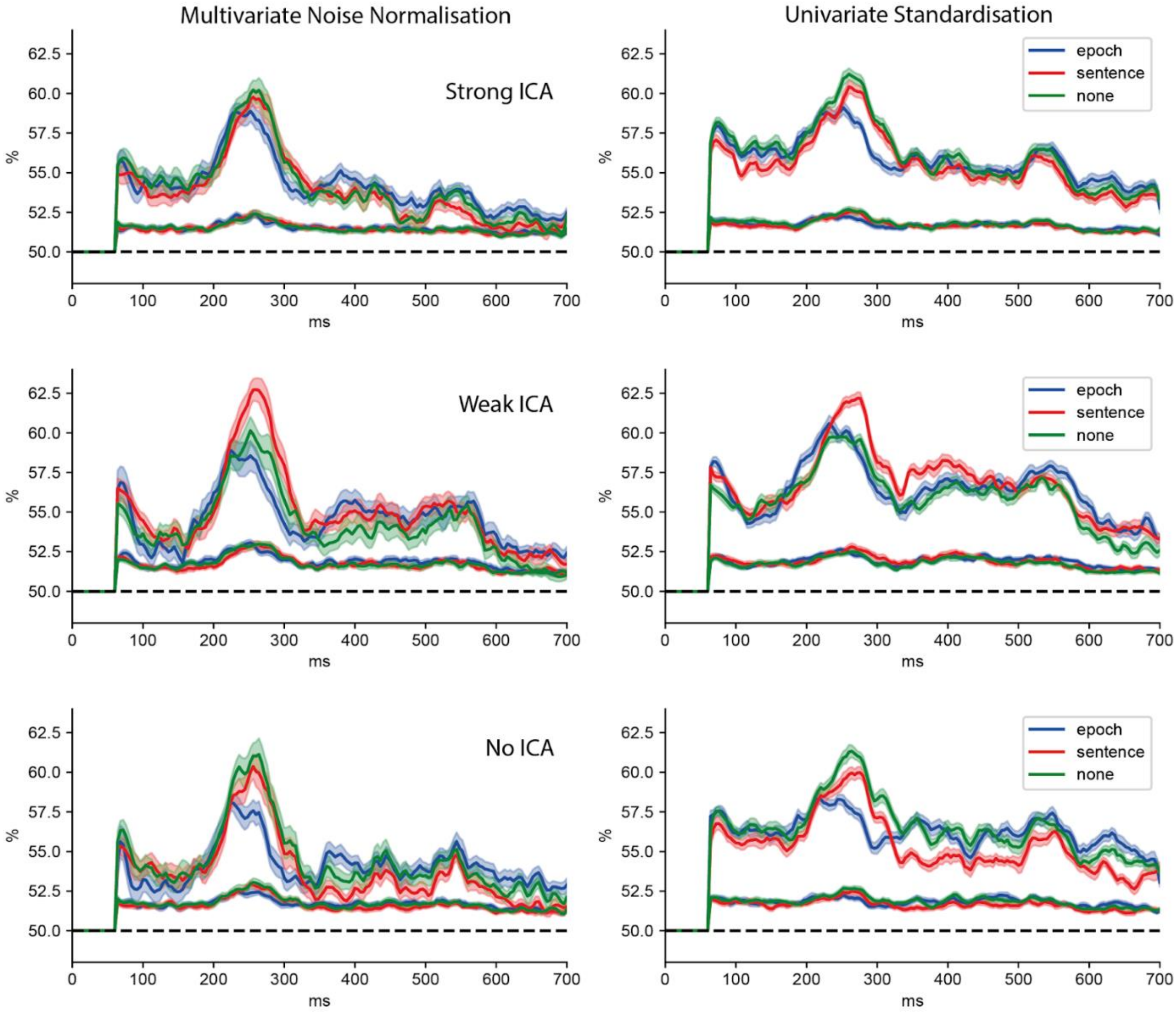

Figure 3.5. Development set sliding window decoding traces for word class (open vs closed-class) decoding. The top three traces of each sub-figure refer to 10-averaged pseudotrials and the bottom three refer to single trials. Each figure modulates the baseline correction method over a specified setting of ICA correction (rows) and feature scaling (columns), i.e. the first column contains 3 plots which have all been scaled via MNN, and the second column, univariate scaling

In all combinations of scaling, baseline correction and ICA-correction type, the response peak is centred around 250 ms. The decoding accuracy immediately after this peak then drops away, more rapidly in the case of multivariate noise normalisation, while the response is sustained using univariate standardisation of the data. This could arise due to overfitting since early stopping during the training of the linear SVM selected the models that performed best on the development set. If multivariate noise normalisation is more sensitive to noisy channels and can correct them, this could indicate that perhaps scaling noisy channels in the data does lead to such overfitting. The period between 350-500 ms appears to indicate that epoch baseline correction results in the best decoding of the binary word class distinction after the main peak. With weakly ICA-corrected data, it appears that sentence baseline correction performs better than other forms of baseline correction.

3.2.3.3.1 - Statistical Analysis

The plots for temporal decoding of word class on the development set suggest the following:

1. From 200-300 ms, there appears to be an interaction between ICA correction type and baseline correction type, specifically driven by weak ICA correction and sentence baseline correction. Epoch baseline correction during this peak appears to be lower in all other combinations except for weak ICA correction with univariate standardisation as the feature scaling procedure.
2. The 300-500 ms varies quite a lot among different combinations of scaling, baseline and ICA settings, but most prominently there appears to be an effect of scaling that occurs over all combinations of ICA-correction and baseline correction

The corresponding decoding traces on the test set are given in Appendix E.

3.2.3.3.2 - Pseudotrials

<u>200-300 ms</u>

The only significant effects in this time period were the main effects of scaling and baseline correction: scaling ($F(1,15) = 28.25$, $p < 0.05$, $\eta_p^2 = .0014$), baseline ($F(2,15) = 9.50$, $p < 0.05$, $\eta_p^2 = .00094$). Post-hoc tests of the three levels of baseline correction did not determine any statistically significant pairwise contrast. These results do not support any generalisation of the observed development set hypothesis referenced above in (1).

<u>300-500 ms</u>

A significant interaction of ICA correction and baseline correction was found in this time window ($F(4,15) = 9.67$, $p < 0.05$, $\eta_p^2 = .0002$), with a significant main effect of scaling, too ($F(1,15) = 85.46$, $p < 0.05$, $\eta_p^2 = .0009$). Post-hoc tests on the significant interaction revealed two pairwise contrasts supporting the effect: strong ICA with sentence baseline correction with (i) no ICA and epoch baseline correction and (ii) weak ICA and epoch baseline correction. The significant main effect of scaling supports (2) from the development set.

3.2.3.3.3 - Single trials

<u>200-300 ms</u>

A significant interaction between ICA-correction and baseline correction exists in this time window ($F(4,15) = 8.05$, $p < 0.05$, $\eta_p^2 = .000017$). Post-hoc tests determined that the significant contrasts principally concerned strong ICA correction with sentence baseline correction. Furthermore, the main effect of scaling was found to be significant in this window ($F(1,15) = 17.78$, $p < 0.05$, $\eta_p^2 = .000095$). Similarly to the pseudotrial case, these results do not support the hypotheses (1) referenced above.

300-500 ms

The same effects reported above for the pseudotrial results were also present for the single trial results in this time window, namely a significant interaction between ICA correction and baseline correction ($F(4,15) = 7.39$, $p < 0.05$, $\eta_p^2 = .000082$), with a significant main effect of scaling ($F(1,15) = 139.21$, $p < 0.05$, $\eta_p^2 = .000039$). Post-hoc tests on the pairwise contrasts revealed only a single significant contrast, that between strong ICA with sentence baseline correction and no ICA correction with epoch baseline correction. This was also a significant pairwise contrast in the pseudotrial case. The significant main effect of scaling supports (2) from the observations on the development set.

3.2.4 – Topographic analysis

In order to gain further insight into the information content carried by the preprocessed EEG data under various preprocessing choices, this section will take a statistically significant finding from the sections relating to word length, frequency and class mentioned in the previous section, visualise the effects on the preprocessed EEG data for the relevant difference wave over a scalp topography. Each of the aforementioned linguistic features were previously divided into binary partitions (see Section 3.2.2.4). Each contrast therefore represents the average signal of one partition after subtracting the other, which if displayed as a topographic map, reveals the systematic differences among the processed EEG data and therefore allows for hypotheses to be made on how a linear classifier can take advantage of differences in order to infer a correct prediction within a given window. 2,500 samples per level of each binary contrast (length: short vs long, frequency: high vs low, class: open vs closed) was randomly sampled across the entire dataset and then averaged together to form a representative ERP. This number of samples was chosen in order to provide a reasonable generalisation across the entire dataset, without the danger of averaging out too many subtle distinctions that are not present in every sample. The EEG data used in decoding is often noisy and could contain systematic differences that are not visible if an average over too many samples is taken. A smaller number of samples might not be representative across the entire dataset, so approximately 5% of the data used in Section 2 (2,500 samples) was used to create representative difference waves, which are mapped to electrode locations in order to visualise the respective topographic maps.

3.2.4.1 – Word Length

In the decoding window of 450–500 ms, two main effects were found in the test data (see Section 3.2.3.1.2), that were hypothesised to be present based on observations of the same temporal window in the development data results. These main effects were of ICA correction type and baseline correction type. Figure 3.6 shows the topographic maps relating to each combination of ICA correction and baseline correction type.

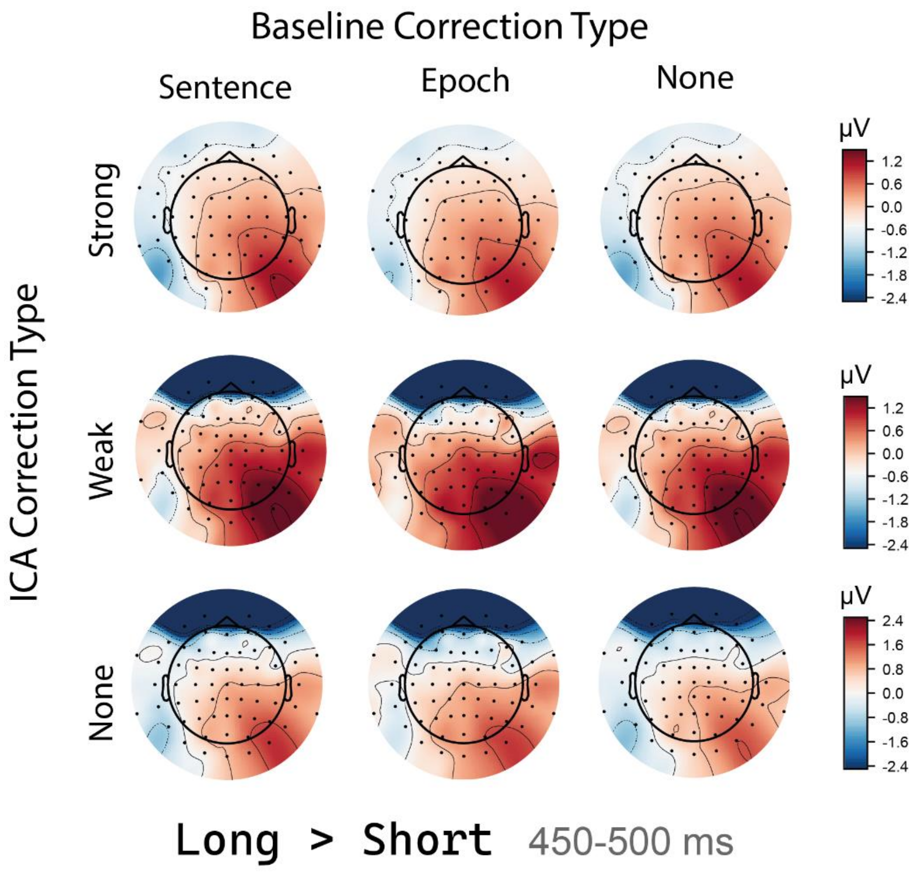

Figure 3.6. Topographies of the difference wave (Long > Short) over the combination of possible settings of ICA correction and baseline correction

Figure 3.6 demonstrates the lack of an interaction in between ICA and baseline preprocessing strategies, while highlighting the prominent differences between each type in the form of the main effect observed in the previous section. The lack of a complete removal of eye movement information (in the cases of no ICA or weak ICA correction) reveal that there is a clear negative frontal polarity that exists between the classes, likely influenced by high-magnitude early visual processing of word length features that take time to recover, over the frontal electrodes. A strong version of ICA correction removes this feature of the topographic maps. The effect of baseline correction in this window is less easy to discern visually in these topographic maps, but given the presence of visual responses to these features, which are typically between 80-150 ms, this (450-500 ms) is a relatively late window that likely is affected by earlier word length effects that spill over into later time periods and interact in complex ways with the subsequent stimuli and / or with other confounds that are not directly modelled here. The effect of eye movements more generally, however, are expected to be different between long versus short words as longer words require more processing time and are fixated more extensively. Long words also typically require saccadic movements to fixate multiple points in a word, while shorter words are processed with more ease. These effects can lead to large discrepancies between these two classes with regard to strong activity at the frontal electrodes, which is demonstrated above in the cases where not all eye movement components were removed via ICA.

3.2.4.2 – Word Frequency

The classifier sliding-window accuracy trace in the previous section, when modelling high-frequency versus low-frequency words, suggested that the window between 280-320 ms contained a region which was differentially affected by both ICA correction type and baseline correction type, which was verified by statistical testing. Post-hoc tests revealed that the interaction between ICA and baseline correction type were responsible for driving the effect. This window is also where the typical decoding peak of frequency information is seen, so a further investigation into what electrophysiological features are helpful in decoding frequency in this window is very useful, as frequency information is a strong signal that can be decoded and provides a lot of information that correlates with other linguistic features of words, such as part-of-speech. Figure 3.7 shows the topography of the 9 conditions in the average of this post-stimulus window (280-320 ms), namely, the topography of the difference wave by subtracting the ERP of high-frequency trials from low-frequency trials.

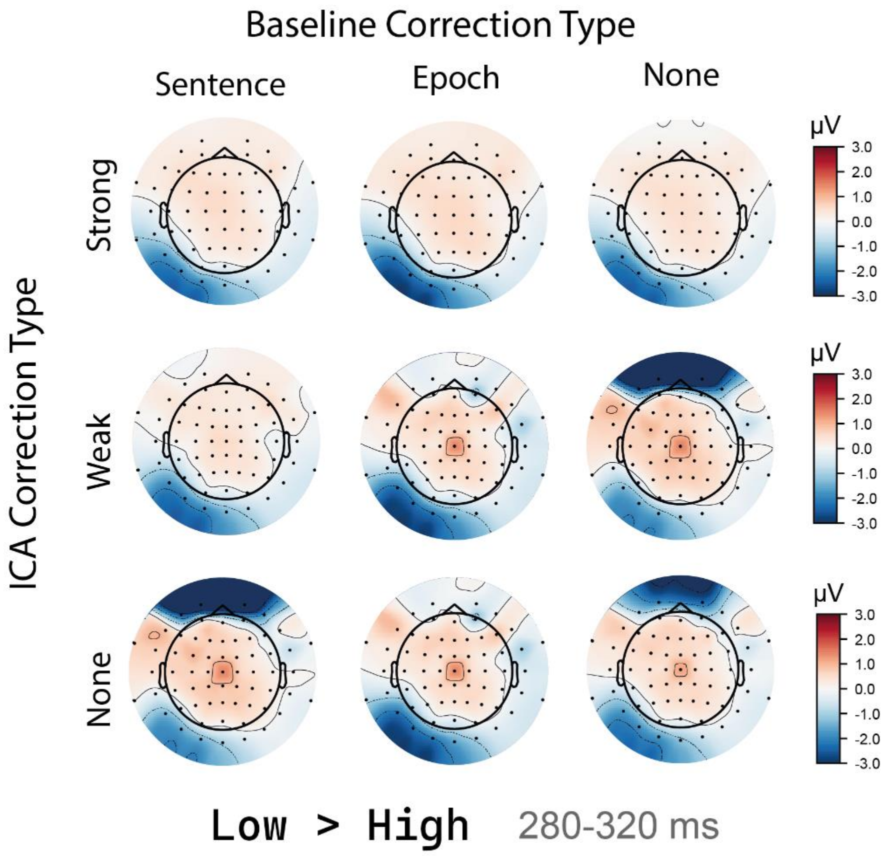

Figure 3.7. Topographies of the difference wave (Low > High) over the combination of possible settings of ICA correction and baseline correction

The topographical distribution of the difference wave intuitively supports the effect of a significant interaction between these groups. An interesting pattern emerges relating to the effects of frontal eye movements. If strong ICA is used to correct the data, then the effect of baseline correction does not have a profound effect on the topographic difference wave (top row). However, in the case of weak ICA correction, it appears this preserves a pronounced difference between high and low frequency words in the frontal electrodes – unless corrected with a form of baseline correction. If no baseline correction is applied on weakly ICA corrected data (middle row, column three) then frontal negativity is preserved, which is removed when either sentence-level (first column) or epoch-level (second column) is applied to the EEG data during preprocessing. This interaction is however, not sufficient to account for the other topographies. If no ICA correction of eye movements is applied, then sentence-level baseline correction preserves a strong frontal negativity between the two classes, yet epoch-level baseline correction, closer in temporal proximity, is effective at removing this effect, which classifiers can use in order to boost decoding performance. The topographies in the third column of Figure 3.7 support the earlier observation in the decoding and statistical analysis (Section 3.2.3.2.1) that not applying baseline correction results in an increased capacity of a classifier to learn the difference between long and short words. As the `Long > Short` difference topography shows, this is supported by a large discrepancy in the frontal electrode polarity between the classes.

3.2.4.3 – Word Class

The temporal window between 300-500 ms for the closed versus open word levels of the word class distinction (Section 3.2.3.3.1) was observed (and statistically confirmed) to be differentially affected by the choices of ICA and baseline correction type, as well as more broadly by the type of normalisation. The latter point relates more to the processing of the data during the learning stage of the classifier and is less interpretable via ERPs or scalp topographies, but an examination of baseline correction and ICA correction types will likely shed further light on the effect these preprocessing steps have on decoding open versus closed class trials. The results are shown in Figure 3.8.

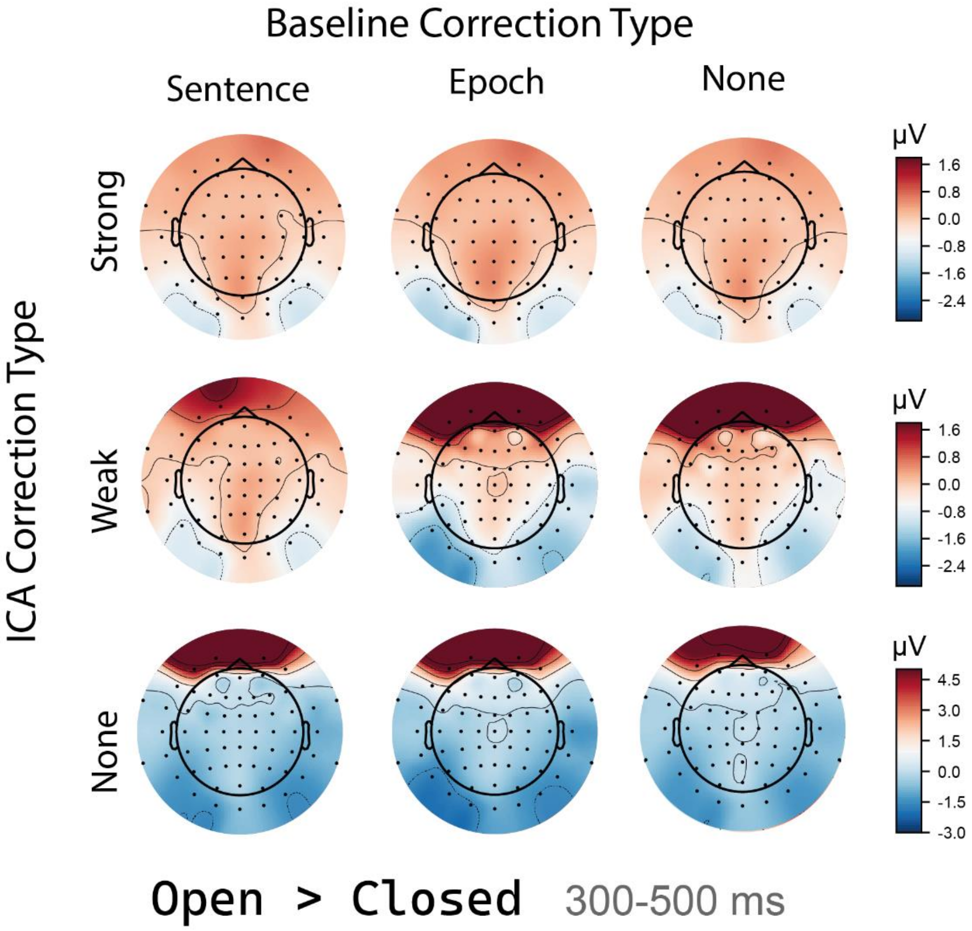

Figure 3.8. Topographies of the difference wave (Open > Closed class) over the combination of possible settings of ICA correction and baseline correction

In comparison to the reported topography in Figure 2.6(B), where confound-correction had been applied (see Section 2.1.6 for details), the topographies here are from samples across the dataset and therefore are not balanced with respect to word length and frequency. Therefore, it's likely the confounding influence of these other factors, accounted for over the previous two sections, likely play a role here. The status of preprocessing with regards to the word class (open versus closed) distinction is relevant, however, for cases where it is not possible to carefully balance confounding dimensions, as is usually the case in naturalistic and likely future applications of such technology.

As can be seen in Figure 3.8 above, when applying strong ICA correction, baseline correction has no real effect, as sentence-level and epoch-level correction result in similar topographies as the case when none are applied. Conversely, when no ICA correction is applied, the principle difference between open and closed class words is centred around the frontal electrodes due to the electrophysiological effects of eye movements, whose ERP traces are magnitudes larger than the electrode traces from elsewhere across the head. This is not an ideal solution, because the data used for decoding is then primarily driven by the effects of confounding eye movements. When only weakly-correcting the EEG data during preprocessing with ICA, the topographies between the two classes are stronger across a broader range, revealing a visually-recognisable difference according to baseline correction method. The topographies for the cases where epoch-level and no baseline correction are applied look relatively similar, showing a dominant frontal distinction. However, when baseline correction happens at the sentence level, the magnitude of the positive frontal activity resembles more the case when strong ICA is applied, differing only by a more subtle positivity in frontal electrodes. This appears to be a stable medium, that would allow decoding to take advantages of the confounding eye movements without them dominating the rest of the signal, with relatively broad positivity down the midline of the head. The temporal decoding results reported in Figure 3.5 for weak ICA (middle row) show that it is exactly this combination of preprocessing, across both types of normalisation, that results in a higher peak value that is visually distinct from the other methods. An analysis of the resulting topographies between the classes, i.e. difference waves, allows introspection that can lead to good hypotheses to understand the nature of such large differences in decoding performance.

3.2.5 - Discussion

For the binary decoding of word length and frequency, observations on the development set with regard to preprocessing decisions were borne out when tested on completely independent test data, showing that these decisions likely do have a true effect which can be used to inform decisions on data preprocessing if the end goal or principal tasks would benefit from decoding those features. It was more difficult to find robust effects for the generally weaker responses to word class. However, it did appear that there was a positive effect for feature scaling type, in which univariate scaling was significantly higher for decoding in the 300-500 ms post-stimulus period. A strong and reproducible response when discriminating between low and high frequency words was found for the condition of not applying baseline correction, irrespective of ICA correction type or feature scaling. For word length, the variable results showed a lot of significant interactions which requires more nuanced investigation in order to understand the interplay between these factors.

A few limitations are important to consider for these results. Some significant main effects did not reveal any significant pairwise contrasts during post-hoc testing, but Tukey's HSD's method of controlling multiple comparisons can bring about nonsignificant effects if many contrasts are performed or if the results provide relatively weak evidence against the null hypothesis. Furthermore, visualising single trials on the same scale as highly-averaged pseudotrials hides a lot of the differences that were found in the statistical analysis and potentially underrepresents the effect size within the context of single-trial decoding effects due to the scale required of the y-axis to accommodate the increased decoding accuracy from data with a higher signal-to-noise ratio. In order to understand the different decoding traces and the impact each preprocessing step has on those results, it is sensible to look at the effects on the preprocessed data itself that enters the model.

By examining, for each of word length, frequency and class, a topographic map of a difference wave that was originally hypothesised and then statistically verified, one is afforded the opportunity to better understand the mechanisms by which a classifier can learn distinctions between a set of classes. This is the goal of the previous section, which provided support for the observations both in the statistical analysis, i.e. in the form of visually demonstrating the spatial profile of interaction effects, but also supporting the observations in the decoding traces of Sections 3.2.3.1-3.

Strong ICA often results in fairly broad and consistent low-magnitude difference waves between the levels of the classes reported in this section. Frontal electrodes vary considerably in cases where eye movement information is available in the signal, and in some cases, the interaction between ICA correction and baseline correction can be explained by finding a middle ground between removing all high-magnitude frontal eye movement artefacts and allowing the signal to be dominated by these, such that lower-magnitude reflections of linguistic processing are not detectable by a decoding classifier. These results show that it is not always best to remove all ICA activity from eye movements when the goal is that of decoding linguistic information from EEG signals.

3.3 - Visualisation of EEG Preprocessing Effects

3.3.1 - Introduction

The previous section analysed the effects of various preprocessing steps on the temporal decoding of three data sets, in which careful steps had been taken to balance a 2-class problem with respect to confounding variables. In this section, the main focus will be placed on visualising the EEG data that was used in the previous section. This involves both the channel-wise and spatially-distributed grand average potential over the time period of interest, as well as a selection of the topographies as they unfold dynamically. This aids a visual understanding both of (i) what differences in the data could lead to differences in the decoding accuracy, and (ii) the effects that preprocessing steps have (or don’t have) on such data.

3.3.2 - Method

ERP traces and topographies from the training sets of each of the 27 datasets was calculated, i.e. from each of (1) linguistic variable being examined (word length, frequency, class), (2) baseline correction method (epoch-based, sentence-based or none) and (3) ICA correction type (strong, weak, none). This resulted in grand-averaged ERPs over 87,270 samples in the word length data, 51,940 samples in the word frequency data and 41,160 samples in the word class dataset. These results are organised by word length, frequency and class in order to better compare the visual effects of each preprocessing step. This is then used as supporting evidence to add further insight into explaining the set of results on the decoding traces presented in the previous section.

3.3.3 – Results

3.3.3.1 - Word length

Figures 3.9 – 3.11 show the grand averaged ERP over all trials used in the confound-corrected word length dataset (N = 87,270), where each figure contains the epoch-based baseline correction in the top row, sentence-based baseline correction in the middle row and no baseline correction last.

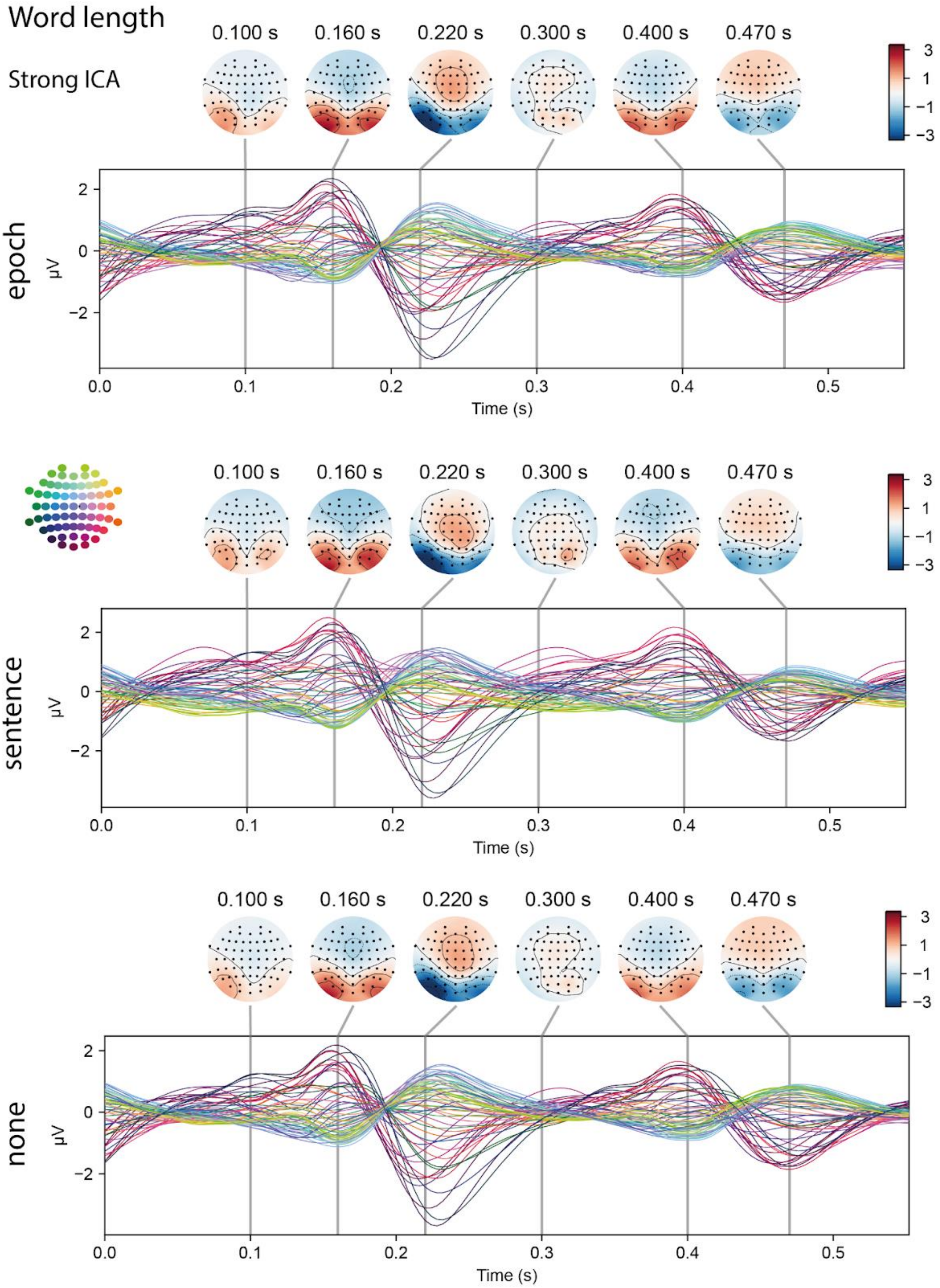

Figure 3.9. Grand average ERPs over all trials in the training portion of the word length dataset (N = 87,270) which had strong ICA correction. Each row visualises a specific baseline setting (top: epoch baseline correction, middle: sentence baseline correction, bottom: no baseline correction)

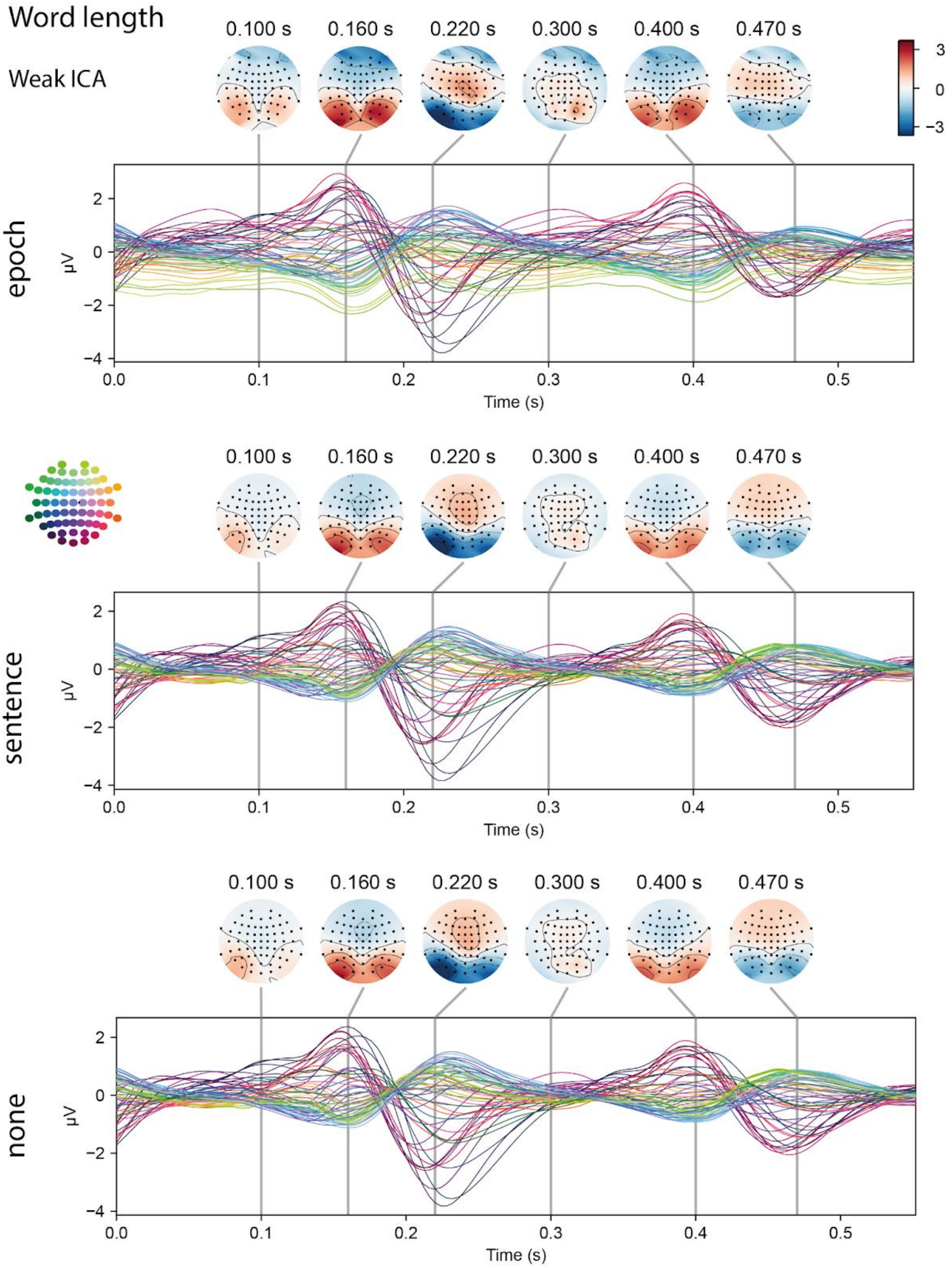

Figure 3.10. Grand average ERPs over all trials in the training portion of the word length dataset (N = 87,270) which had weak ICA correction. Each row visualises a specific baseline setting (top: epoch baseline correction, middle: sentence baseline correction, bottom: no baseline correction)

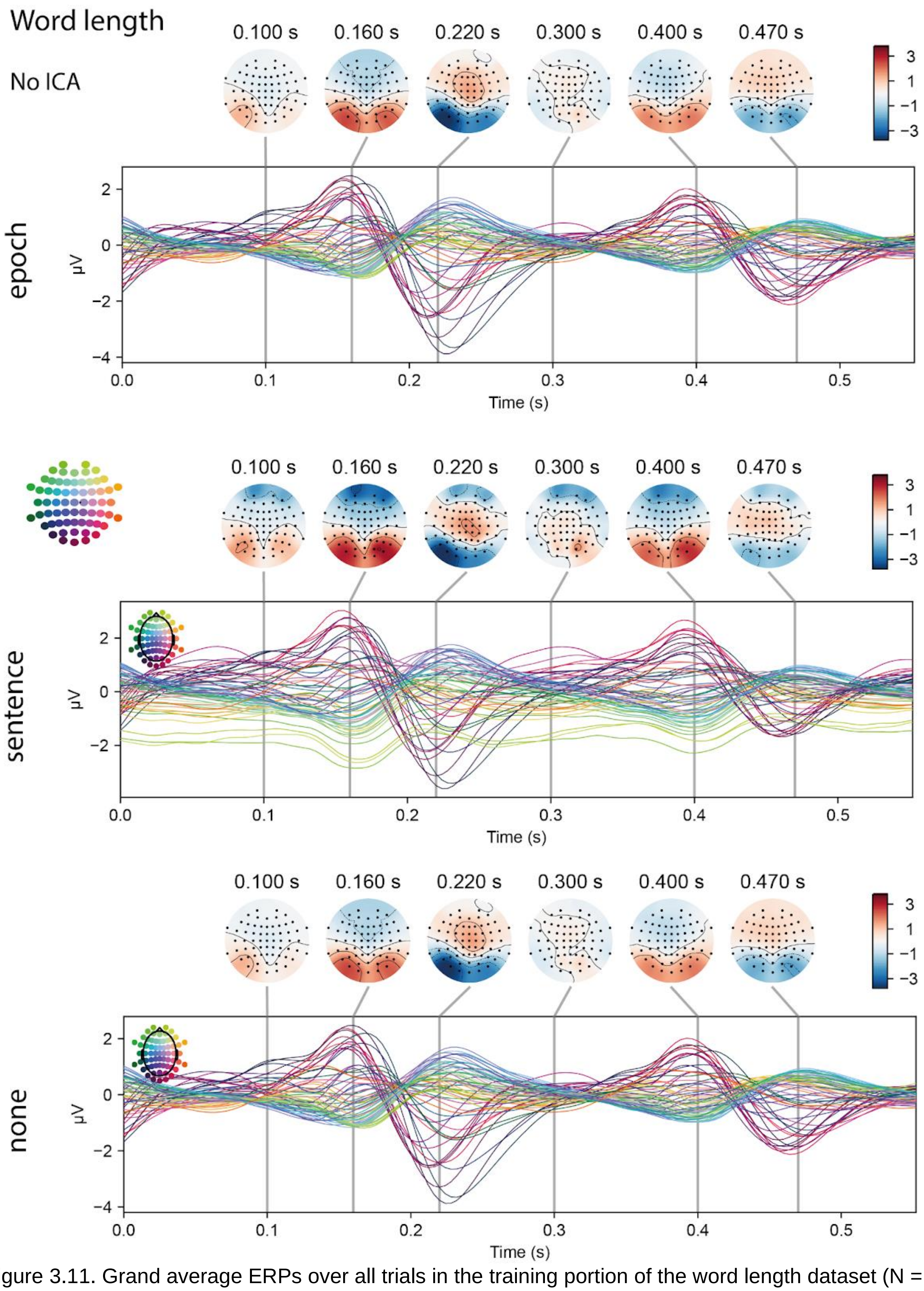

Figure 3.11. Grand average ERPs over all trials in the training portion of the word length dataset (N = 87,270) which had no ICA correction. Each row visualises a specific baseline setting (top: epoch baseline correction, middle: sentence baseline correction, bottom: no baseline correction)

A clear observation is that the grand average ERP where no ICA correction has been applied, with no baseline correction, looks similar to the epoch-based baseline correction. This is undoubtedly because the average over random fluctuations seen in the single-trial data are inconsistent among trials and therefore disappear during the averaging process. In this regard, there is an underappreciation for these figures of how noisy the signals are at the single-trial level and at the pseudotrial level (averages of just 3 / 10 single trials). However, if sentence baseline correction is used, this has a systematic effect which is still present over large-scale averaging, in which there is stronger consistent frontal negativity typical of ocular signals present during reading, i.e. the saccadic eye movements (in left to right reading), in which the negative corneoretinal dipole projects a more negative polarity forwards and is detected by the frontal EEG electrodes. Strong ICA correction results in no discernible ocular effects as the primary purpose is to remove these effects. In weak ICA correction, using epoch baseline correction, a relative frontal negativity is present in the early post-stimulus period.

3.3.3.2 - Word frequency

Figures 3.12 – 3.14 show the grand averaged ERP over all trials used in the confound-corrected word frequency dataset (N = 51,940), where each figure contains the epoch-based baseline correction in the top row, sentence-based baseline correction in the middle row and no baseline correction last.

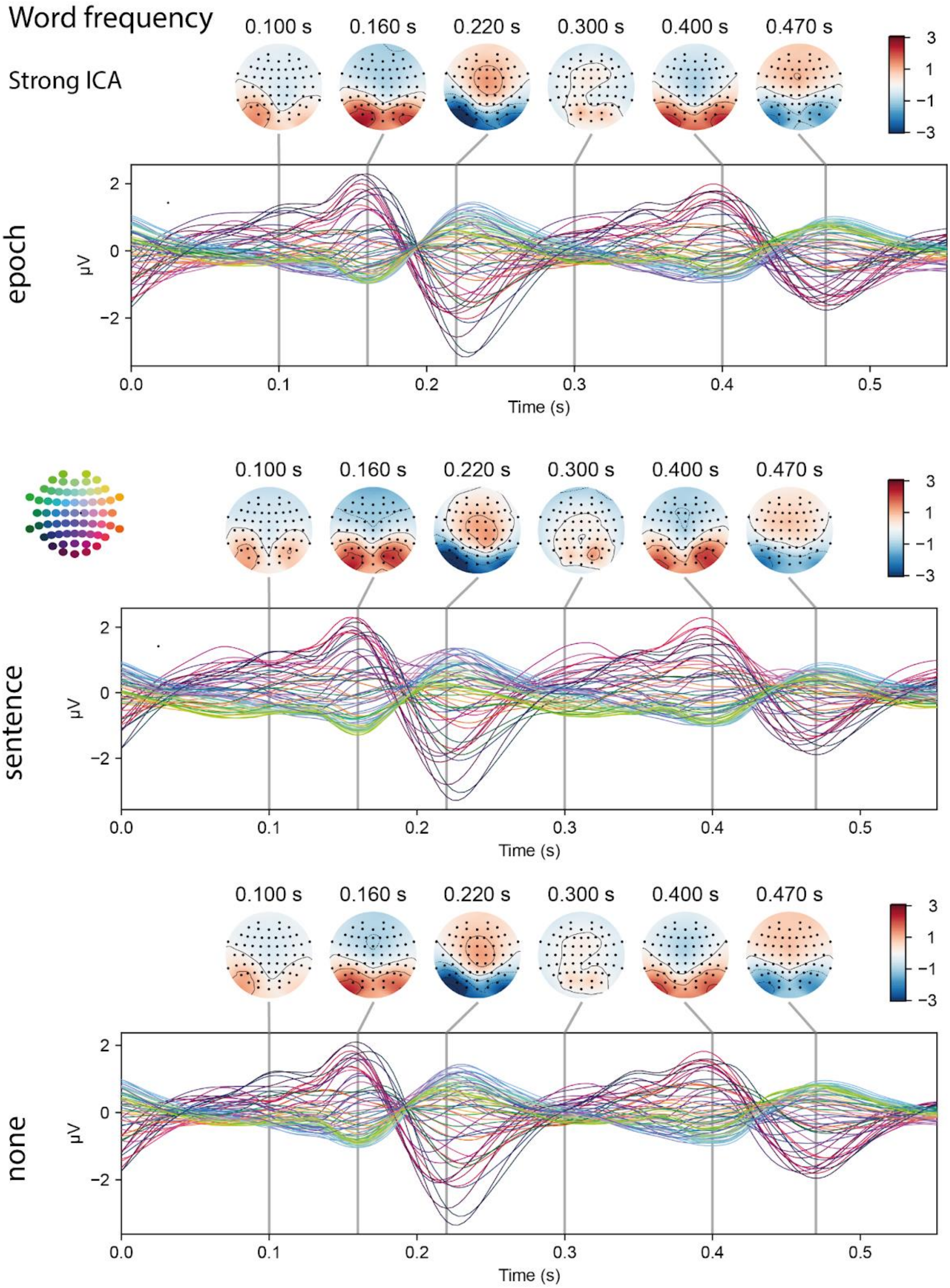

Figure 3.12. Grand average ERPs over all trials in the training portion of the word frequency dataset (N = 51,940) which had strong ICA correction. Each row visualises a specific baseline setting (top: epoch baseline correction, middle: sentence baseline correction, bottom: no baseline correction)

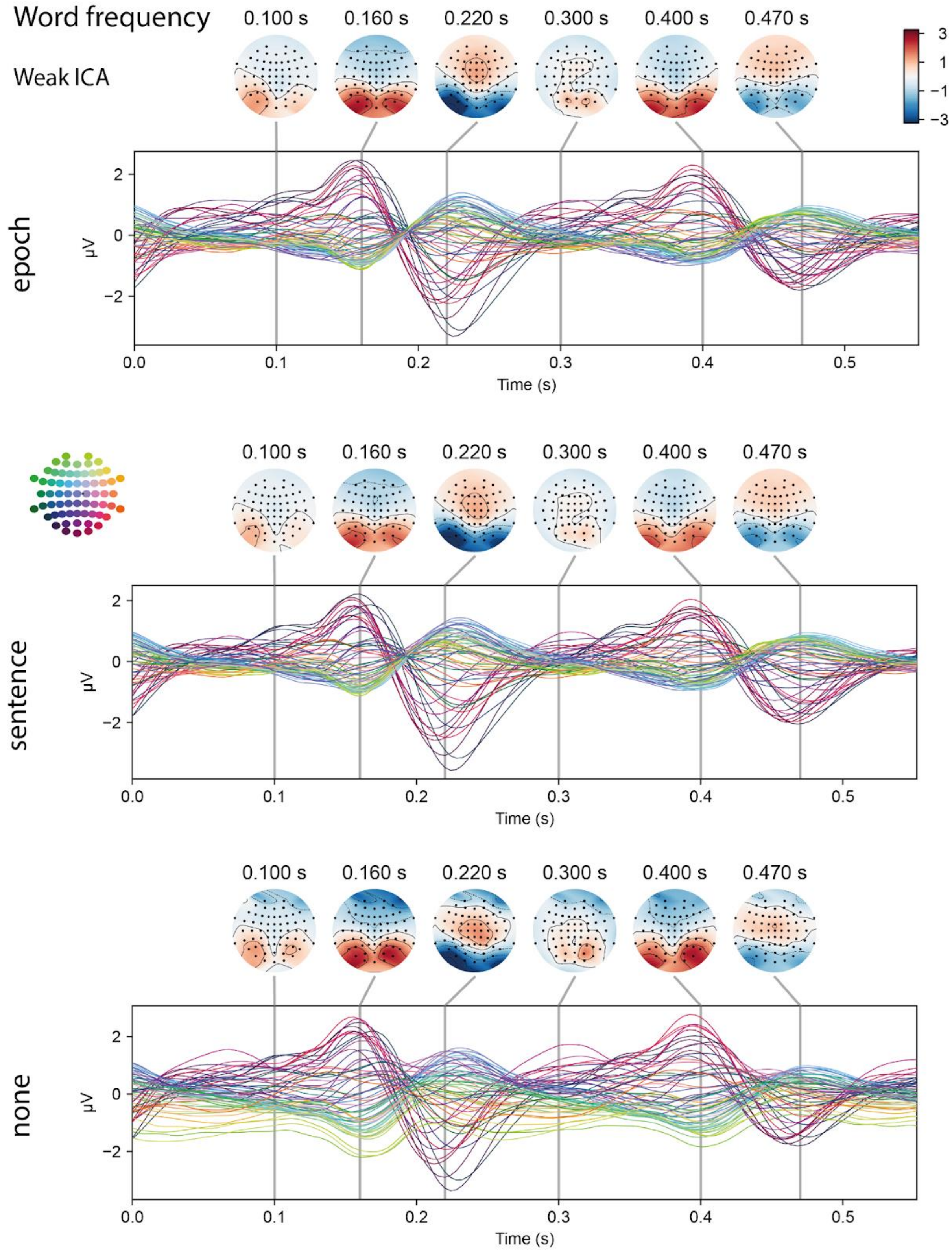

Figure 3.13. Grand average ERPs over all trials in the training portion of the word frequency dataset (N = 51,940) which had weak ICA correction. Each row visualises a specific baseline setting (top: epoch baseline correction, middle: sentence baseline correction, bottom: no baseline correction)

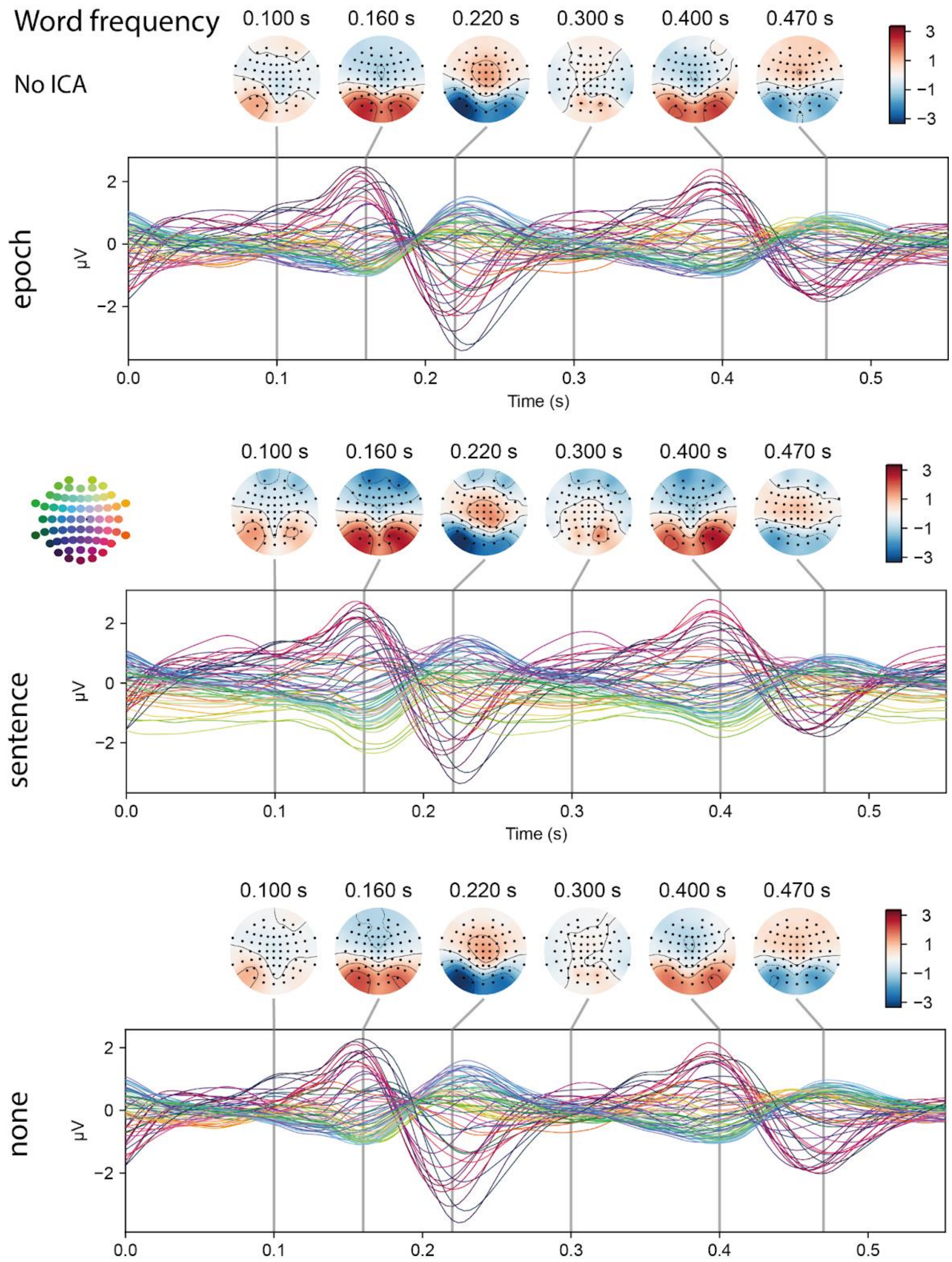

Figure 3.14. Grand average ERPs over all trials in the training portion of the word frequency dataset (N = 51,940) which had no ICA correction. Each row visualises a specific baseline setting (top: epoch baseline correction, middle: sentence baseline correction, bottom: no baseline correction)

The results are similar to word length in that strong ICA correction prevents frontal negativity seen due to eye movements, which manifests itself differentially in the cases where some of those signals remain (weak ICA & no ICA). The same caveats regarding averaging visualisations apply as were mentioned above.

3.3.3.3 - Word class

Figures 3.15 – 3.17 show the grand averaged ERP over all trials used in the confound-corrected word class dataset (N = 46,160), where each figure contains the epoch-based baseline correction in the top row, sentence-based baseline correction in the middle row and no baseline correction last.

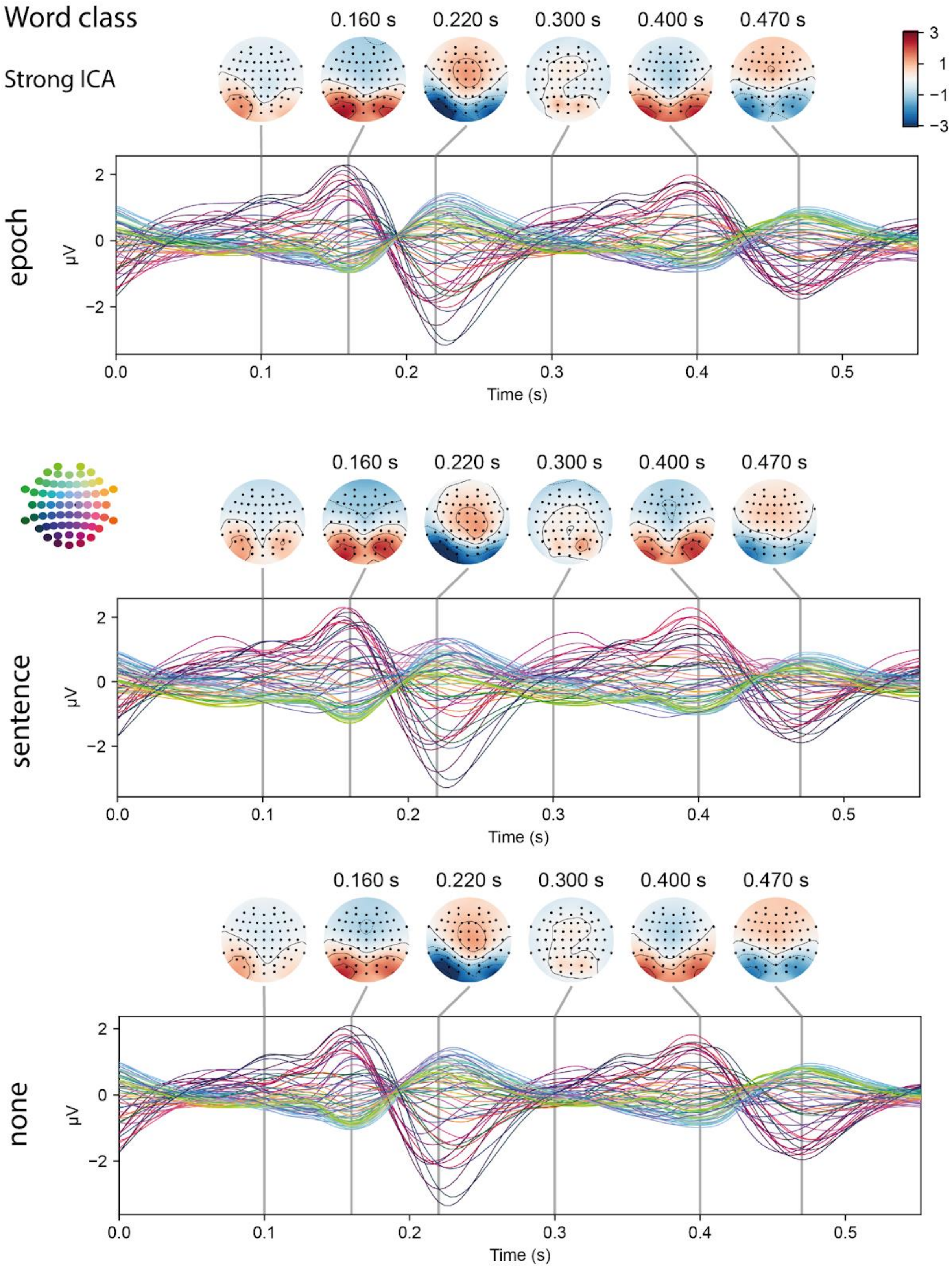

Figure 3.15. Grand average ERPs over all trials in the training portion of the word class dataset (N = 46,160) which had strong ICA correction. Each row visualises a specific baseline setting (top: epoch baseline correction, middle: sentence baseline correction, bottom: no baseline correction)

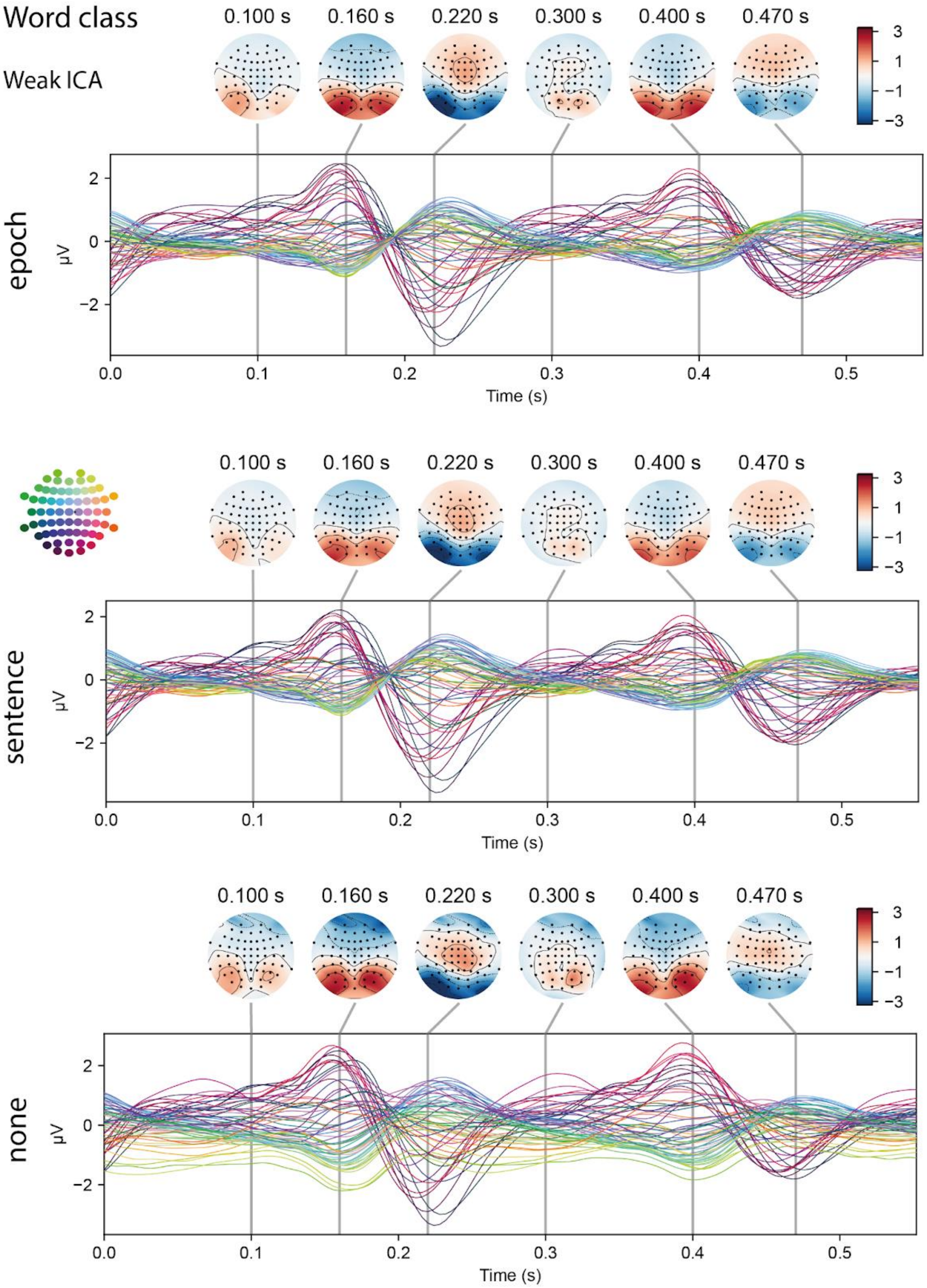

Figure 3.16. Grand average ERPs over all trials in the training portion of the word class dataset (N = 46,160) which had weak ICA correction. Each row visualises a specific baseline setting (top: epoch baseline correction, middle: sentence baseline correction, bottom: no baseline correction)

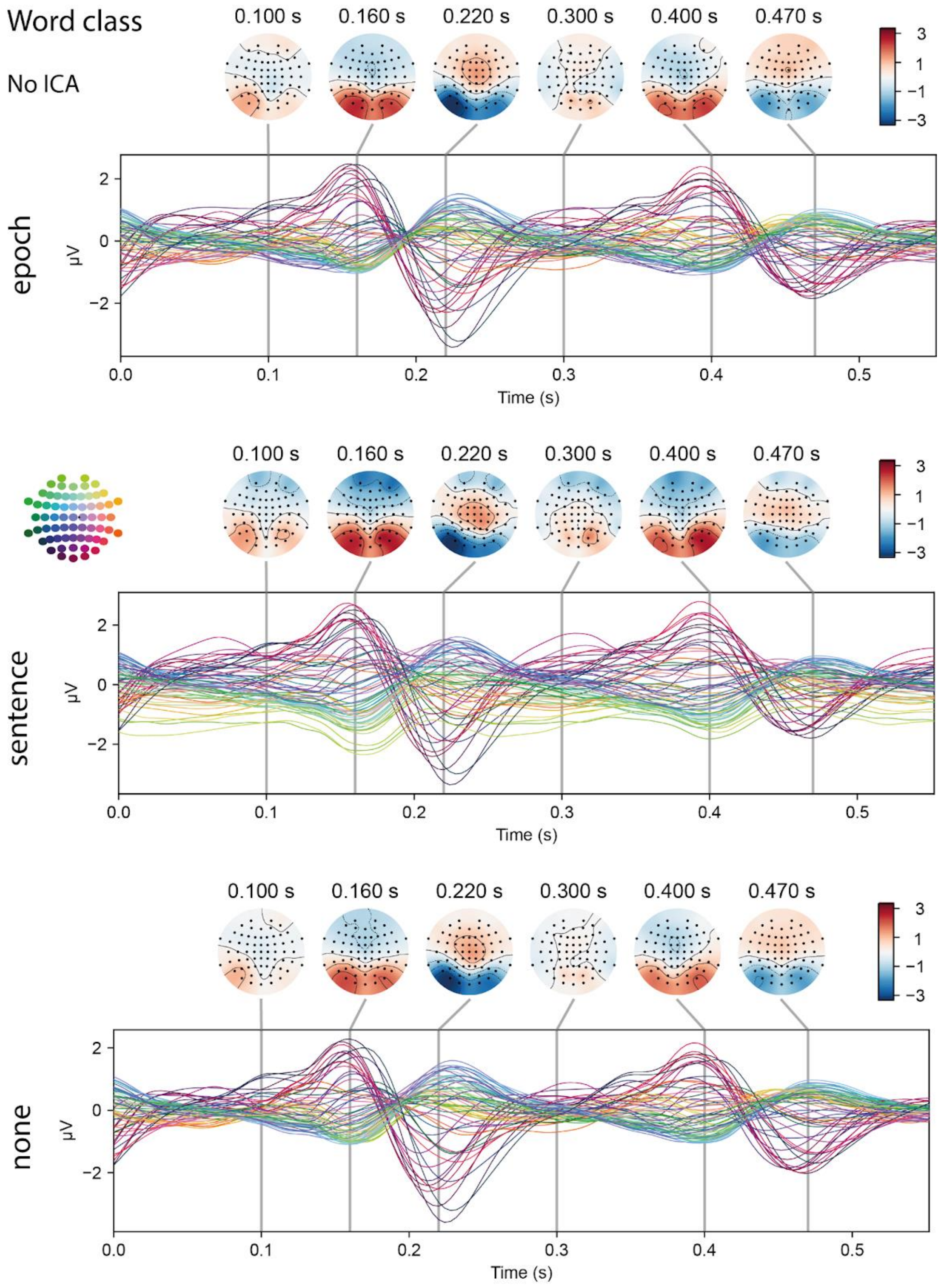

Figure 3.17. Grand average ERPs over all trials in the training portion of the word class dataset (N = 46,160) which had no ICA correction. Each row visualises a specific baseline setting (top: epoch baseline correction, middle: sentence baseline correction, bottom: no baseline correction)

3.3.4 - Discussion

Various effects relating to the preprocessing of EEG data are visible both in the temporal decoding results and also the grand-average ERP effects. These results point to further interactions that could be expanded in future work. These results point to interesting hypotheses that help to further add detail to the emerging picture that when decoding linguistic stimuli, care and attention to the extracted windows of data, feature scaling, type of ICA and baseline correction can help to extract more useful signals used for downstream classification. The inference from decreased decoding performance can also be used to inspect what information is lost in the raw EEG data itself. A critical limitation here is that what can be observed at the macro scale after averaging with many thousands of other data samples. It’s possible that the interaction is more pronounced at a lower-level, but unfortunately one that would be prohibitively large to visually inspect. Broad trends can still be observed between the settings of each different EEG preprocessing pipeline; however it needs to be emphasised that at this scale, similarity of the butterfly plots displaying the ERPs does not imply similarity at the level of the single-trial or few-averaged trials, which is fed to the machine learning classifiers during model training. Overall, the ERPs across word length, frequency and class look remarkably similar, yet lead to very different decoding traces. A limitation of this analysis is that all trials across both classes are used to generate these plots, but separate plots for each condition across both classes could reveal more systematic differences that are lost during the averaging process here.

Chapter 4: General Discussion

In this thesis, I introduce an EEG dataset, which is richly annotated with linguistic features and acquired multiple times by the same subject, which can be used in linguistic experiments and natural language processing tasks many diverse lines of enquiry (such as the stability of the neural response over multiple syntactic contexts) as well as for training models that can decode linguistic information from single-trials. This thesis has focused mainly on the latter, showing that part-of-speech information is present in the EEG response, as demonstrated when accounting for various confounds of word length and frequency. Neural signals can be used directly from recordings to predict up to 6 different part-of-speech classes as well as to decode bigrams, which require complex integration over the time series. I have suggested the self-attention mechanism of the Transformer is a capable tool, capable of modelling such complex interactions towards the goal of linguistic decoding of neural data. Having established this, following work will entail the joint modelling of EEG data with textual inputs to assess the extra contributions that EEG data provide when applied to sequence-based NLP tasks such as part-of-speech tagging or dependency-parsing. Furthermore, a large assessment of the EEG preprocessing pipeline with regard to the ability to decode word length, frequency and open vs closed class was also given in this thesis, under the working hypothesis that eye movements, which are typically removed via ICA correction, are informative of such features and are therefore useful when decoding linguistic information from EEG. The implications, limitations and future research of this work are now considered.

4.1 - Implications

Linguistic status of PoS classes in the brain

I have demonstrated in this thesis that there is a benefit to be conveyed in working with trial-averaged data during the training scheme of both neural network-based and traditional classifiers such as linear support vector machines. Higher signal-to-noise, as measured by a boost in decoding accuracy, has been observed both when using carefully confound-controlled data (section 2.1.2) as well as data averaged from the naturalistic distribution of the text corpus (sections 2.3.2 and 2.3.3). In both of these cases, if responses to linguistic stimuli were primarily driven by confounded responses of word length and frequency, as has been claimed in the literature, then averages over multiple instances of varying word length and frequencies would likely balance out and result in a more noise-like signal that is not associated with a specific part-of-speech class. However, what has been demonstrated is that the signal is boosted in relation to the amount of averaging (trial-averaging from three single trials conveys better performance than single trials; trial-averaging from ten single trials conveys better decoding performance than trial-averages from three single trials). In these averages, across the variable distributions of the potential confounds, the identity of the morphosyntactic status of each part-of-speech remains intact and therefore when trial-averaging is performed over larger samples of single trials, the confounding distributions are expected to become noisy and interact destructively during the averaging procedure. While the claim can’t be made that machine learning classifiers are decoding exactly the morphosyntactic status of a single word, it can be assumed that there is more being detected than just neural responses to confounding variables. This opens up the possibility that EEG-based NLP mechanisms are able to detect linguistic responses on a finer-grained scale than just detecting confounds and have sufficient scope to perform functions such as part-of-speech tagging and potentially parsing mechanisms, too, by identifying the neural responses to concurrent structure building.

Emerging NLP technologies

A major goal of this work was to lay a foundation upon which future research can be based, in terms of verifying what is reasonably decodable from EEG signals using the latest machine learning techniques, which are also used to train state-of-the-art NLP models. This was carefully considered by consistently comparing Transformer results with a linear SVM baseline, in order to dissociate what was linearly decodable in a more traditional fashion, versus what the expected gains were of employing computationally intensive classifiers such as the Transformer model. The key implications of this work with regard to the future of NLP, particularly as a new subfield emerges with cognitive neuroscience, are that linguistic information is decodable from EEG and the type of decodable information is congruent with what state-of-the-art text models extract from their own large corpora of training data. If both input sources are jointly modelled, then this could lead to breakthroughs that advance the field of NLP by incorporating brain signals into their inner workings. Adding EEG information directly to these models without a thorough analysis of the time scales, expected decoding performance and an investigation into what lies behind these effects leads to a situation where any gains are not directly associated with the specifics of EEG, making it unclear what information is being extracted. This thesis aims to specify this upfront, so that the next steps have a solid foundation.

4.2 - Limitations

While it is important to recognise that analysing anything tangentially connected to language is inherently associated with numerous confounds, it is important that care be taken to try to correct for this whenever possible, if one aims to surmise a claim based on observed data. A lot of effort was put into the way I split text data and the neural recordings of these data in order to propose that observed effects, such as the detectability of morphosyntactic responses in EEG signals, whether single-trial or in trial-averaged data, is detectable using machine learning. This also applies when I analysed word length and frequency responses. A key limitation is that it is impossible to completely control for all confounds in an experimentally robust manner. This consideration is important because it leaves open the interpretation that the explained effects that were reported could in theory have been produced by a confound that I failed to properly take into account. A lot of effort was spent into ensuring this was not the case, but this is an important limitation that one cannot escape from. In some sections of this thesis, an engineering (rather than scientific) approach was taken, in which naturalistic confounds of language are welcome as long as they provide information towards helping to predict the correct class. In many sections, the aforementioned limitation does not apply because by demonstrating generalisation onto independent test data, with less of a focus on the causal contributions, this is the end goal.

A related limitation is that we have left open the issue of semantics and aspects of conceptual structure and lexical access which could be causally contributing to some of the observed effects reported in this thesis. A large emphasis was placed on trial-averaging over randomly sampled words such that pseudotrials used in the analyses contain a wide range of different words relating to various semantic concepts, thus blurring any strong semantic signals in the dataset and highlighting commonalities relating to morphosyntactic status. However, since this was not addressed in as systematic a fashion as other known confounds which give rise to strong EEG responses during reading (word length and frequency), the effects of semantic confounding cannot be ruled out.

A further limitation is that our analysis is performed from neural signals from a single subject. Care must be taken towards making generalising statements when a large cohort of multiple subjects isn't present. The ability to recruit multi-subject data in order to expand on some of my earlier work was not achievable, due to global extenuating circumstances. However, given the spatial smoothness of EEG data, I highly suspect that what has been reported here is reproducible in other subjects. The initial portion of the first experimental chapter showed that I managed to replicate ERP responses already reported in the literature (Section 2.1.2) and this hints at the fact that it's reasonable to hypothesise that the reported results reported in this thesis are not specifically unique to the subject used in the acquired EEG dataset reported within this thesis.

4.3 - Future research

The co-dependence of NLP and cognitive neuroscience is mutually beneficial, in that results from either domain can be translated into the other for scientific experimentation, model testing and hypothesis generation. A large component of this thesis has been centred on using EEG signals to set the stage for potential applications to improve NLP systems, but the reverse is also true. NLP methods are equally good testing grounds to develop and simulate neurolinguistic theories, providing the assumed model is suitable to act as a proxy for neural processing. This point is contentious in the wider literature, but some novel research is beginning to show that this is possible (Toneva & Wehbe, 2019). The EEG dataset collected for this thesis contains a viable testing ground to further such ideas. For example, given the rich linguistic annotation, the temporal dynamics and scalp topographies can easily be inspected with respect to the linguistic status of classes of words, categorised according to either linguistic or semantic grounds and used to inform research into the cognitive neuroscience of language.

The successful demonstration of decoding part-of-speech information from single-trial EEG data shows that the combination of this data, along with traditional (state-of-the-art) machine learning methods, such as Transformer-based neural networks, is a promising next step. This research entails training a system that jointly models the textual input features as well as the neural responses, as recorded by EEG, of a human subject reading the same textual inputs. It remains an open question how fine-grained the cardinality of the output classes can be, but successful demonstration of this then allows for experimental research where models of neural data over vocabulary sets can then replace neural signals specific to each textual input. This decoupling of textual input with neural responses is important to allow models to be aided by the inclusion of human neural signals during linguistic processing, without requiring accompanying signals for each textual input. The first steps of such an idea are emerging in the literature for the NLP tasks of relation extraction and named entity recognition. The continuation of the work presented in this thesis will go even further and move to NLP tasks such as part-of-speech tagging and dependency parsing, which contain a highly temporally dynamic feature set.

4.4 - Concluding remarks

The worlds of cognitive neuroscience and NLP have been stepping into each other's territory for many years, but now it appears that the fields are soon to give rise to a joint field of study, in which it is common practice to use neural network models as models of the brain in which hypotheses can be simulated, alongside neural network modelling of neural signals (with or without accompanying input sources such as text, images, audio etc.) This thesis has highlighted the potential for EEG-based NLP systems that can identify part-of-speech alongside showing that other features can be reliably decoded from single trial data. The modelling problem of neural signals is specialised and unlike other common input domains that neural networks are commonly used with, special considerations with regard to the signal-to-noise ratio need to be taken into account. To this end, I have demonstrated that neural networks can implement specialised training procedures to help aid the modellability of neural signals and improve generalisation performance on data not seen during testing.

For the development of potential systems that aid in gold-standard corpus-generation (i.e. fully labelled data which can be useful for model development) for low-resource languages, in which native speakers would read texts and online systems could help to correctly tag each word, a difficult impediment with traditional EEG preprocessing techniques is the computationally costly data decomposition into independent components, identification of noise components and/or ocular 'artefacts', reassembling of the data matrix for use in downstream pipeline stages. I have shown that, depending on the end goal, such steps might not only be unnecessary, but decoding performance might be enhanced by not performing this step, under the working hypothesis that ocular information in the signal is correlated with the type of variable being decoded. These results also apply equally to systems that do not suffer from issues with computational complexity, but also ones that have a high focus on maximum generalisation to new data.

While the results presented in this thesis only scratch the surface of potential new directions which can be taken in the emerging world of integrating human signals with machine learning applications to language, it is my sincere hope that the presented results serve as a useful groundwork for future research in this direction.

Appendix A

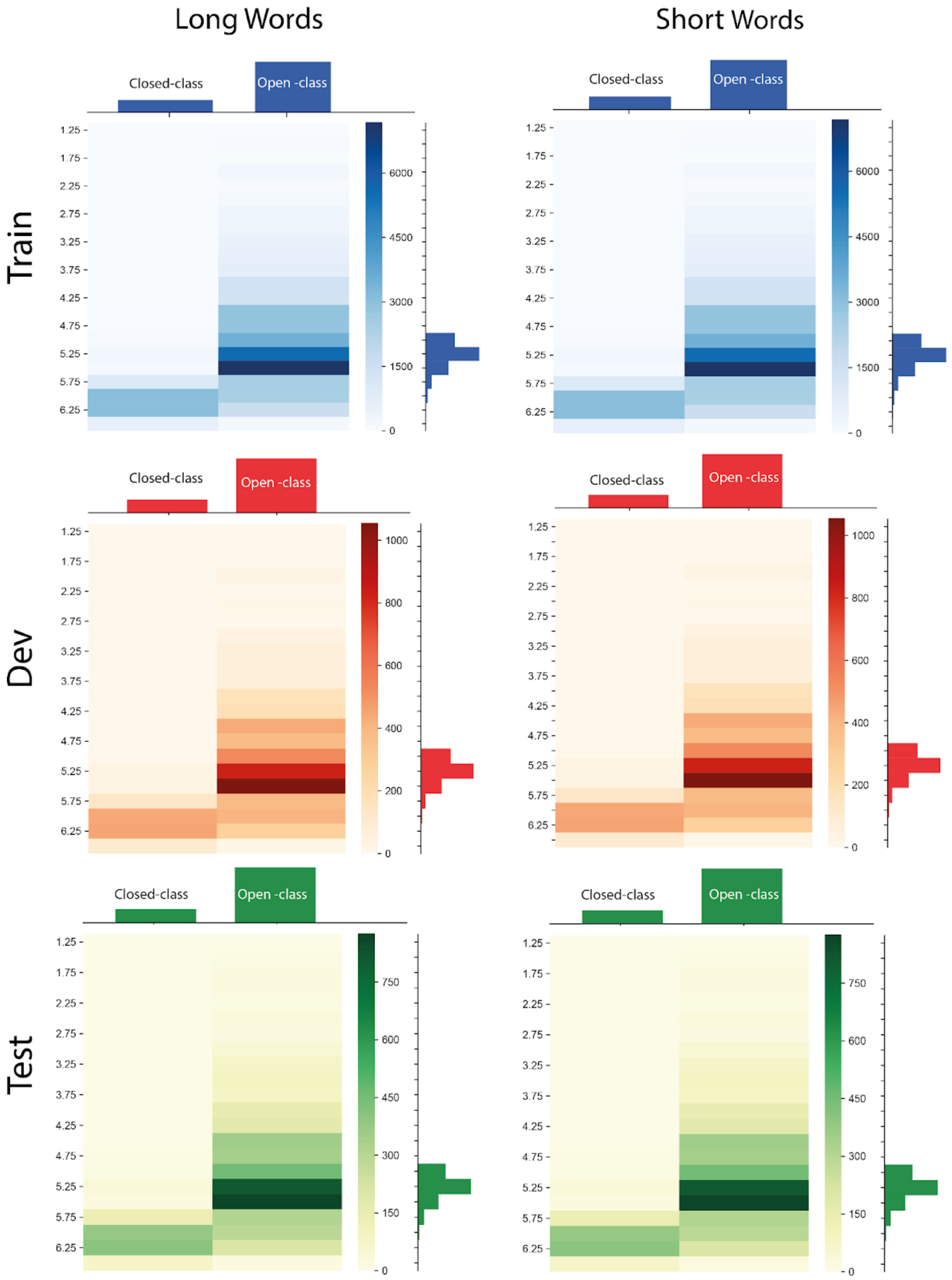

Figure 1. Joint histogram of confounding variables (word class, frequency) with respect to word length as the central variable, i.e. word length is equally balanced across open and closed-class words and word frequency, both equal with respect to the distributions of the confounds. This process is done internally to each data split, which is given by each row (train: blue, development: red, test: green). This process is also applied separately for each of three sentential positions (sentence-start, sentence-middle and sentence-end). Marginal distributions are given along the axes.

Appendix B

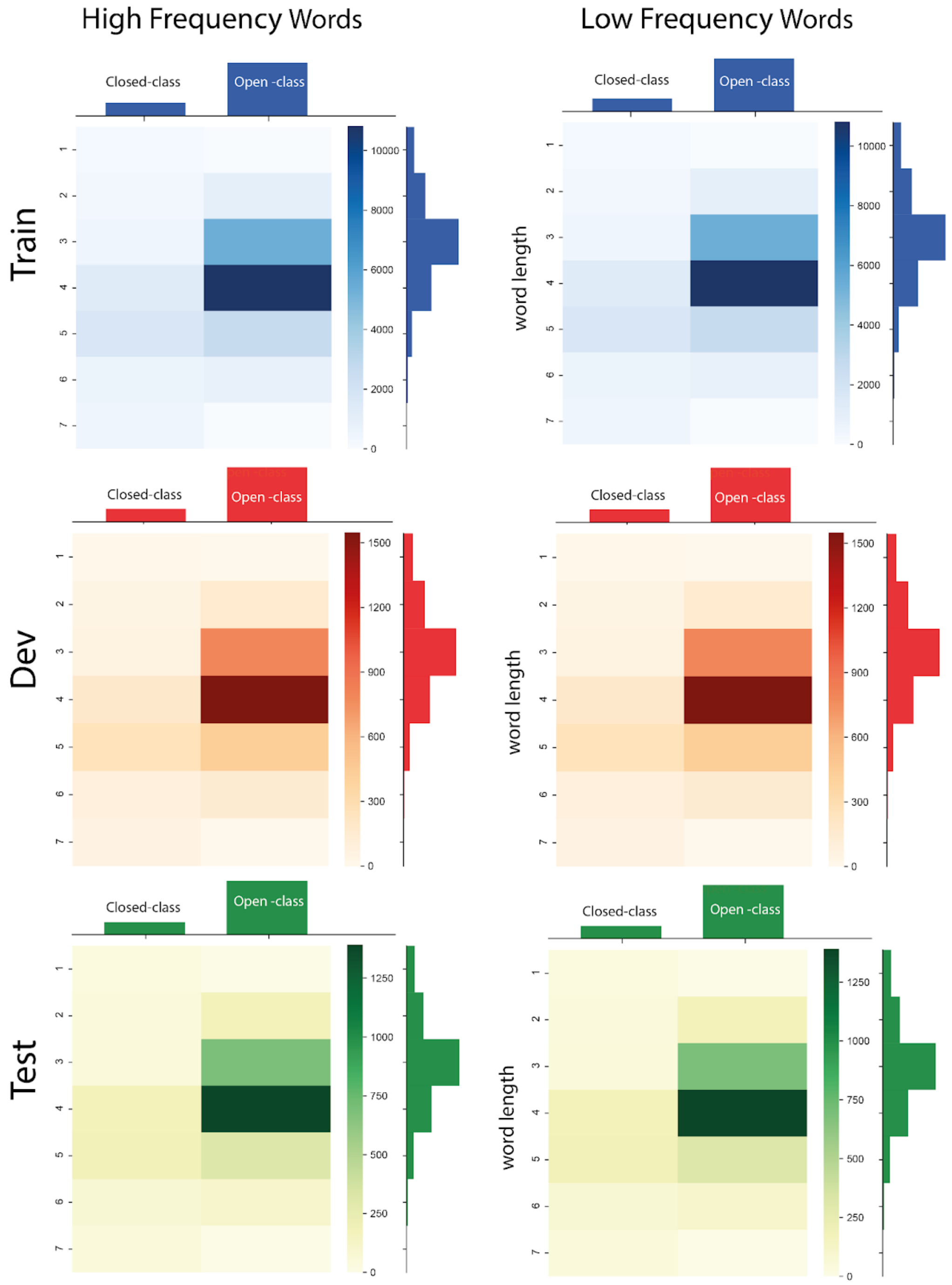

Figure 1. Joint histogram of confounding variables of word length and class with respect to word frequency as the central variable, i.e. word frequency is equally balanced with respect to the distributions of the confounds. This process is done internally to each data split (train: blue, development: red, test: green), which is given by each row. This process is also applied separately for each of three sentential positions (sentence-start, sentence-middle and sentence-end). Marginal distributions are given along the axes.

Appendix C

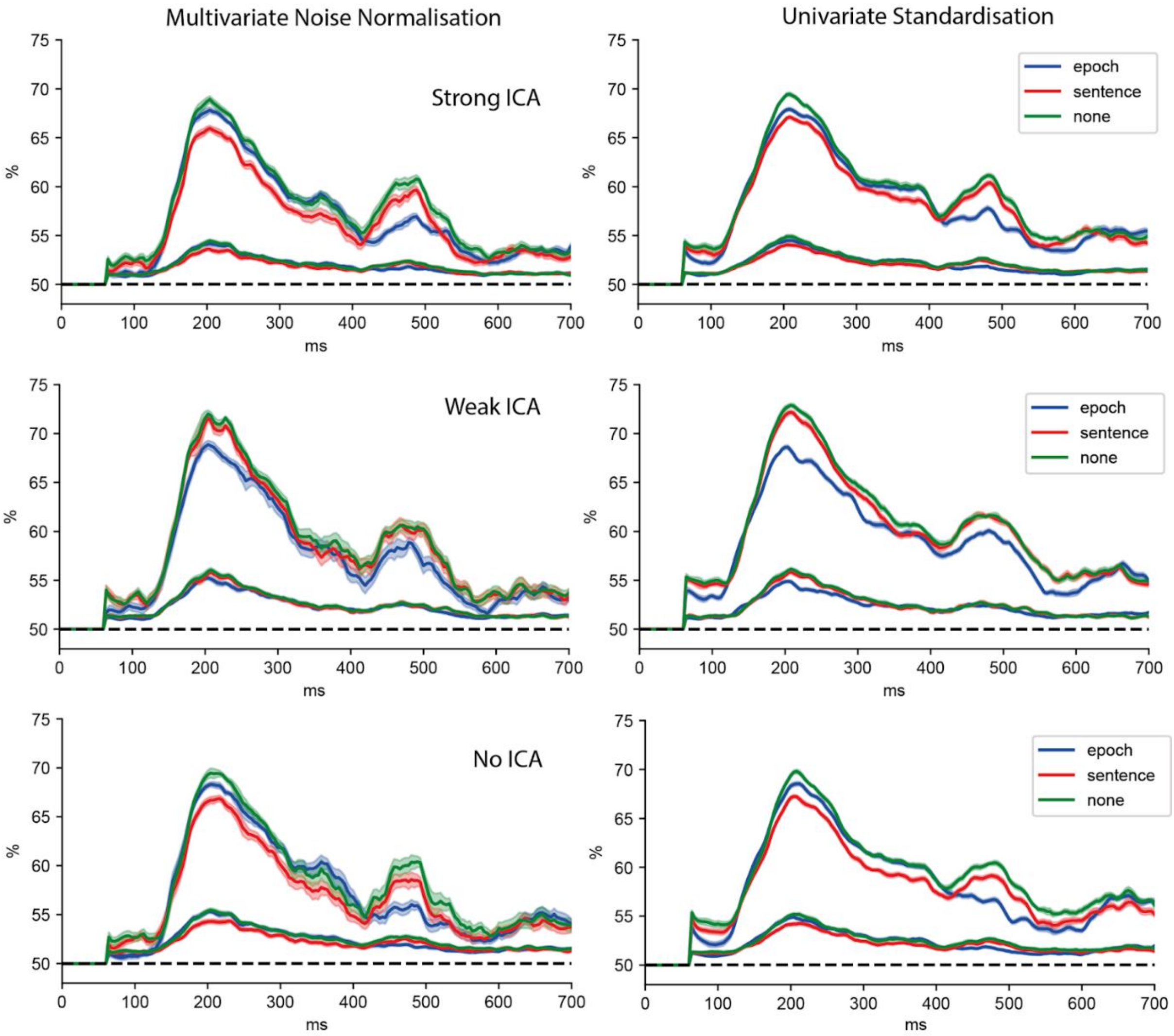

Figure 1. Test set sliding window decoding traces for word length decoding. The top three traces of each sub-figure refer to 10-averaged pseudotrials and the bottom three refer to single trials. Each figure modulates the baseline correction method over a specified setting of ICA correction (rows) and feature scaling (columns)

Appendix D

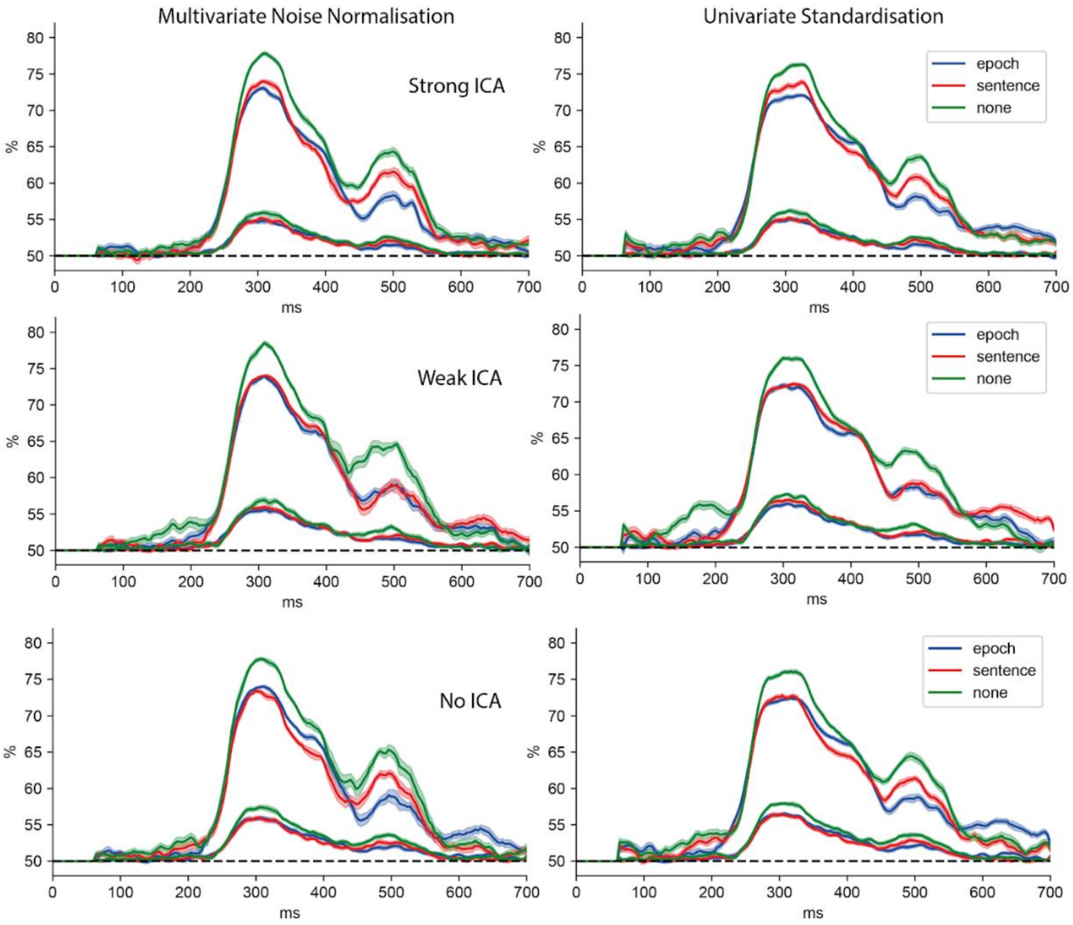

Figure 1. Test set sliding window decoding traces for word frequency decoding. The top three traces of each sub-figure refer to 10-averaged pseudotrials and the bottom three refer to single trials. Each figure modulates the baseline correction method over a specified setting of ICA correction (rows) and feature scaling (columns)

Appendix E

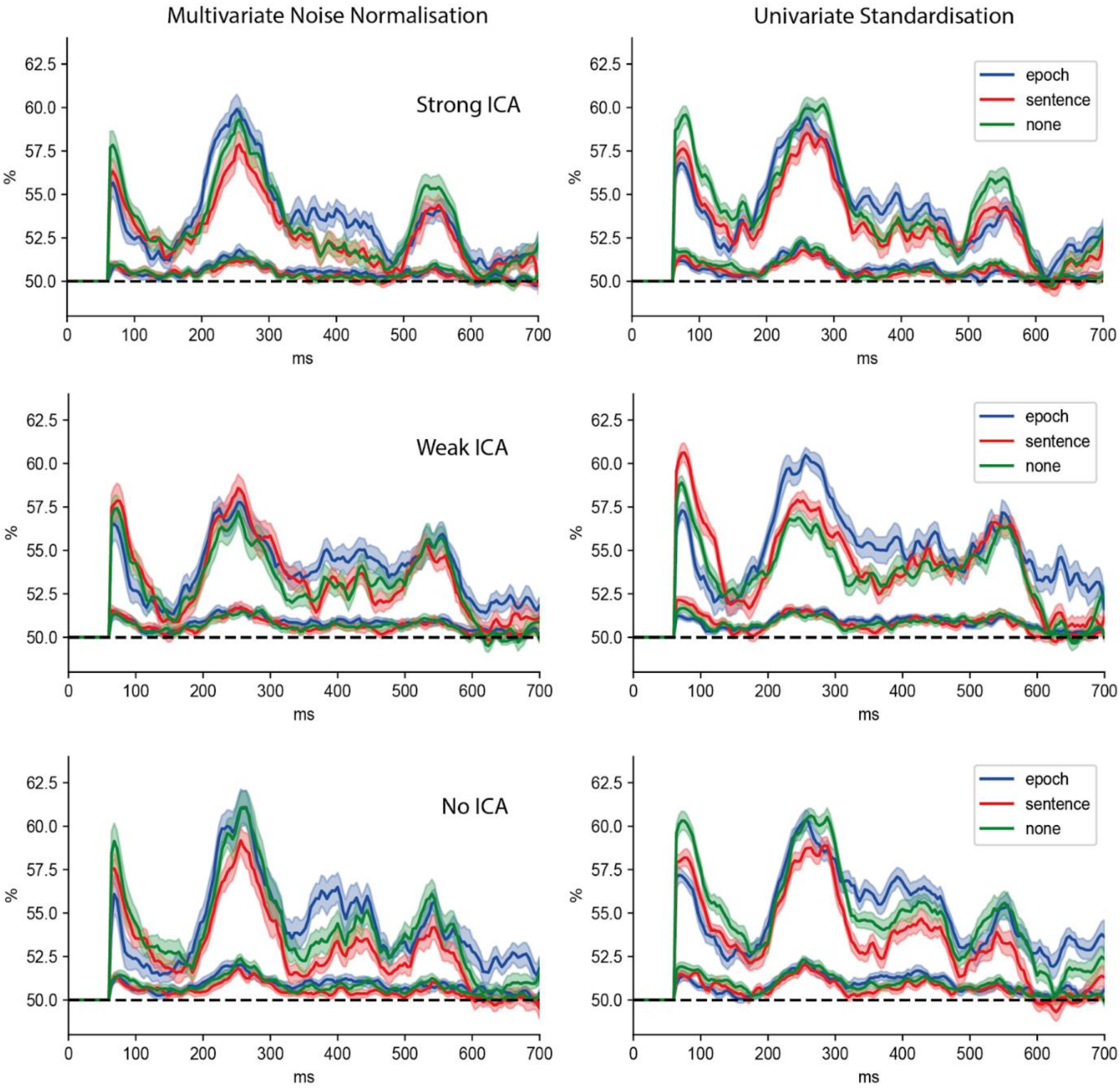

Figure 1. Test set sliding window decoding traces for word class (open vs closed-class) decoding. The top three traces of each sub-figure refer to 10-averaged pseudotrials and the bottom three refer to single trials. Each figure modulates the baseline correction method over a specified setting of ICA correction (rows) and feature scaling (columns)

References

Akmajian, A., Demers, R. A., & Harnish, R. M. (2001). Linguistics, an introduction to language and communication. (5th ed.) Cambridge, Mass: MIT Press.

Alday P. M. (2019). How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting its limits. *Psychophysiology*, *56*(12), e13451. https://doi.org/10.1111/psyp.13451

Amsel, B. D. (2011). Tracking real-time neural activation of conceptual knowledge using single-trial event-related potentials. Neuropsychologia, 49(5), 970-983, DOI:10.1016/j.neuropsychologia.2011.01.003.

Barrett, M., Bingel, J., Keller, F. & Sogaard, A. (2016). Weakly Supervised Part-of-speech Tagging Using Eye-tracking Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association of Computational Linguistics. Berlin: Germany. DOI:10.18653/v1/P16-2094

Bastiaansen, M., Van Der Linden, M., Ter Keurs, M., Dijkstra, T. & Hagoort, P. (2005). Theta Responses Are Involved in Lexical–Semantic Retrieval during Language Processing. J. Cognitive Neuroscience 17, 3, 530–541. DOI:/10.1162/0898929053279469

Bies, A., Mott, J., Warner, C. & Kulick, S. (2012). English web treebank. Linguistic Data Consortium.

Bigdely-Shamlo, N., Mullen, T., Christian, K., Kyung-Min, S. & Robbins, K. A. (2015). The PREP pipeline: standardised preprocessing for large-scale EEG analysis. Frontiers in Neuroinformatics, 9. https://doi.org/10.3389/fninf.2015.00016

Bingel, J., Barrett, M. & Sogaard, A. (2016). Extracting token-level signals of syntactic processing from fMRI with an application to PoS-induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 747-755. Berlin, Germany. Association for Computational Linguistics. DOI:10.18653/v1/P16-1071

Bishop, C.M. (2006) Pattern Recognition and Machine Learning. Springer, Berlin.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., Benoit, R. G., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Boye, K. & Harder, P. (2012). A usage-based theory of grammatical status and grammaticalization. Language, 88(1), 1-44.

Boye, K., & Bastiaanse, R. (2018). Grammatical versus lexical words in theory and aphasia: Integrating linguistics and neurolinguistics. Glossa: A Journal of General Linguistics, 3(1), 1–18. https://doi.org/10.5334/gjgl.436

Brown, C. M., Hagoort, P., & ter Keurs, M. (1999). Electrophysiological signatures of visual lexical processing: open- and closed-class words. *Journal of cognitive neuroscience*, *11*(3), 261–281. DOI:10.1162/089892999563382

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. & others (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165

Burges, C.J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2, 121–167. DOI:10.1023/A:1009715923555

Caramazza, A., Hillis, A. (1991). Lexical organisation of nouns and verbs in the brain. *Nature* 349, 788–790. https://doi.org/10.1038/349788a0

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 1–27.

de Cheveigné, A., & Nelken, I. (2019). Filters: When, Why, and How (Not) to Use Them. Neuron, 102(2), 280–293. https://doi.org/10.1016/j.neuron.2019.02.039

Croft, R. J., & Barry, R. J. (2000). Removal of ocular artefact from the EEG: a review. *Clinical neurophysiology*, *30*(1), 5–19. DOI:10.1016/S0987-7053(00)00055-1

Dambacher, M., Dimigen, O., Braun, M., Wille, K., Jacobs, A. M. & Kliegl, R. (2012). Stimulus onset asynchrony and the timeline of word recognition: Event-related potentials during sentence reading, Neuropsychologia 50(8), 1852-1870. https://doi.org/10.1016/j.neuropsychologia.2012.04.011

Degno, F., Loberg, O., Zang, C., Zhang, M., Donnelly, N., et al. (2019). A co-registration investigation of inter-word spacing and parafoveal preview: Eye movements and fixation-related potentials. PLOS ONE 14(12). DOI:10.1371/journal.pone.0225819

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv, DOI: 1810.04805

Dehaene, S., Naccache, L., Cohen, L., Le Bihan, D., Mangin, J. F. & Poline, J. B. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. Nature Neuroscience, 4, 752–758.

Dehaene, S., Jobert, A., Naccache, L., Ciuciu, P., Poline, J. B. , Le Bihan, D. (2004). Letter binding and invariant recognition of masked words: behavioural and neuroimaging evidence. Psychol Sci 15, 307– 313

Devlin, J., Chang, M-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1), 4171-4186

Dimigen, O. (2020). Optimising the ICA-based removal of ocular EEG artifacts from free viewing experiments. *NeuroImage*, *207*, DOI: 10.1016/j.neuroimage.2019.116117

Dufau, S., Grainger, J., Midgley, K. J., & Holcomb, P. J. (2015). A Thousand Words Are Worth a Picture: Snapshots of Printed-Word Processing in an Event-Related Potential Megastudy. Psychological science, 26(12), 1887–1897. DOI: 10.1177/0956797615603934

Faísca, L., Reis, A. & Araújo, S. (2019). Early Brain Sensitivity to Word Frequency and Lexicality During Reading Aloud and Implicit Reading. Frontiers in Psychology 10, DOI:10.3389/fpsyg.2019.00830

Fiorenzo, A., Delorme A., & Makeig, S. (2018). Applying dimension reduction to EEG data by Principal Component Analysis reduces the quality of its subsequent Independent Component decomposition. *NeuroImage* 175, 176–187, DOI:10.1016/j.neuroimage.2018.03.016.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. Brain and language, 140, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Friederici, A. D., (2002). Towards a neural basis of auditory sentence processing. Trends in Cognitive Sciences, 6(2), 78-84, https://doi.org/10.1016/S1364-6613(00)01839-8.

Frisch, S., & Schlesewsky, M. (2001). The N400 reflects problems of thematic hierarchizing. Neuroreport. 12(15), 3391-4. doi: 10.1097/00001756-200110290-00048.

Garraffa, M. & Fyndanis, V. (2020) Linguistic theory and aphasia: an overview, Aphasiology, 34(8), 905-926, DOI: 10.1080/02687038.2020.1770196

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Parkkonen, L. & Hämäläinen, M. (2014). MNE software for processing MEG and EEG data, NeuroImage, 86(1), 446-460, ISSN 1053-8119

Gratton, G. (1998). Dealing with artifacts: The EOG contamination of the event-related brain potential. Behaviour Research Methods, Instruments, & Computers, 30, 44–53

Griffiths, T. L., Vul, E., and Sanborn, A. N. (2012). Bridging Levels of Analysis for Probabilistic Models of Cognition. Curr. Dir. Psychol. Sci. 21, 263–268. DOI:10.1177/0963721412447619

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. Journal of Cognitive Neuroscience, 29(4), 677–697. DOI:10.1162/jocn_a_01068

Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage*, *173*, 434–447. DOI:10.1016/j.neuroimage.2018.02.044

Hagoort, P., Brown, C. & Groothusen, J. (1993) The syntactic positive shift (sps) as an erp measure of syntactic processing, Language and Cognitive Processes, 8(4), 439-483. DOI:10.1080/01690969308407585

Hale, J., Dyer, C., Kuncoro, A. & Brennan, J. (2018). Finding syntax in human encephalography with beam search. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2727-2736. Melbourne, Australia. Association for Computational Linguistics. DOI:10.18653/v1/P18-1254

Hauk, O., & Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology, 115(5), 1090–1103. DOI:10.1016/j.clinph.2003.12.020

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage*, *30*(4), 1383–1400. DOI:10.1016/j.neuroimage.2005.11.048

Heim, S. (2005). The structure and dynamics of normal language processing: insights from neuroimaging. *Acta neurobiologiae experimentalis*, *65*(1), 95–116.

Hollenstein, N., Rotsztejn, J., Troendle, M. *et al.* (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Sci Data* 5, https://doi.org/10.1038/sdata.2018.291

Hollenstein, N., Barrett, M., & Beinborn, L. (2020). Towards Best Practices for Leveraging Human Language Processing Signals for Natural Language Processing. In Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources, 15-27. European Language Resources Association. https://www.aclweb.org/anthology/2020.lincr-1.3

Ishkhanyan, B., Sahraoui, H., Harder, P., Mogensen, J., & Boye, K. (2017). Grammatical and lexical pronoun dissociation in French speakers with agrammatic aphasia: A usage-based account and REF-based hypothesis. Journal of Neurolinguistics, 44, 1–16. https://doi.org/10.1016/j.jneuroling.2017.02.001

Ishkhanyan, B., Michel Lange, V., Boye, K., Mogensen, J., Karabanov, A., Hartwigsen, G., & Siebner, H. R. (2020). Anterior and Posterior Left Inferior Frontal Gyrus Contribute to the Implementation of Grammatical Determiners During

Language Production. Frontiers in Psychology, 11, 685. https://doi.org/10.3389/fpsyg.2020.00685

Iwasaki, M., Kellinghaus, C., Alexopoulos, A. V., Burgess, R. C., Kumar, A. N., Han, Y. H., Lüders, H. O., & Leigh, R. J. (2005). Effects of eyelid closure, blinks, and eye movements on the electroencephalogram. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, *116*(4), 878–885. https://doi.org/10.1016/j.clinph.2004.11.001

Kemmerer, D. (2014). Word classes in the brain: Implications of linguistic typology for cognitive neuroscience. Cortex, 58, 27-51, https://doi.org/10.1016/j.cortex.2014.05.004.

Keren, A.S., Yuval-Greenberg, S. & Deouell, L. Y. (2010). Saccadic spike potentials in gamma-band EEG: Characterization, detection and suppression. NeuroImage, 49(3), 2248-2263, https://doi.org/10.1016/j.neuroimage.2009.10.057.

King, J. R., Charton, F., Lopez-Paz, D. & Oquab, M. (2020). Back-to-back regression: Disentangling the influence of correlated factors from multivariate observations, NeuroImage, 220. DOI:10.1016/j.neuroimage.2020.117028.

King, J. W. & Kutas, M. (1995). A Brain Potential Whose Latency Indexes the Length and Frequency of Words. CRL Newsletter, 10(2), 3-9

Kornrumpf, B., Niefind, F., Sommer, W., & Dimigen, O. (2016). Neural Correlates of Word Recognition: A Systematic Comparison of Natural Reading and Rapid Serial Visual Presentation. Journal of cognitive neuroscience, 28(9), 1374–1391. DOI: 10.1162/jocn_a_00977

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annual review of psychology, 62, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. Science, 207(4427), 203–205. https://doi.org/10.1126/science.7350657

Lins, O. G., Picton, T. W., Berg, P., & Scherg, M. (1993). Ocular artifacts in recording EEGs and event-related potentials. II: Source dipoles and source components. *Brain topography*, *6*(1), 65–78. https://doi.org/10.1007/BF01234128

Luck, S. J. (2005). An introduction to the event-related potential tech-nique. Cambridge, MA: MIT Press.

Mellem, M., Bastiaansen, M., Pilgrim, L., Medvedev, A. & Friedman, R. (2012). Word Class and Context Affect Alpha-Band Oscillatory Dynamics in an Older Population. Frontiers in Psychology, 3. http://doi.org/10.3389/fpsyg.2012.00097

Miceli, G., Silveri, C. & Nocentini, U.C. (1988). Patterns of dissociation in comprehension and production of nouns and verbs. Aphasiology 2, 351–358.

Münte, T. F., Wieringa, B. M., Weyerts, H., Szentkuti, A., Matzke, M., & Johannes, S. (2001). Differences in brain potentials to open and closed class words: class and frequency effects. *Neuropsychologia*, *39*(1), 91–102. https://doi.org/10.1016/s0028-3932(00)00095-6

Neville, H. J., Mills, D. L., & Lawson, D. S. (1992). Fractionating language: different neural subsystems with different sensitive periods. *Cerebral cortex (New York, N.Y. : 1991)*, *2*(3), 244–258. https://doi.org/10.1093/cercor/2.3.244

Nielsen, S. R., Boye, K., Bastiaanse, R., & Lange, V. M. (2019). The production of grammatical and lexical determiners in Broca's aphasia. Language, Cognition and Neuroscience, 34(8), 1027–1040. DOI:10.1080/23273798.2019.1616104

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. Journal of Memory and Language, 31(6), 785–806. DOI:10.1016/0749-596X(92)90039-Z

Osterhout, L., Bersick, M., & McKinnon, R. (1997). Brain potentials elicited by words: word length and frequency predict the latency of an early negativity. Biological psychology, 46(2), 143–168. DOI:10.1016/s0301-0511(97)05250-2

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behaviour made easy. Behavior Research Methods. DOI:10.3758/s13428-018-01193-y

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. NeuroImage, 45(1 Suppl), S199–S209. https://doi.org/10.1016/j.neuroimage.2008.11.007

Pulvermüller, F., Lutzenberger, W., & Birbaumer, N. (1995). Electrocortical distinction of vocabulary types. Electroencephalography and clinical neurophysiology, 94(5), 357–370. https://doi.org/10.1016/0013-4694(94)00291-r

Pulvermuller, F., Mohr, B., Sedat, N., Hadler, B. & Rayman, J. (1996) Word class-specific deficits in Wernicke's aphasia, Neurocase, 2:3, 203-212, DOI: 10.1080/13554799608402397

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. , Semantic Scholar.

Rayner, K., Schotter, E. R., Masson, M. E. J., Potter, M. C., & Treiman, R. (2016). So Much to Read, So Little Time: How Do We Read, and Can Speed Reading Help? Psychological Science in the Public Interest, 17(1), 4–34. DOI:10.1177/1529100615623267

Ren, Y. & Xiong, D. (2021). CogAlign: Learning to align textual neural representations to cognitive language processing signals. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Joint Conference on Natural Language Processing (Volume 1: Long Papers), 3758-3769. Online. Association for Computational Linguistics.

Ritchie, J. B., Masson, H. L., Bracci, S. & Op de Beeck, H. P. (2021). The unreliable influence of multivariate noise normalisation on the reliability of neural dissimilarity. NeuroImage, 245, DOI:10.1016/j.neuroimage.2021.118686.

Rouam, S. (2013) False Discovery Rate (FDR). In Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (eds) *Encyclopaedia of Systems Biology.* Springer, New York, NY. DOI:10.1007/978-1-4419-9863-7_223

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. Journal of neural engineering, 16(5), 051001. DOI:10.1088/1741-2552/ab260c

Sanei, S. & Chambers, J. A. (2021) EEG preprocessing and Machine Learning. 2022. Wiley. John Wiley & Sons ltd.

Segalowitz, S. J., & Lane, K. C. (2000). Lexical access of function versus content words. Brain and language, 75(3), 376–389. https://doi.org/10.1006/brln.2000.2361

Sereno, S. C., Hand, C. J., Shahid, A. Mackenzie, I. J. & Leuthold, H. (2020) Early EEG correlates of word frequency and contextual predictability in reading. Language, Cognition and Neuroscience, 35(5), 625-640. DOI: 10.1080/23273798.2019.1580753

Schuster, S., Hawelka, S., Hutzler, F., Kronbichler, M., & Richlan, F. (2016). Words in context: The effects of length, frequency, and predictability on brain responses during natural reading. Cerebral Cortex, 26(10), 3889–3904. https://doi.org/10.1093/cercor/bhw184

Schwartz, D. & Mitchell, T. (2019). 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), 43-57. Minneapolis, Minnesota. Association for Computational Linguistics.

Speer, R., Chin, J., Lin, A., Jewett S., & Nathan, L. (2018). LuminosoInsight/wordfreq: v2.2. Zenodo. https://doi.org/10.5281/zenodo.1443582

Sun, L., Liu, Y. & Beadle, P. J. (2005). Independent component analysis of EEG signals Proceedings of 2005 IEEE International Workshop on VLSI Design and Video Technology, 219-222, doi: 10.1109/IWVDVT.2005.1504590.

Swinney, D. A., Zurif, E. B., & Cutler, A. (1980). Effects of sentential stress and word class upon comprehension in Broca's aphasics. Brain and language, 10(1), 132–144. https://doi.org/10.1016/0093-934x(80)90044-9

Synigal, S. R., Teoh, E. S. & Lalor, E. C. (2020). Including Measures of High Gamma Power Can Improve the Decoding of Natural Speech From EEG. Frontiers in Human Neuroscience, 14. DOI:10.3389/fnhum.2020.00130

Tanner, D., Norton, J. J. S., Morgan-Short, K., & Luck, S. J. (2016). On high-pass filter artifacts (they're real) and baseline correction (it's a good idea) in ERP/ERMF Analysis. Journal of Neuroscience Methods, 266, 166–170. https://doi.org/10.1016/j.jneumeth.2016.01.002

ter Keurs, M., Brown, C. M., Hagoort, P. & Stegeman, D. F. (1999). Electrophysiological manifestations of open- and closed-class words in patients with Broca's aphasia with agrammatic comprehension: An event-related brain potential study, Brain, 122(5), 839–854. DOI:10.1093/brain/122.5.839

Toneva, M. & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Proceedings of the 33rd International Conference on Neural Information Processing Systems.* Curran Associates Inc., Red Hook, NY, USA, Article 1339, 14954–14964

Tuckute, G., Therese, S., Hansen, N. P., Steenstrup, D., Hansen, L. K. & Maex., R. (2019). Single-Trial Decoding of Scalp EEG under Natural Conditions. Intell. Neuroscience. DOI:10.1155/2019/9210785

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems 5998--6008

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2011). Nouns and verbs in the brain: a review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience and biobehavioral reviews*, *35*(3), 407–426. https://doi.org/10.1016/j.neubiorev.2010.04.007

Yudes, C. Domínguez, A., Cuetos, F. & Vega, M. (2016). The time-course of processing of grammatical class and semantic attributes of words: Dissociation by means of ERP. Psicológica. 37.

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič, J (Jr)., Hlaváčová, J., Kettnerová, V., Urešová, Z. et al. 2017. *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, 1–19, Vancouver, Canada. Association for Computational Linguistics.

Zhang, P., & Peng, J. (2004). SVM vs regularised least squares classification. Proceedings - International Conference on Pattern Recognition, 1, 176-179.