Title

Examining the Time-course of Information Retrieval During Predictive Processing in Human Language Comprehension

Author
Trammel, Timothy George

Publication Date
2024

Peer reviewed|Thesis/dissertation

Examining the Time-course of Information Retrieval During Predictive Processing in Human Language Comprehension

By

TIMOTHY GEORGE TRAMMEL

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PSYCHOLOGY

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

____

Tamara Swaab, chair

____

Matthew Traxler

____

Steven Luck

Committee in Charge

2024

Chapter 2 was published in its entirety in NeuroImage*:*

Trammel, T., Khodayari, N., Luck, S. J., Traxler, M. J., & Swaab, T. Y. (2023). Decoding semantic relatedness and prediction from EEG: A classification method comparison. NeuroImage, 277, 120268. https://doi.org/10.1016/j.neuroimage.2023.120268

Acknowledgements

This dissertation is the culmination of a years-long journey, and it could not have been completed without guidance and support from many people. First, I would like to thank my major advisor, Tamara Swaab, for twice allowing me the opportunity to join her lab and then being the kindest and most supportive mentor that a graduate student could ever have. Tamara granted me the freedom to pursue my interests in machine-learning methods while teaching me about the field of neurolinguistics, about which I knew almost nothing. Above all else, she showed a great deal of patience and understanding during difficult times throughout my time in her lab. I am also grateful to Matt Traxler and Steve Luck for their invaluable mentorship and support over the past five years on my advising committee. I am especially thankful for their understanding as plans changed unexpectedly over the past year. Thank you also to Dave Corina and Emily Morgan for asking tough questions and providing insightful feedback during my qualifying exams. I am also appreciative of Emilio Ferrer whose multivariate statistics seminar inspired me to perform a performance comparison between classification models (Chapter 2). I am also incredibly grateful for my past mentors for inspiring me to pursue a career in scientific research: Tom Carter, Dawn Strongin, Kelly Cotter, and Gary Williams. I consider myself quite lucky to have worked so closely with, and learn from, these talented scientists.

I also want to thank my collaborators and lab mates from the Swaab-Traxler Lab: Agnes Gao, Natalia Khodayari, Zoe Yang, Katie Sendek, Elaina Jahanfard, Eleonora Beier, Noemi Diaz, and Sarah Wang. Their support and camaraderie have been priceless and have done wonders for my mental well-being! I am particularly thankful to Natalia for being such a wonderful collaborator and sounding board over the years. I would also like to extend a special thanks to Trevor Brothers. My work is quite literally built upon the foundations that he laid.

A special thanks goes out to the Dim Sum Group. We started way too late, but I am grateful that we began.

Above all, I want to thank my wife, Kimberly, and my son, George, for their unwavering support and belief in me throughout this journey. Without them, I could not have achieved this, nor would the journey have been worth it to me.

Abstract

Prediction plays a critical role in comprehending human language. Many theoretical and computational models have attempted to characterize how we use context to facilitate language processing in noisy environments either with or without relying on predictive processing. Despite these attempts, we do not yet have a complete understanding of the role of prediction in language processing. Predictive coding models have recently gained popularity as potential architectures for the role of prediction during language comprehension. These models suggest that predictions about bottom-up inputs are continuously generated from higher cortical levels to lower levels in a hierarchical manner – i.e., a particular level generates predictions about the next lower level. As bottom-up input is encounter by each level of processing, prediction error is computed by comparing the input with the top-down prediction. The goal of this dissertation was to assess whether predictive coding models can account for the time course of information retrieval during predictive language processing. Specifically, the studies described examine the time course of pre-activations of lexical and sub-lexical features in both monolinguals and bilinguals using a combination of decoding electroencephalogram (EEG) with machine-learning classifiers and mass univariate event-related potential analysis.

Chapter 1 describes an experiment that compared three frequently used models for signal classification – support vector machines (SVM), linear discriminant analysis (LDA), and random forest (RF) to determine which is best-suited for analyzing word pair prediction paradigms. Results showed that SVM was the best performing classifier of the three within two data sets from separate visual word priming paradigms. Chapter 2 describes an experiment which used EEG decoding with SVM classifiers and mass univariate ERP analyses to identify the time course of information retrieval prior to the onset of accurately predicted, related but inaccurately

predicted, and unrelated target words during a visual word priming prediction paradigm. In addition to this pre-stimulus information retrieval, these analyses were used to investigate the effects of prediction error. The results of this study showed that semantic information, such as concreteness, is retrieved earlier than visual feature information, like word length, and that unrelated words had greater prediction error than predicted or related but inaccurately predicted words. Finally, Chapter 3 describes an experiment that extends the results of Chapter 2 by using the same paradigm and analyses with Spanish-English bilingual participants. The results of this study showed that bilinguals reading words in their second language (L2) retrieve anticipated information in a similar fashion as monolinguals. Semantic information preceded visual information and unrelated words showed evidence of greater prediction error than did predicted or related words that were not accurately predicted.

Together, these experiments support predictive coding models of language processing in both monolinguals and bilinguals during word recognition. Both groups predict higher-level features (concreteness) before lower-level features (word length) of anticipated words and calculate prediction error when they make inaccurate predictions.

Table of Contents

Chapter 1

1. Introduction

Language processing is hard. Gwilliams and Davis (2020) characterize this difficulty by pointing out the great challenges faced, and resources invested, in creating state-of-the-art artificial language processors. Nevertheless, humans have the remarkable ability to comprehend language rapidly and accurately in several modalities (written, spoken, or sign), often dealing with both internally and externally noisy environments. Current theories and models of language processing assume that efficient language processing therefore relies on predictive processing, defined here as the ability to use the context and prior knowledge to pre-activate upcoming linguistic features (and, potentially, non-linguistic features) prior to encountering the sensory input. The notion that prediction leads to pre-activation of imminent linguistic representations is currently debated, in part because it has been methodologically challenging to measure the presence and content of pre-activated features.

In this chapter, I will provide some evidence from prior research that a complete model of language comprehension should be cross-modal, allow for contextual effects beyond those that are purely linguistic, and be driven, at least in part, by prediction. A critical mechanism of prediction – potentially among several – is priming. Priming happens when one input facilitates the processing of a later input. The present work used machine learning approaches to examine whether predictive pre-activation of various linguistic features can be decoded from electrophysiological signals recorded during language processing during word pair priming studies.

Before introducing the studies that I have performed to address this question, I will provide an overview of psycholinguistic models of language processing at multiple processing levels (phonemic/orthographic, lexical, sentential, syntactic, and semantic) and across visual and spoken modalities. This overview is followed by a discussion of prediction frameworks. Next, I will review two important methodologies – event-related potentials (ERPs) from electroencephalography (EEG) and multivariate pattern analysis (MVPA) of EEG using machine-learning classifiers (frequently referred to as decoding) – used to investigate language comprehension, predictive processing, and their relation to other domain general cognitive processes. These sections are accompanied by a discussion of how to use these techniques in complementary ways to further explore prediction in language. The final section discusses the implications of the reviewed literature and motivations for the present work.

2. Models of language comprehension

Each of the various modalities of language comprehension present challenges that must be overcome by the comprehender. This includes speech, written language, and sign language. This review will focus primarily on speech and written language. While the studies presented in the following chapters focus on visual word recognition, many models of language comprehension treat word recognition as one part of the hierarchy of language processing. Indeed, in models that rely upon top-down predictions, higher levels of representation are relevant for word recognition regardless of input modality. Furthermore, several models of word recognition assume that both on phonological and orthographic representations are activated, even during visual word recognition. Therefore, a more complete picture of word recognition requires a basic understanding of different models. Speech comprehension poses a challenging problem in that the signal is continuous and noisy. The speech signal must be segmented so that the parts bearing meaning – for example, words – can be identified. This must be done while handling the noise created by the speaker – in the form of accents, speech impairments, and other individual speaker characteristics –by the environment – such as ambient sound, interruptions, and distortions created by obstacles – and internally due to ambiguous linguistic input or lack of focus. To compound these challenges, all this processing must occur in a matter of a few hundred milliseconds to maintain discourse comprehension. Written language is already discretized, and readers can also easily re-read prior parts of the message when comprehension is difficult. However, it poses its own challenges to comprehenders in the forms of noise in the writing – sloppy handwriting, misspellings, and poor grammar – and ambiguity in words, phrases, and sentences, for example when words are lexically ambiguous, or when sentences have an ambiguous structure. Due to these challenges in processing language, several theories have proposed that comprehenders rely on the prior context to facilitate the processing of incoming words. Some theories assume that comprehenders use context to predict upcoming input while others suggest that the prior context facilitates integration of new input only after encountering said input (see Traxler, 2014 and Traxler & Swaab, 2024 for further review). Here I focus on seminal models of speech and reading comprehension that have been, or could be, implemented computationally to provide views on how language may be processed using contextual constraints either with or without a mechanism for predictive processing.

2.1 Spoken word recognition

Presented here are three seminal models of speech processing. These models were chosen to provide a non-exhaustive sample of models that range from mostly ignoring predictive processing (*cohort model*) to relying on it heavily (*lossy context surprisal model and information exchange model)*. Additionally, this selection of models includes both isolated word recognition (*cohort*) and word recognition within a sentence context (cohort, *lossy context surprisal and information exchange model*). It is important to note that while these are speech recognition models, many of the principles could apply to visual word recognition with little or no adaptation. Several of the models which I discuss here and in the following section are connectionist models. Connectionist models are designed, in some fashion, to resemble the workings of the human nervous system (Rohde & Plaut, 2003). These models are based on parallel, distributed processing (Rumelheart & McClelland, 1988). Connectionist models are the precursors to current prediction models; particularly those based on neural networks. The first model discussed here is one of the earliest connectionist models. The *cohort model* (Marslen-Wilson & Welsh, 1978; Marslen-Wilson & Tyler, 1980; Marslen-Wilson, 1987, Gaskell & Marslen-Wilson, 1997), initiates word processing at the onset of the incoming word and uses context to facilitate processing of the word and integrate it into the context. The second model, the *lossy context surprisal model* (Futrell et al., 2020), is a probabilistic model that uses a representation of the prior context to predict upcoming words. The third model, referred to here as the *information exchange model,* is also a probabilistic model that goes beyond word recognition and suggests that probabilistic prediction is the main driver of speech comprehension.

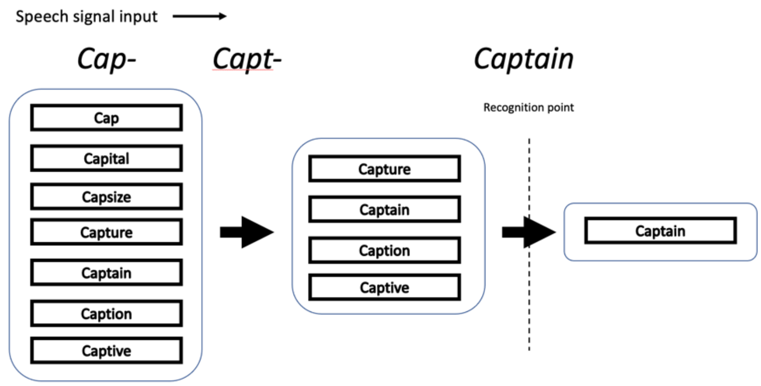

Figure 1. An interpretation of the cohort model. A word initial cohort is formed after about 100-150ms of the speech input. As more speech input is heard, the cohort is narrowed down until the point of recognition is reached. Note that the actual word initial cohort would be much larger than pictured here.

Cohort Model. The first model presented here is the *cohort model* (see figure 1, Marslen-Wilson & Welsh, 1978; Marslen-Wilson & Tyler, 1980; Marslen-Wilson, 1987). Marslen-Wilson's cohort model (Marslen-Wilson & Welsh, 1978; Marslen-Wilson & Tyler, 1980; Marslen-Wilson, 1987) evolved from prior models of the organization of the mental lexicon, including the logogen model (Figure 2), according to which each word is represented as a separate abstract unit – a logogen – in a network that is organized as a function of semantic relations between words (Morton, 1969). Another prominent model of the organization of the mental lexicon was the serial lexical frequency ordered bin model (Forster, 1976), which assumed that information about a given word form is part of a single master file, or bin, and addressed by domain-specific codes (Allport et al., 1981). Whereas this latter model assumes a serial frequency ordered search, where each candidate must be discarded before another candidate can be activated, the logogen model assumed that sensory input would activate the intended representation, and that this activation would spread to semantically related representations, thereby facilitating their recognition should they subsequently be encountered. It will become clear that the *Cohort model* requires parallel spoken word recognition as I review the model below.

Marslen-Wilson (1987) asserts that spoken word recognition breaks down into three basic functions: *access*, *selection*, and *integration*. *Access* is the relationship between the recognition process to the sensory input. Any system for recognizing spoken words must be capable of mapping speech signals to representations in the mental lexicon. *Integration* is the relationship of the recognized word to higher-level representations of the preceding utterance. To complete recognition, the system must allow for the integration of the syntactic and semantic information

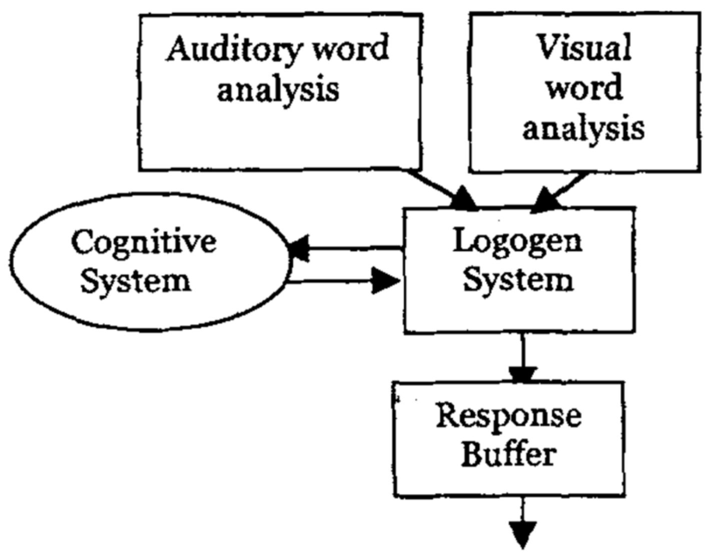

Figure 2. Early version of Morton's (1969; figure from Coltheart et al, 2001) logogen model for reading aloud. In this model auditory and visual word analyses feed forward to activate logogens that are mental lexical representations of words which feed forward to generate a verbal response.

of the word being recognized into the existing sentence or discourse context. *Selection* is the interface between access and integration. The system must discriminate between accessed word-forms to select the word-form that best matches the input. During this process of lexical selection, the initial cohort is narrowed down as more bottom-up speech input becomes available and those representations that no longer match the speech input are dropped out of the cohort. This cohort narrowing continues until a *recognition point* is reached in which only one mental lexical representation is consistent with the signal input. These three required functions must be performed within the short timeframe during which listeners are shown to be capable of correctly identifying words (Marslen-Wilson, 1973, 1985, 1987; Marslen-Wilson & Tyler, 1975, 1980).

Cohort assumes that context can facilitate lexical selection: When a word is heard, the recognition point of words can occur prior to the entire word being heard. This early selection constraint rules out serial models and forces a parallel model of selection and access. This results in three functional requirements for models of the spoken word recognition process: *multiple access*, which is the ability to access multiple candidate word forms from the input signal to their representations in mental lexicon; *multiple assessment*, which is the ability to evaluate the syntactic and semantic appropriateness of candidate words at a time at which multiple candidate words are compatible with the input; and *real-time efficiency*, which is the need for the access and selection processes to complete within a real timeframe of ~200ms from word onset (Marslen-Wilson, 1987). The cohort model assumes that the initial access to word-form representations is entirely driven by the sensory signal, even when words are embedded in a highly constraining context. The model hypothesized that a "word-initial cohort" of lexical representations that match the initial sensory input is activated within 100-150ms of the onset of the incoming speech signal. This corresponds to *multiple access.* The process of narrowing down the word initial cohort to those candidates that are still compatible with the speech signal can be facilitated in constraining sentence contexts, i.e., the zooming in to the lexical candidate that represents the speech signal is facilitated. This corresponds to *multiple assessment*. Marslen-Wilson (1987) argues that this system allows for optimal real-time efficiency as words will be recognized as soon as the auditory signal permits according to the currently available contextual constraints; thus, satisfying the requirement for *real-time efficiency*. The early versions of the model (Marslen-Wilson & Welsh, 1978; Marslen-Wilson & Tyler, 1980) could not account for word frequency effects and assumed matching in an all-or-nothing fashion which required unrealistically perfect matches to sensory and contextual inputs, given the great variability in the acoustic realization of phonemes in the speech signal. These issues necessitated amendments to the model to allow for an activation-based representation of the cohort (Marslen-Wilson, 1987)

In a seminal study, Zwitserlood (1989) performed a series of experiments to examine the two main predictions of the cohort model, namely that 1) only the sensory input determines access to and activation of candidates in the cohort, and 2) that a biasing sentence context can speed up the process of lexical selection such that less of the speech signal is needed to zoom in on the correct word form representation In this study participants heard sentences and were asked to perform a visual lexical decision task at a critical point in the sentence. The contextual bias of the sentence contexts was manipulated, as shown in the examples below.

1. The next word is *captain.*

2. They mourned the loss of their *captain.*

3. With dampened spirits the men stood around the grave. They mourned the loss of their *captain.*

In addition to the context manipulation, the probing point at which the lexical decision stimulus was presented was varied, in 4 conditions, 1) before the onset of the critical word to probe potential pre-activation of the critical probe word, 2) at word onset, to examine effects of context on lexical access 3) before the recognition point to examine effects of context on lexical selection 4) after the recognition point to examine the effects of context on lexical integration. The recognition point of spoken words in isolation was determined in a gating task prior to the actual experiment. Participants were presented with increasing stretches of the speech signal of individual words, until most participants converged on the same word, i.e. the recognition point. The visual target words for the lexical decision task were semantically related either to the critical word of the priming stimulus or a competing word – Zwitserlood gives the example of Dutch words for “captain” (kapitein – critical word) and “capital” (kapitaal – competitor word) with “ship” (schip) and “money” (geld) as the respective related visual probe target words. The probing points were determined based on the results of the gating task. Differences in facilitation of the critical-related target words versus the competitor-related target words at a given probing point indicated at which point in time the critical word was selected. Zwitserlood found no evidence of differential facilitation prior to the onset of the critical word or when the first phoneme was presented. However, lexical decision was facilitated before the recognition point. This suggests that context does not pre-activate contextually appropriate lexical items as both appropriate and inappropriate items were activated after word onset, but contextual effects do occur prior to the sensory input being sufficient to disambiguate the activated lexical items.

In contrast to Zwitserlood’s findings, a more recent, two-experiment study found evidence that semantic features of the upcoming word are indeed activated prior to the onset of the critical word during auditory sentence processing (Heikel et al., 2018). While listening to auditory sentences that were either highly constraining regarding concreteness (experiment 1) or regarding animacy (experiment 2), participants sometimes encountered an unexpected delay directly preceding the critical word. Electroencephalogram (EEG) was recorded from the subjects while listening to these sentences. Heikel and colleagues used machine-learning techniques to classify the EEG signal from the silent period prior to critical word onset according to the animacy or the concreteness of said word, and they were able to reliably decode animacy and concreteness features. These findings suggest that context can indeed predictively pre-activate features of an upcoming word. The next model features prediction as a driving factor in speech processing.

Lossy Context Surprisal Model. While the cohort model relegates contextual effects to constraints on selection of the candidate that matches the (partial) speech input, other models assume that contextual effects can be used to predict imminent input. One such model is the *lossy context surprisal model* (figure 3), which is a combination of predictive expectation-based, models that are rooted in information theory and the idea of imperfect, or *lossy,* representations of the current context.

The lossy surprisal model was proposed by Futrell and colleagues (2020), to integrate expectation and memory-based models of spoken word recognition. Expectation-based models treat processing difficulty as a function of word predictability, positing that optimal processors will have already performed much of the work related to processing the word prior to encountering it when context makes words predictable (Futrell et al., 2020; Hale 2001; Jurafsky, 2003). A prime example of expectation-based models is surprisal theory, which is based on information theory (Shannon, 1948) and suggests that a word's processing cost is determined by how predictable the word is given its context. (Futrell et al., 2020; Hale, 2001; Levy, 2008). Futrell et al. suggest that surprisal, serves as a link to prediction and predictive coding (Clark, 2013; Friston, 2009; 2010; Friston & Keibel, 2009; Rao & Ballard, 1999), which is discussed in more detail in the section on predictive processing. Memory-based models posit that integration of words in sentence contexts requires retrieval of representations of those words from working memory and that this retrieval process can be the source of incremental sentence processing difficulty (Futrell et al., 2020; Gibson 1998; Lewis & Vasishth, 2005). The example of memory-based models presented by Futrell and colleagues is the dependency locality theory which suggests that processing difficulty is, in part, due to difficulty in integrating words that are linearly far apart from each other in the sentence with the cost of integration increasing with distance (in the sentences below: integrating *threw out* in 4 and 6 will be less costly than in 5 and 7, Gibson 1998; 2000, example given in Futrell et al., 2020).

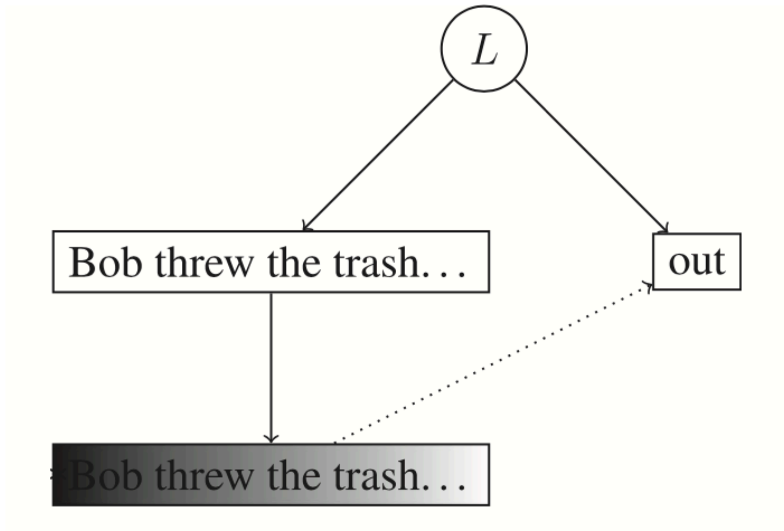

Figure 3. Lossy context surprisal model (Futrell et al., 2020). Within this model, the prior linguistic input (Bob threw the trash...) is not used to predict the upcoming word. Rather, a lossy, or incomplete, memory representation is used as the prior context.

4. Bob threw out the trash.
5. Bob threw the trash out.
6. Bob threw out the old trash that had been sitting in the kitchen for several days.
7. Bob threw the old trash that had been sitting in the kitchen for several days out.

Futrell and colleagues present their model, *lossy context surprisal*, and suggest that it unifies traditional surprisal theory with dependency locality theory. Their model extends surprisal theory with a memory model that includes four characteristics: *incrementality of memory, linguistic knowledge, inaccessibility of content,* and a *linking hypothesis.* By *incrementality of memory,* Futrell and colleagues mean that working memory during sentence processing can be characterized as a probabilistic memory encoding function that takes a memory representation and combines it with the current word to produce the next memory representation. *Linguistic knowledge* is defined as the ability of listeners to access some probability model – potentially Bayesian – providing the distribution of potential upcoming words given a context. Although referred to as linguistic knowledge, this probability model may incorporate non-linguistic context – such as information about the speaker. For example, a Bayesian version of this model may incorporate the prior context of the current discourse, existing knowledge of what words are semantically related to the given context, the mood of the speaker, and their proclivity toward sarcasm to determine a distribution of future input. *Inaccessibility of* content means that comprehenders do not have access to the true linguistic contexts, but only to their memory representations. Finally, the *linking hypothesis* is that incremental processing difficulty for any given word is proportional to the surprisal of that word given the previous memory representation.

To support the claim that *lossy context surprisal* theory integrates the prediction of surprisal theory with those of the dependency locality theory, Futrell and colleagues present evidence showing that *lossy context surprisal* can account for structural forgetting. Structural forgetting happens when comprehenders forget, or misremember, the beginning of a sentence by the time they reach the end. This can sometimes cause ungrammatical sentences to appear more understandable and acceptable than grammatical sentences (Futrell, 2020; Vasishth, 2010). Futrell and colleagues provide the sentences below as an example of this effect.

8. The apartment that the maid who the cleaning service sent over was well-decorated.
9. The apartment that the maid who the cleaning service sent over cleaned was well-decorated.

English speakers reliably rate Sentence 8 to be equally or more acceptable than Sentence 9, despite Sentence 8 being ungrammatical (Frazier, 1985; Gibson & Thomas, 1999).

To summarize, the *lossy context surprisal* model provides a model of word recognition in context that is driven by prediction and incorporates context effects from memory. This model therefore provides a bridge between expectation-based and memory-based theories of speech processing and can account for a phenomenon, structural forgetting, that neither family of models could resolve alone. While the lossy context surprisal model incorporates predictive effects and memory effects into word recognition, it does not explain how acoustic-phonetic, phonological, morphological, and semantic information interact during speech processing. The next model attempts to do exactly that.

Information Exchange Model. The prior models discussed here focus primarily on how context influences word recognition in sentence or discourse contexts, but they do not provide a mechanism for how a representation of the overall discourse is developed. Gwilliams and Davis (2020) aim to address this by taking a broader perspective, formally describing the speaker-listener interaction as a *communication system* (see figure 4, defined by Shannon, 1948). The goal of this system is to exchange information between the speaker and the listener, where the speaker is the *source* of the information. The purpose of the speaker is to formulate an intended message and produce an auditory signal via the human vocal apparatus or *transmitter,* that *encodes* that message. Along the channel between the speaker and the listener, the auditory signal will encounter both internal and external noise that distorts the signal. The listener serves as the *destination* of the information. The purpose of the listener is to *decode* the distorted auditory signal. The combined auditory and language processing systems serve as the *receiver*, allowing the listener to reconstruct an interpretation of the originally intended message.

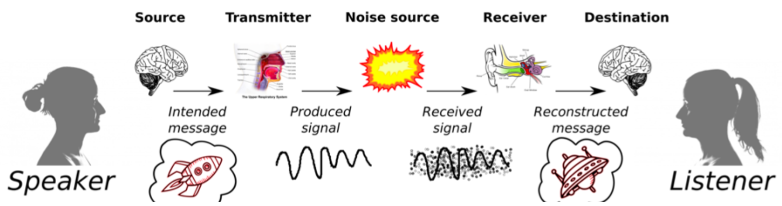

Figure 4. The speaker-listener interaction as a communication system (Gwilliams & Davis, 2020).

Gwilliams and Davis developed a simple, high-level network graph model (see figure 5) of speech comprehension from this *information theory*-driven view of the speaker-listener interaction (see Gwilliams et al., 2018 for a lower-level spoken word recognition model). Although Gwilliams and Davis do not formally name their model, I will refer to it as the *information exchange model* for ease of reference. The *information exchange model* suggests that the goal of speech comprehension can be construed as identifying a sequence of morphemes, the smallest meaningful units of speech, from the auditory signal using multiple sources of noisy information – for example, slips of the tongue and misheard words – and, ultimately, combining those morphemes with prior knowledge to understand the intended message. Prior studies have shown that surprisal – and its mathematical equivalent – entropy are frequently observed to modulate neural responses (Gwilliams & Davis, 2020; Donhauser & Baillet, 2020; Di Liberto et al., 2019; Brodbeck et al., 2018; Gaston & Merantz, 2018; Gwilliams et al., 2017; Gwilliams & Merantz, 2015; Ettinger et al., 2014; Gagnepain et al., 2012). Informed by these findings, Gwilliams and Davis suggest that Bayesian Inference provides a plausible framework for integrating the aforementioned noisy information sources. In this framework, the acoustic input is processed into speech sounds, or phonemes, using a combination of top-down predictions and bottom-up prediction error calculations. These

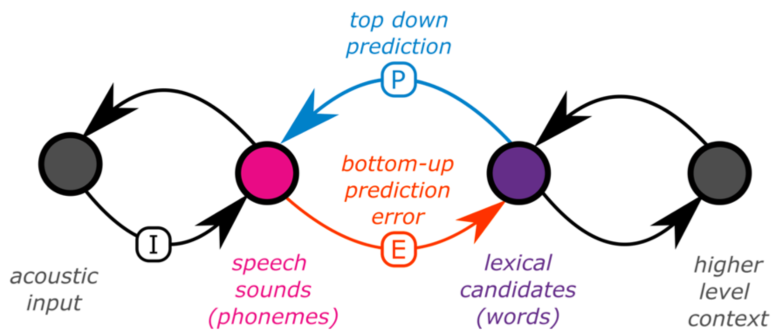

Figure 5. Network graph model of information exchange theory of speech comprehension (Gwilliams & Davis, 2020). The listener uses an internal statistical language model to form top-down predictions of lower levels of speech representation (P). Bottom-up acoustic input (I) is compared to these top-down predictions (P – I) to generate prediction error (E). Prediction error is used to update the probabilities of possible words iteratively throughout the word processing time course to converge on the best possible lexical candidate. A similar process is used at every level of representation from phonemes to high-level concepts.

phonemes are then inferred into lexical candidates, or words, and these inferences continue into higher-level interpretations. Overall, the *information exchange model* attempts to explain the mechanisms behind speech processing across all levels of representation using predictive Bayesian Inference to drive the process.

This section has provided an overview of computational models of speech processing that address word recognition within context. Now with a better understanding of how language comprehension can be modeled within speech, I turn our attention to reading.

2.2 Visual word recognition

As in speech comprehension, there are many theoretical and computational models of reading comprehension. Here I review a selection of three models that address reading in context – two of which do not use prediction as a driving factor and one which assumes prediction. I start our exploration with the older of the two models that do not explicitly use prediction.

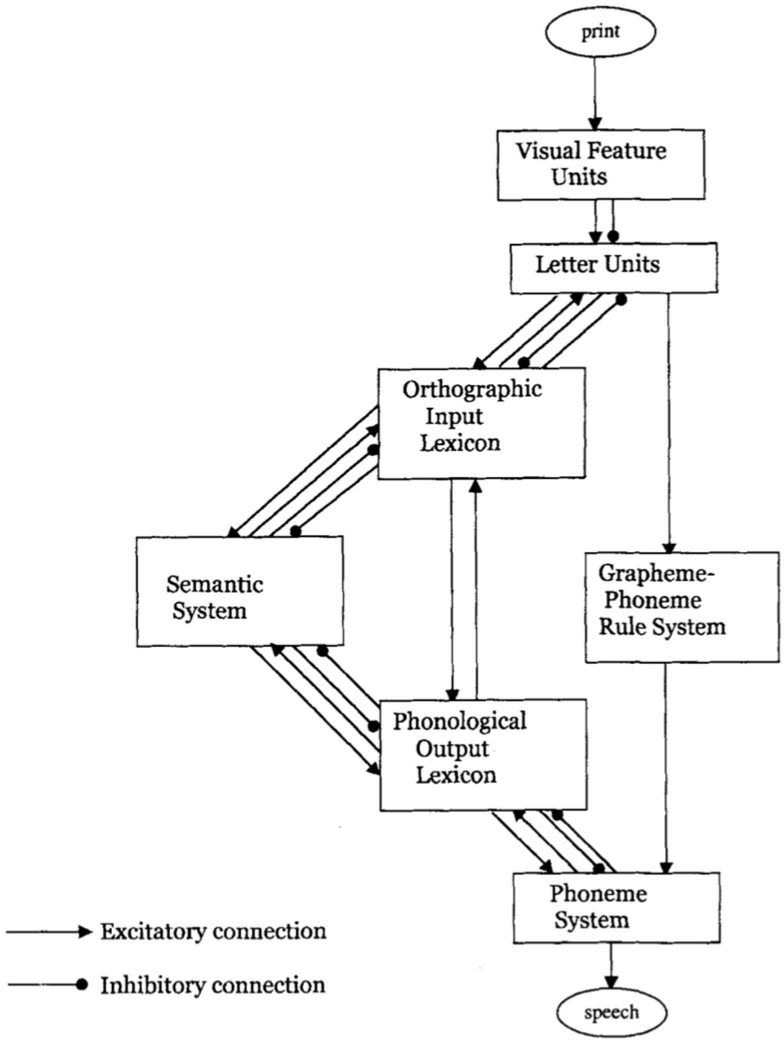

Figure 6. Dual-route cascade model (Coltheart et al., 2001)

Dual-Route Cascade Model. Coltheart and colleagues (2001) developed the *dual-route cascade model* (see figure 6) of visual word recognition and reading aloud as an evolution from logogen models (Morton, 1969). Dual routes were proposed because beginning readers almost exclusively rely on a sub-lexical route of translating graphemes into phonemes until they have read the word, while experienced readers typically skip this process unless encountering an unfamiliar word. Coltheart and colleagues used a principle called nested modeling, in which each model builds on prior models that have not been disproved and, consequently, account for all phenomena previously explained by the prior models. The model uses graded activation and is interactive allowing for both bottom-up and top-down processing for both excitation and inhibition. Although it is called "dual-route", this model technically has three routes for reading aloud: the lexical non-semantic route, the lexical semantic route, and the grapheme-phoneme rule system route. In the lexical non-semantic route, graphemes are activated in parallel, subsequently activating visual word representations. When reading aloud, these visual word representations activate phonological lexical representations, which can be used to generate the pronunciation of a word. In the lexical semantic route, orthographic lexical representations directly activate the meanings of words in the semantic system. When reading aloud, semantic representations then activate phonological lexical representations. To account for the ability to read pseudowords and generate spoken output, the third route allows for direct translation of graphemes to phonemes without activation of lexical wordforms. From a predictive processing perspective, it is critical to note that, while the *dual-route cascade model* does not explicitly mention a mechanism for predictive processing effects, the theoretical framework implemented by this computational model does include top-down input. The semantic system is undefined, and it is assumed to have both excitatory and inhibitory connections to the orthographic lexicon. This leaves room for future extensions of the model to incorporate specific models of predictive effects in the form of top-down excitation or inhibition of lexical items based on whether they fit with the current semantic context.

In summary, the *dual-route cascade model* was created specifically to account for reading aloud behaviors of individual words while also accounting for reading without naming. The model does this by allowing visual input to take one of three routes – the lexical non-semantic, lexical semantic, or phoneme-grapheme rule route. However, there is evidence from fast masked phonological priming (Grainger & Holcomb, 2009; Rastle & Brysbaert, 2006) that may disqualify this model. As I will show while discussing the following model, evidence from masked priming studies (e.g., Grainger et al., 2003) in which auditory words presented rapidly after masked visual words benefit from priming by cannot be explained by the *dual-cascade route model,* because it the dual-route cascade places phonological access too long after orthographic input. To be able to account for their results, Grainger and Holcomb proposed the Bi-modal Interactive Activation model. The next model presented is not only a dual-route model, but also incorporates parallel phonological activation, rather than the sequential activation (orthographic followed by phonological) used in *dual-route cascade model.*

Bi-modal Interactive Activation Model. Like Coltheart and colleagues (2001), Grainger and Holcomb recognize that there are cross-modal influences in word recognition. However, according to Holcomb and Grainger (2009), findings from their masked visual priming studies cannot be explained by Coltheart's dual cascade model. In one masked priming experiment (Grainger et al., 2003), homophone word or pseudoword primes (e.g., maid–MADE, brane–BRAIN) were presented for very short durations (<60s), and a mask was (#####) presented immediately after the prime. Masking of the prime in this manner prevents expectation induced priming and allows for assessment of automatic spread of activation. The question they raised in their studies is if the orthographic form of the primes could activate the phonological form of the target words, even though their visual word forms were not identical. In addition to the homophone and pseudo-homophone primes, participants were shown unrelated word or non-word primes. The findings showed that effects of phonological priming occurred only 20-30ms later than effects of visual priming. This finding of very rapid phonological regularization errors cannot be easily accounted to by Coltheart's model but can be accommodated by the *bi-modal interactive model.* In the latter model, visual words first activate visual features outside of the word recognition system which in turn activate sub-lexical orthographic units, or O-units. O-units activate a central interface between orthographic and phonologic representations that allows the O-units to be mapped onto corresponding phonological units, or P-units. In parallel, the O-units activate orthographic word forms while the P-units activate higher-level phonological representations that can, in turn, modulate the processing of higher-level orthographic representations. While Holcomb and Grainger focus primarily on visual word recognition, it is critical to note that the model can also accommodate speech processing via a spoken word input path.

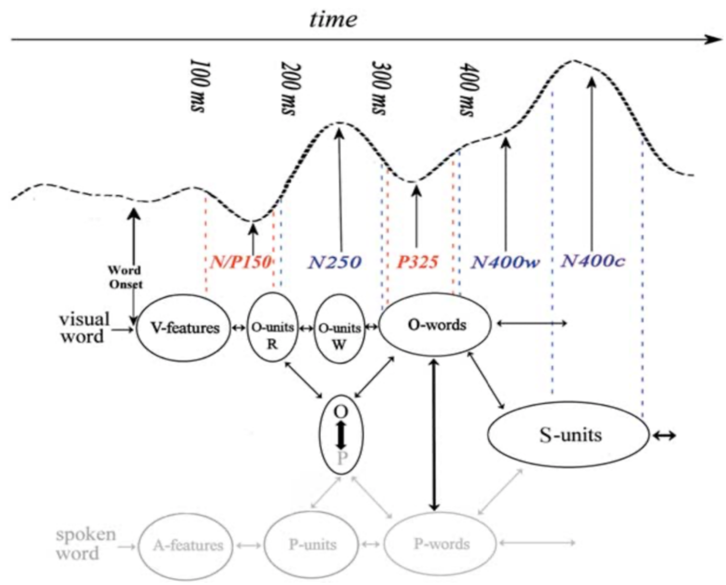

Figure 7. Time course of the bimodal interactive activation model (Grainger and Holcomb, 2009)

The *bimodal interactive activation model* accounts for cross-modal effects between phonology and orthography during word recognition in parallel. This results in the model making predictions about the time course of word recognition (Figure 7) – including predicting fast phonological priming effects that lag slightly behind orthographic effects (Grainger & Holcomb, 2009; Diependaele et al., 2009). These predictions account for the empirical observations in the aforementioned masked priming study (Grainger et al., 2003). The predictions are also supported by findings in other masked repetition priming and semantic priming experiments (Chauncey et al., 2008; Grainger et al., 2003; Holcomb & Grainger, 2006, 2007, 2009) using *event-related potentials* (ERPs) – measures of brain responses to specific events generated by averaging electrophysiological data across similar trials locked to said events (see Luck, 2014 for detailed explanation of ERP techniques). Grainger and Holcomb identify four critical ERP components to this time course. The first component is the N/P150, a posteriorly distributed component that peaks around 150ms after stimulus onset. It is sensitive to visual feature mismatches and thought to be driven by featural overlap between prime stimuli and target stimuli (Grainger & Holcomb, 2009; Chauncey et al., 2008). The second component, the N250, is negative going wave that can begin as early as 110ms after stimulus onset, peaks around 250ms, has a wide scalp distribution that is maximal over midline and anterior left hemisphere electrode sites, and is sensitive to the degree to which prime words and target words overlap orthographically (Grainger & Holcomb, 2009). The third component is the P325 (Holcomb & Grainger, 2006) which is a positive going component that peaks around 325ms and is thought to be sensitive to repetition of *words* but not to repetition of pseudowords (Grainger & Holcomb, 2009; Holcomb & Grainger, 2006). The final ERP component is the N400 which is a widely distributed, negative going component that is maximal over centroparietal electrodes that peaks between 300-500ms and is sensitive to the amount of effort involved in interfacing between whole-word representations, both orthographic and phonological, and higher-level semantic representations (Grainger & Holcomb, 2009).

Overall, the *bimodal interactive activation model* provides explanations for cross-modal priming effects in word recognition and is supported by ERP evidence. However, the model does not explicitly use prediction as a driving mechanism for processing. The next model does use predictions generated by top-down contextual processes within a recurrent neural network to drive language comprehension.

Simple Recurrent Network Model. Altmann and Mirković (2009) recognize a critical, but often overlooked, distinction that speech unfolds as a function of time, yet written language unfolds across space. Due to this distinction, Dienes, Altmann, and Gao (1999) and Altmann (2002) extended the simple recurrent network model proposed by Elman (1990). Elman's original model processes structure in time and was only trained on linguistic information. Dienes and colleagues built upon this simpler model to allow for mapping across domains. I will first explain Elman's model and then discuss how it was extended.

Elman's (1990) model took input as a combination of sensory input and the system's prior internal state. The network then learned the time-varying structure by attempting to predict, at each time point, the upcoming word input. In this way, when given a sequence of words, the network would learn some range of words which would be most probable to follow the current word. However, humans learn language as a symbolic representation for external world events and Elman's model can only map linguistic information to predict linguistic events. In contrast, the extension (Dienes et al., 1999; Altmann, 2002) allows the model to first train on non-linguistic information, such as real-world events or scenes, and subsequently train on linguistic information. With this extension, Altmann and Mirković (2009) argue that the *simple recurrent network* embodies four principles that underlie human sentence comprehension: *Mapping across domains*, which is the principle that language processing is a mapping of sentence internal linguistic context to real-world events being described; *prediction,* which refers to real-world mapping manifesting as the ability to make predictions as to how the sentence will unfold as well as how the real-world event would unfold; *context,* which is the principle that both the sensory input and the current internal state of the comprehension system drive the predictive process; *representation across time,* which refers to making predictions across various time frames and with variable levels of abstraction.

2.3 Evaluation of speech and reading processing models

Each model presented here has supporting evidence that indicates they work, heuristically, to model human language comprehension, within their respective scopes. However, with increasing empirical evidence, it has become clear that some models are better at modeling word recognition than others. For instance, of the models presented here, only *dual-route cascading* and *bi-modal interactive models* can directly account for cross-modal interactions, and the *simple recurrent network,* can account for these interactions indirectly. Furthermore, many studies now support the idea that both linguistic and non-linguistic memory and contextual information can be used to generate top-down inferences about the incoming linguistic input. Only two models, *lossy context surprisal* and *simple recurrent network*, account for context beyond purely linguistic information and only the latter extends that context to real-world event knowledge. Although all the presented models account for contextual constraint effects, the *cohort model* assumes that these constraint effects only begin after new input has been encountered and does not allow the system to activate information about an upcoming stimulus prior to its onset. Taken together, these discrepancies indicate that a complete model of language comprehension should be cross-modal, allow for contextual effects beyond those that are purely linguistic, and be driven, at least in part, by prediction. The next section explores the roles of prediction and predictive processing in language comprehension.

3. Prediction and predictive processing

Many researchers have many different ideas about what constitutes prediction in cognition in general and more specifically within language processing (see Kuperberg & Jaeger, 2016 for additional review). In fact, many still debate whether prediction is an a-priori characteristic of language processing, or whether it can be flexibly engaged or disengaged as a function of prediction success. Older approaches viewed integration as the primary mechanism driving language comprehension – here integration is the linking of novel information to information that is already known (Ferreira & Chantavarin, 2018; Gernsbacher, 1991; Kintsch & Van Dijk, 1978). For example, Kintsch, (1988; see Ferreira & Chantavarin, 2018 for review) assumed incoming sentences are parsed into propositions, temporarily stored into working memory, which then activate long-term memory representations. Comprehenders use these long-term representations to establish both linguistic and semantic relationships which are then combined to create a rich representation of the overall discourse context. If there is no immediate link between the current sentence and previous content, then the comprehender must go back through the discourse-level representation to connect the current sentence with the ones that preceded it. In contrast, newer approaches – such as the *lossy context surprisal model* (Futrell, 2020), the *information exchange model* (Gwilliams & Davis, 2020) discussed above – lean heavily into Bayesian and information theory approaches to comprehension. Ferreira and Chantavarin (2018) suggest that there is a way to reconcile these two views of language comprehension and their approach seems to be compatible with predictive coding (Clark, 2013; Friston, 2009; 2010; Friston & Keibel, 2009; Rao & Ballard, 1999) which I will discuss later in this section. However, before I discuss how they can be reconciled, it is necessary to explore what is meant by prediction regarding language processing.

Ferreira and Chantavarin (2018) argue that many of the empirical results that have been taken as evidence of prediction could also be explained in terms of facilitation or inhibition of the *integration* of newly encountered input into a context (see also Foss, 1982). Facilitated integration is conceptualized as top-down contextual processes, facilitating word recognition when the incoming word matches in meaning with the contextual representation, and hindering word recognition when representation of the context does not match with the meaning of the incoming word. In this view, this can only occur after the onset of the new input. They therefore proposed three strict criteria for prediction effects to be considered legitimate: 1) Assure that facilitated processing is not due to facilitated integration. Highly contextually constrained words and phrases are not sufficient to support prediction as these items might be easier to process once encountered. 2) Assure that facilitated processing is not due to interlexical priming effects in sentences. Highly constraining sentences often containing words that are associatively or semantically related, and results in facilitated processing not due to sentence level prediction effect, because spreading activation, which is thought to drive associative priming, is a passive mechanism that occurs within a single level of representation (Duffy et al., 1989). 3) Assure that facilitated processing in language is not due to tasks that encourage prediction, because this may result from strategic mechanisms that may not be necessary of natural language processing.

In line with this strict interpretation, prediction was often seen as an all-or-nothing process (Kuperberg & Jaeger, 2016). However, the minimal sense of the term prediction, as it relates to language processing, is that context changes the language processing system's state prior to encountering new input which facilitates the processing of said input (Kuperberg & Jaeger, 2016). In this view of prediction as a system state, a more intuitive interpretation is that context predictively pre-activates linguistic representations to process language more efficiently.

With a more solid concept of what prediction is, the focus now turns to how it may be implemented within theoretical and computational models. Many of the models discussed in the previous sections implicitly addressed prediction by including top-down feedback. But some previously discussed models explicitly suggest Bayesian inference as a mechanism for prediction. In line with these probabilistic models, Kuperberg and Jaeger (2016) suggested a *hierarchical multi-representational generative framework* in which the goal of the comprehender is to infer – as certainly as possible – the message-level representation that the producer intends to communicate. This is accomplished by incremental cycles of propagating predictions – generated by Bayesian inference – down to successively lower levels of representation via predictive pre-activation. As bottom-up input reaches each of these levels, the Bayesian surprise or prediction error generated will be, on average, reduced due to pre-activation. Within this framework, pre-activation is weighted such that the degree to which representations are pre-activated corresponds to the relative reliability of the top-down beliefs. Therefore, predictions are used less when they are perceived to be less useful. This interpretation could be instantiated at the neural level as a form of predictive coding (Clark, 2013; Friston, 2009; 2010; Friston & Keibel, 2009; Rao & Ballard, 1999) and is compatible with the idea of prediction as *preparedness* (Ferreira & Chantavarin, 2018).

Ferreira and Chantavarin (2018) suggest that the older and newer views of predictive processing can be reconciled by thinking of prediction as a form of *preparedness* (see figure 8) – the idea that rich high-level representations place the language processing system in a ready state that improves receptivity to certain semantic, syntactic, and phonological features. This state of preparedness may take the form of pre-activation. The preparedness framework suggests that as new input is encountered, both redundant and novel information are integrated bottom-up into the preceding context to develop a rich representation of the discourse or intended message. This rich representation then generates anticipation of upcoming information which, potentially through pre-activation, prepare the system to process anticipated features. The

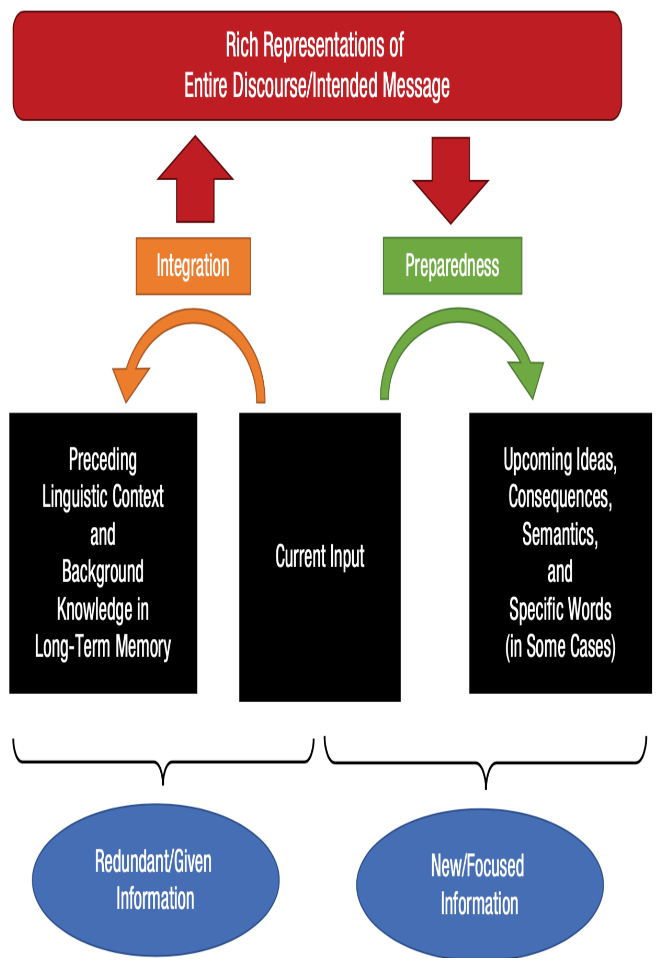

Figure 8. Preparedness model of language processing (Ferreira & Chantavarin, 2018)

preparedness framework could potentially be implemented computationally as a probabilistic model. The rich high-level representations act as the prior distribution. The posterior distributions are fed down to pre-activate lower-level representations. The prediction error generated by comparing the bottom-up input with the top-down predictions represents the cost associated with integrating the input into the preceding context to update the higher-level representations.

Taken together, the concepts of preparedness, predictive coding, and early definitions of prediction, associative priming, and integration seem to have some missing links. Early strict definitions of prediction as described above would make true prediction effects impossible by disallowing attribution to associate priming and requiring that prediction occur prior to receiving the input (Ferreira & Chantavarin, 2018).

Neely (1991; Neely & Keefe, 1989) reviewed evidence in support of pre-activation of lexical items in semantic priming studies and suggested expectancy (Posner & Snyder, 1975) as a prospective mechanism for this priming. Under the Posner-Snyder theory of expectancy, when subjects encounter a prime word – such as "cat' – they develop an expectancy for "cat"-related items and, therefore, have facilitated processing for related targets like "dog". One caveat to this is that the expectancy induced priming effects are strategically related to the task and such expectancy sets are not generated during a naming task (Grainger, 1990; Jared et al., 1990; Strain et al., 1995). However, when considering ideas of preparedness, predictive coding, and associative priming together one could propose a link between prediction at higher representation levels and word priming. I suggest that within the preparedness model of language comprehension, priming is one mechanism that drives putting together context at any level of representation. A word that we read or hear is semantically associated with other potential inputs and primes the comprehender anticipates for potentially upcoming words. As more words are added to the discourse, they are integrated into the overall context which then collectively primes for additional input. This collective priming is pruned by the comprehender's models of higher-level meaning. Predictive coding is the architecture by which these constraining hierarchical models are developed. To properly test whether a preparedness model is driven by priming and predictive coding, we must be able to investigate what happens both before and after encountering a particular input. In the next two sections, I will discuss potentially well-suited methods and avenues for testing the predictions of a preparedness computational model, as well as other models for predictive processing.

4. Electrophysiological methods

Many methodologies have been used to gain valuable insights into language processing, and more specifically, predictive processing within language. From purely behavioral techniques – which rely on modulated RTs and accuracy as a function of association strengths, and cloze probabilities as indices of predictive effects to functional neural imaging techniques - such as functional magnetic resonance imaging (fMRI) that rely on hemodynamic responses as a measure of brain activity in response to tasks – researchers have many options for approaches to studying predictive effects in language comprehension. For this chapter, I focus primarily on electrophysiology – specifically, electroencephalography (EEG) and magnetoencephalography (MEG) – techniques due to their ability to measure responses that occur prior to stimulus onset with relatively high temporal resolution compared to alternative measures – particularly in the case of EEG and MEG versus fMRI – while being minimally invasive. The milliseconds level temporal precision of electrophysiology measures such as EEG and MEG provide a critical tool for exploring the time course of language processing and the sensitivity of several ERP components (and their MEG equivalents), such as the ones discussed earlier in this review, to a variety of perceptual, orthographic, phonological, and semantic features of language provide a means for designing clever experiments such as those discussed below. With that purpose in mind, this section explores electrophysiology techniques that are highly relevant to investigating predictive processing in language.

There is a long history of electrophysiology providing insights on effects of context during language processing, particularly from the use of ERPs. The ERP component that is likely most relevant to investigating prediction is the N400, which has been briefly discussed earlier in this chapter, but I will expand on it here. As mentioned before, the N400 is a negative deflecting ERP component that is maximal between 300 and 500 ms. It is widely distributed, but maximal over centroparietal electrodes. It has been shown to be sensitive to levels of cognitive effort, orthographic and phonological features, and semantic features. However, its sensitivity to semantic violations (Kutas & Hillyard, 1980) is perhaps the most intriguing feature when discussing predictive processing.

In a recent study, Eddine and colleagues (2023) used a computational predictive coding model to tie prediction error – as defined in predictive coding theory – to various N400 findings. The model was built to approximate Bayesian inference and was able to simulate N400 effects consistent with priming effects – repetition and semantic – and with contextual effects – lexical probability and lexical violation – as well as interactions between these effects. The findings from this study provide computational evidence that the N400 is, at least in part, reflective of prediction error during language processing.

In a study that examined prediction during sentence processing, Grisoni and colleagues (2020) presented participants with either high constraint sentences or low constraint sentences that ended with an animal word or a tool word (see sentences 10-13 below) while recording

EEG. Participants showed more negative N400 amplitude to critical words in the low relative to high constraint sentences. However, the ERP recordings also showed a negative deflecting pre-stimulus component called the semantic prediction potential, or SPP, that begins approximately 150ms before the stimulus onset that is inversely correlated with the N400 – the SPP is more negative in high constraint sentences and less negative in the low constraint condition and the N400 is the reverse. These findings further support the pre-activation of contextually predictable words prior to the comprehender encountering them.

10. HC animal: “The emblem of Germany is the eagle.”
11. LC animal: “The emblem of my family is the eagle.”
12. HC tool: “The logo of the German post office is a horn.”
13. LC tool: “The logo of the company is a horn.”

In addition to N400, the N250 ERP component – which was discussed earlier in this chapter while describing the *bimodal interactive activation model* and masked priming studies – is another potential measure of prediction error. The N250 has been shown to be sensitive to repetition priming, particularly in relation to manipulations of orthographic features of words (Holcomb et al., 2002; Holcomb & Grainger, 2006, 2007). This means that the N250 has the potential to be used as a measure of prediction error for sub-lexical features.

Recently, machine learning approaches have been successful implemented to classify EEG data into two or more categories depending on the experimental conditions (Bae & Luck, 2018, 2019;Hong et al., 2020; Nadra et al., 2023; Noah et al., 2020;, Trammel et al., 2023). The brain signals could be from EEG, MEG, or fMRI – among other possibilities. To illustrate machine learning decoding approaches to electrophysiological signals

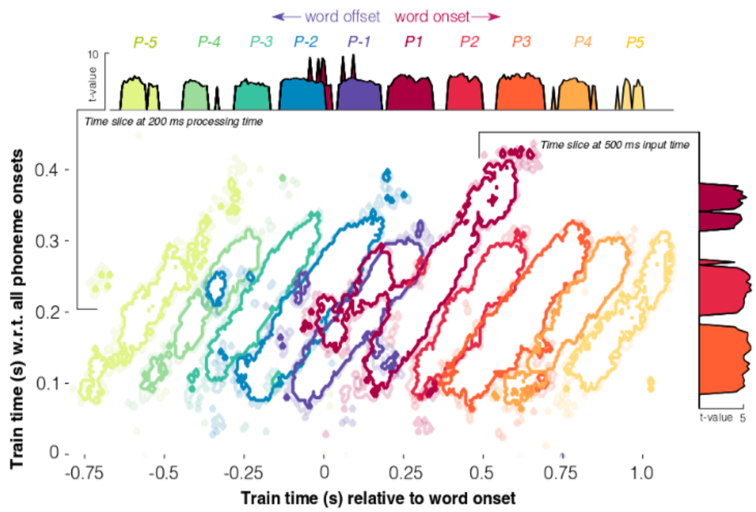

Figure 9. Results of temporal generalization analysis (Gwilliams et al., 2022)

I will discuss a study by Gwilliams and colleagues (2022). These authors recorded MEG while participants listened to stories that contained lexically ambiguous words. The MEG recordings were then analyzed using temporal generalization – a technique in which a machine learning classifier is trained at one point and then used to decode all other time points in the time frame. The results (see figure 9) indicate that speech signal is transformed at the rate of phoneme duration and that phonemes representations are sustained until lexical ambiguity is resolved.

Temporal precision is important to studying predictive processing and ERP and MEG measures are temporally precise in the order of milliseconds. MEG has an additional advantage of better spatial resolution than EEG (Nakasato et al., 1994), which means that it may also be possible to source localize predictive effects using MEG, to gain a better understanding of which cortical regions are involved in the processing of predictive effects. Taken together these methodologies provide an increasingly useful toolbox for studying predictive processing in language. I will further review the use of decoding methods to classify EEG signals and some of their potential benefits in the next section.

5. Multivariate Pattern Analysis using machine learning

As discussed in the previous section, machine-learning can be used to decode MEG or EEG data to classify different aspects of language, such as different phonemes (Gwilliams et al., 2022). A chief advantage of these decoding techniques is their multivariate nature. Machine-learning algorithms can readily accept dozens to hundreds of variables on which to perform their classifications. This allows the analysis to be done on multiple electrode sites simultaneously even when very dense electrode arrays are used. This can prove useful when performing an investigation in which there are no *a priori* assumptions for when and where an effect will take place.

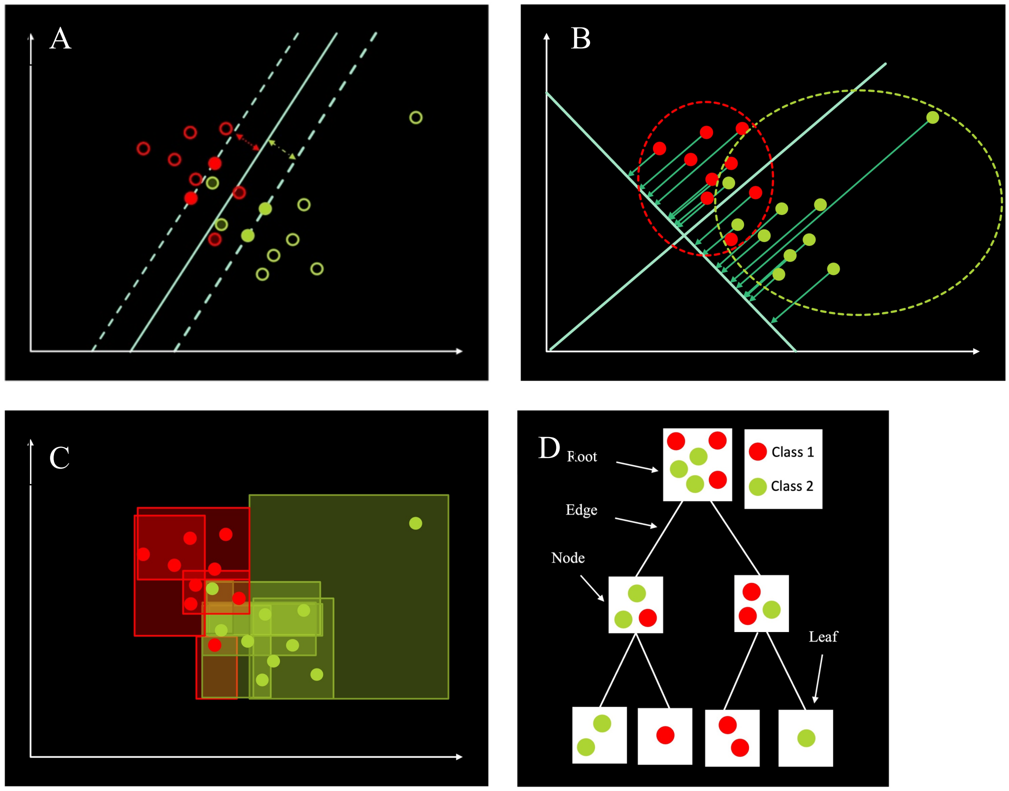

Figure 9. Examples of three different machine learning models that can be used for classification: A) Support Vector Machine (SVM), B) Linear Discriminant Analysis (LDA), and C - D) Random Forest (RF). SVM uses data points (support vectors; filled dots) closest to the decision boundary to create a maximal margin boundary to classify; some training data are allowed to cross this boundary within a set limit to improve classification generalization. LDA projects the data points onto a one-dimensional space and then chooses a boundary which maximizes the distance between the means of the classification groups. RF utilizes an ensemble of decision trees to split the data according to some pre-determined information gain function until only one class remains in a node. Test data that ends up within a given node are classified as the same group. Multiple trees classify in this fashion and vote for the overall class of the data. This process allows RF to have multiple decision boundaries.

There are a multitude of classification algorithms that can be used for decoding. Some of these – especially those based on neural networks – can be computationally very costly. I present here three varieties of machine learning classifier that readily accessible for most language research. These three models will be compared in Chapter 2. Support vector machines (SVM, figure 11A; Boser et al., 1992) are a category of binary classifiers that are based on the idea of a separating hyperplane. The basic principle of a support vector machine is to represent the data in some n-Dimensional space. The dots in the figure each represent one data point. A cost optimization function, called a kernel, is then used to generate a dividing plane through the data. The data points that define this plane are called support vectors. SVMs are robust classifiers because only those data points closest to the marginal boundary will influence the boundary. However, SVMs are naturally binary classifiers and work well for that function. They require adaptation to work for multiple classes. Linear discriminant analysis (LDA or sometimes called discriminant function analysis, figure 11B), is a generalization of Fisher's linear discriminant (Fisher, 1936) and creates linear combinations of predictors and latent variables for each function. The generated canonical functions are then used for classification of the data. Random forests (Figure 11C; Breiman, 2001) are an ensemble classifier method. The random forest consists of multiple classifiers that come together to make the final classification. In this case, the base classifier is a decision tree. At each of the nodes (Figure 11D), the tree randomly selects a subset of predictor features and uses a function to generate a split. At the next node, a new subset of features is selected for the split. This continues until all that remains is a single class in a final node called a leaf. Usually for computational performance and to avoid overfitting, there are tree depth constraints and other rules placed on the random forest. A defined number of trees are grown in this way from the training data, then the test data are put through each tree and classified according to which leaf they end in. The classifications of the majority of trees within the RF determines to which class the data is assigned.

I have now introduced some background on language processing models including more recent prediction-based models. Additionally, I have provided some overview of EEG and decoding methods that can be potentially used to investigate these models. Next, I will discuss the implications these discussions have on the present work.

6. Implications for the present work

In this review, I presented an overview of seminal models on processing language in context in both speech and reading. Next, I explored the role of prediction during language processing and discussed a model that could potentially unify predictive and integrative accounts of language processing. Finally, I discussed some of the most promising methods for future investigations of predictive processing in language comprehension. These discussions form the bases for how I aimed to study which linguistic and non-linguistic features are predictively pre-activated during language comprehension, to test assumptions of hierarchical predictive coding models.

Of the two anticipatory models described, predictive coding and preparedness, the easier of the two to test is predictive coding. The reason is that predictive coding models make a much easier to approach falsifiable claim: that predictions occur in a top-down fashion at all levels of processing. This suggests that during predictive processing of language, there should be representations in the brain of different levels of the linguistic hierarchy – such as semantic features which would be higher-level and visual features which would be lower-level – at different times during the prediction. Specifically, higher-level features should be represented earlier than are lower-level features. Using EEG decoding methods data can be labeled according to different content categories and allowing the machine learning classifier to differentiate the data according to those content labels. Combining decoding with more traditional ERP measures, provides an excellent toolbox for investigating predictive coding. In the upcoming chapters, I present how I used decoding of EEG data from a visual predictive priming paradigm to investigate which linguistic features are predictively pre-activated and the time course of those pre-activations. First, I looked at which machine learning classifier was best suited for the task. This study is presented here as it was published within NeuroImage (Trammel et al., 2023).

Next, I utilized that method to study which features are pre-activated in native English speakers. Finally, I investigated whether Spanish-English bilinguals pre-activate features in the same way.

References

Allport, D. A., Funnell, E., Longuet-Higgins, H. C., Lyons, J., & Broadbent, D. E. (1981). Components of the mental lexicon. Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 295(1077), 397–410. https://doi.org/10.1098/rstb.1981.0148

Altmann, G. T. M. (2002). Learning and development in neural networks: The importance of prior experience. Cognition, 85, 43–50.

Altmann, G., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. Cognitive Science, 33(4), 583–609.

Becker, W., & Jürgens, R. (1979). An analysis of the saccadic system by means of double step stimuli. Vision Research, 19(9), 967–983. https://doi.org/10.1016/0042-6989(79)90222-0

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. Cognition, 136, 135–149. https://doi.org/10.1016/j.cognition.2014.10.017

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. Current Biology, 28(24), 3976-3983.e5. https://doi.org/10.1016/j.cub.2018.10.042

Chauncey, K., Holcomb, P. J., & Grainger, J. (2008). Effects of stimulus font and size on masked repetition priming: An event-related potentials (ERP) investigation. Language and Cognitive Processes, 23(1), 183–200. https://doi.org/10.1080/01690960701579839

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. The Behavioral and Brain Sciences, 36(3), 181–204.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. Psychological review, 108(1), 204.

Dienes, Z., Altmann, G. T. M., & Gao, S.-J. (1999). Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. Cognitive Science, 23, 53–82

Di Liberto, G. M., Wong, D., Melnik, G. A., & de Cheveigné, A. (2019). Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. NeuroImage, 196, 237–247. https://doi.org/10.1016/j.neuroimage.2019.04.037

Donhauser, P. W., & Baillet, S. (2020). Two Distinct Neural Timescales for Predictive Speech Processing. Neuron, 105(2), 385-393.e9. https://doi.org/10.1016/j.neuron.2019.10.019

Duffy, S. A., Henderson, J. M., & Morris, R. K. (1989). Semantic facilitation of lexical access during sentence processing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 791–801.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A Dynamical Model of Saccade Generation During Reading. Psychological Review, 112(4), 777–813. https://doi.org/10.1037/0033-295X.112.4.777

Ettinger, A., Linzen, T., & Marantz, A. (2014). The role of morphology in phoneme prediction: Evidence from MEG. Brain and Language, 129, 14–23. https://doi.org/10.1016/j.bandl.2013.11.004

Ferreira, F., & Chantavarin, S. (2018). Integration and Prediction in Language Processing: A Synthesis of Old and New. Current Directions in Psychological Science, 27(6), 443–448. https://doi.org/10.1177/0963721418794491

Foss, D. J. (1982). A discourse on semantic priming. *Cognitive Psychology*, *14*(4), 590–607. https://doi.org/10.1016/0010-0285(82)90020-2

Forster, K. I. (1976). Accessing the mental lexicon. In R.J.Wales & E.Walker (Eds.), New Approaches to Language Mechanisms. (pp. 257-287). Amsterdam: North-Holland.

Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. (2018). Does predictive coding have a future? Nature Neuroscience, 21(8), 1019–1021. https://doi.org/10.1038/s41593-018-0200-7

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1521), 1211–1221. https://doi.org/10.1098/rstb.2008.0300

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. Cognitive Science, 44(3), e12814. https://doi.org/10.1111/cogs.12814

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal Predictive Codes for Spoken Words in Auditory Cortex. Current Biology, 22(7), 615–621. https://doi.org/10.1016/j.cub.2012.02.015

Gaston, P., & Marantz, A. (2018). The time course of contextual cohort effects in auditory processing of category-ambiguous words: MEG evidence for a single “clash” as noun or verb. Language, Cognition and Neuroscience, 33(4), 402–423. https://doi.org/10.1080/23273798.2017.1395466

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. Cognition, 68(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Grainger, J., & Holcomb, P. J. (2009). Watching the Word Go by: On the Time-course of Component Processes in Visual Word Recognition. *Language and Linguistics Compass*, *3*(1), 128–156.

Grainger, J., Diependaele, K., Spinelli, E., Ferrand, L., & Farioli, F. (2003). Masked Repetition and Phonological Priming Within and Across Modalities. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(6), 1256–1269. https://doi.org/10.1037/0278-7393.29.6.1256

Grisoni, L., Tomasello, R., & Pulvermüller, F. (2021). Correlated Brain Indexes of Semantic Prediction and Prediction Error: Brain Localization and Category Specificity. Cerebral Cortex, 31(3), 1553–1568. https://doi.org/10.1093/cercor/bhaa308

Gwilliams, L., & Marantz, A. (2015). Non-linear processing of a linear speech stream: The influence of morphological structure on the recognition of spoken Arabic words. Brain and Language, 147, 1–13. https://doi.org/10.1016/j.bandl.2015.04.006

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In Spoken Word Recognition, the Future Predicts the Past. The Journal of Neuroscience, 38(35), 7585–7599. https://doi.org/10.1523/JNEUROSCI.0065-18.2018

Gwilliams, L., & Davis, M. H. (2020). Extracting language content from speech sounds: An information theoretic approach. In The Auditory Cognitive Neuroscience of Speech Perception. https://hal.archives-ouvertes.fr/hal-03013496

Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2022). Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-34326-1

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In Proceedings of NAACL (Vol. 2,pp. 159–166). Pittsburgh, PA.

Holcomb, P. J., & Grainger, J. (2009). ERP effects of short interval masked associative and repetition priming. *Journal of Neurolinguistics*, *22*(3), 301–312. https://doi.org/10.1016/j.jneuroling.2008.06.004

Hong, X., Bo, K., Meyyapan, S., Tong, S., & Ding, M. (2020). Decoding Attention Control and Selection in Visual Spatial Attention [Preprint]. Neuroscience. https://doi.org/10.1101/2020.02.08.940213

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. Probabilistic Linguistics, 39–96.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. Psychological Review, 95, 163–182.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. Psychological review, 85(5), 363.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? Language, Cognition and Neuroscience, 31(1), 32–59.

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

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

Lowder, M. W., & Ferreira, F. (2016). Prediction in the Processing of Repair Disfluencies: Evidence from the Visual-World Paradigm. Journal of Experimental Psychology. Learning, Memory, and Cognition, 42(9), 1400–1416. https://doi.org/10.1037/xlm0000256

Luck, S. J. (2014). An Introduction to the Event-Related Potential Technique, second edition. MIT Press.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. Nature. 244(5417), 522-523.

Marslen-Wilson, W. D. (1985). Speech shadowing and speech comprehension. Speech Communication, 4(1), 55–73. https://doi.org/10.1016/0167-6393(85)90036-6

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken-word recognition. Cognition, 25(1), 71-102.

Marslen-Wilson, W., & Tyler, L. K. (1975). Processing structure of sentence perception. Nature, 257(5529), 784–786. https://doi.org/10.1038/257784a0

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. Cognition, 8(1), 1–71.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. Cognitive psychology, 10(1), 29-63

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. Psychological Review, 88(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

Morton, J. (1969). Interaction of Information in Word Recognition. Psychological Review, 76, 165. https://doi.org/10.1037/h0027366

Nakasato, N., Levesque, M. F., Barth, D. S., Baumgartner, C., & Rogers, R. L. (1994). Comparisons of MEG, EEG, and ECoG source localization in neocortical partial epilepsy in humans. Electroencephalography and Clinical Neurophysiology, 91(3), 171–178. https://doi.org/10.1016/0013-4694(94)90067-1

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. Journal of Experimental Psychology: General, 106(3), 226–254. https://doi.org/10.1037/0096-3445.106.3.226 4

Neely, J. H. (1991). Semantic priming effects in visual word recognition: a selective review of current findings and theories. Basic Processes in Reading: Visual Word Recognition, 264.

Neely, J. H., & Keefe, D. E. (1989). Semantic Context Effects on Visual Word Processing: A Hybrid Prospective-Retrospective Processing Theory. In G. H. Bower (Ed.), Psychology of Learning and Motivation (Vol. 24, pp. 207–248). Academic Press. https://doi.org/10.1016/S0079-7421(08)60538-1

Posner, M. I., Snyder, C. R. R., & Solso, R. L. (1975). Information processing and cognition: The Loyola symposium.

O'Regan, J. K., & Lévy-Schoen, A. (1987). Eye Movements. From Physiology to Cognition. Amsterdam.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 2(1), 79–87. https://doi.org/10.1038/4580

Rastle, K., & Brysbaert, M. (2006). Masked phonological priming effects in English: Are they real? Do they matter? Cognitive Psychology, 53(2), 97–145. https://doi.org/10.1016/j.cogpsych.2006.01.002

Rayner, K. (1975). The perceptual span and peripheral cues in reading. Cognitive Psychology, 7(1), 65–81. https://doi.org/10.1016/0010-0285(75)90005-5

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. Journal of Experimental Psychology. Human Perception and Performance, 37(2), 514–528. https://doi.org/10.1037/a0020990

Reichle, E. D., & Sheridan, H. (2015). EZ Reader: An overview of the model and two recent applications. The Oxford handbook of reading, 277-290.

Reichle, E. D., & Reingold, E. M. (2013). Neurophysiological constraints on the eye-mind link. Frontiers in Human Neuroscience, 7, 1–6.

Reilly, R. G., & Radach, R. (2006). Some empirical tests of an interactive activation model of eye movement control in reading. Cognitive Systems Research, 7(1), 34–55. https://doi.org/10.1016/j.cogsys.2005.07.006

Reingold, E. M., Reichle, E. D., Glaholt, M. G., & Sheridan, H. (2012). Direct lexical control of eye movements in reading: Evidence from a survival analysis of fixation durations. Cognitive Psychology, 65(2), 177–206. https://doi.org/10.1016/j.cogpsych.2012.03.001

Rumelhart, David E., James L. McClelland, and PDP Research Group. "Parallel distributed processing." *Foundations* 1 (1988).

Schotter, E. R., Angele, B., & Rayner, K. (2012). Parafoveal processing in reading. Attention, Perception, & Psychophysics, 74(1), 5–35. https://doi.org/10.3758/s13414-011-0219-2

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Taylor, W. L. (1953). “Cloze Procedure”: A New Tool for Measuring Readability. Journalism Quarterly, 30(4), 415–433. https://doi.org/10.1177/107769905303000401

Traxler, M. J. (2014). Trends in syntactic parsing: Anticipation, Bayesian estimation, and good-enough parsing. Trends in Cognitive Sciences, 18(11), 605–611. https://doi.org/10.1016/j.tics.2014.08.001

Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. Language and Cognitive Processes, 25(4), 533–567. https://doi.org/10.1080/01690960903310587

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. Cognition, 32(1), 25–64. https://doi.org/10.1016/0010-0277(89)90013-9

Chapter 2

Decoding semantic relatedness and prediction from EEG: A classification method comparison

Abstract

Machine-learning (ML) decoding methods have become a valuable tool for analyzing information represented in electroencephalogram (EEG) data. However, a systematic quantitative comparison of the performance of major ML classifiers for the decoding of EEG data in neuroscience studies of cognition is lacking. Using EEG data from two visual word-priming experiments examining well-established N400 effects of prediction and semantic relatedness, we compared the performance of three major ML classifiers that each use different algorithms: support vector machine (SVM), linear discriminant analysis (LDA), and random forest (RF). We separately assessed the performance of each classifier in each experiment using EEG data averaged over cross-validation blocks and using single-trial EEG data by comparing them with analyses of raw decoding accuracy, effect size, and feature importance weights. The results of these analyses demonstrated that SVM outperformed the other ML methods on all measures and in both experiments.

This chapter was published in its entirety in NeuroImage:

Trammel, T., Khodayari, N., Luck, S. J., Traxler, M. J., & Swaab, T. Y. (2023). Decoding semantic relatedness and prediction from EEG: A classification method comparison. NeuroImage, 277, 120268. https://doi.org/10.1016/j.neuroimage.2023.120268

1. Introduction

Recent developments in machine-learning (ML) classification have been successfully applied to analyze the contents of cognitive computations reflected by EEG signals. These approaches are commonly referred to as multivariate pattern analysis (MVPA) classification.

MVPA approaches are increasingly used in cognitive neuroscience studies that use the EEG method and these studies have identified potential strengths (Hebart & Baker, 2018). First, MVPA can reliably classify patterns of EEG activity at the individual subject level, which could uncover individual differences that are not detectable with univariate EEG analyses. This is important for studies of neurotypical participants, but essential in studies of neuro-divergent individuals. A second potential strength of MVPA is that they can classify the content of information carried in the EEG signal, even prior to the onset of a critical stimulus in the experiment. This is evident from studies that have decoded the contents of working memory (Bae & Luck, 2018), the specific category of objects selected by attention prior to stimulus presentation (Noah et al., 2020), the animacy and size of visually presented objects (Wang et al., 2022), the word information activated and anticipated during reading (Murphy et al., 2022), speech comprehension (e.g., Heikel et al., 2018; McMurray et al, 2022; Murphy et al., 2022), silent naming (Murphy et al., 2011), and imagined speech (Proix et al., 2022).

However, systematic comparison of ML classification methods to quantify which method is best suited to decode EEG data from neuroscience studies of cognition is lacking. Here we will examine the accuracy, reliability, and robustness of three major ML methods that use different algorithms to perform classification of EEG data from two visual word priming experiments: 1) support vector machines (SVM), 2) linear discriminant analysis (LDA), and 3) Random Forest (RF); (for details see: Boser et al., 1992 on SVM; Fisher, 1936 on LDA; and Breiman, 2001 on RF). Although deep learning models implemented using Artificial Neural Networks (ANN) may outperform the ML methods tested in this study, we nevertheless chose to assess these three ML methods. This was done because SVM, LDA and RF are capable of classifying EEG data from cognitive neuroscience studies that do not include vast amounts of data (which is needed for ANNs), are computationally less costly, and can be implemented more easily.

We compared the performance of the three classification methods using EEG data from two visual word priming experiments conducted in our laboratories (Experiment 1, Brothers et al., 2016; Experiment 2, Kappenman et al., 2021). In these studies, we examined the effects of prediction and semantic relatedness on a well-established event-related potential (ERP), the N400, which reflects facilitation of word processing in supportive contexts (e.g., Bentin et al., 1993; Chwilla et al., 1995; Holcomb; 1988; Delaney-Bush et al., 2019; Lau et al., 2013; see Swaab et al., 2012). In both studies, participants read related (e.g., circus – clown) and unrelated (e.g., napkin – clown) word pairs. In the first experiment, the participants' task was to predict the target word after having read the prime. In the second experiment their task was to judge whether the words of each pair were semantically related after the target word was presented.

Although ML methods provide a powerful tool to uncover new data-driven features that could be highly informative to model targeted real-world applications, the aim of the present study was to specifically test the performance of ML approaches in cognitive neuroscience EEG studies, i.e., whether the EEG data can be reliably classified according to the conditions in the two visual word priming experiments. To test this, we compared the SVM, LDA and RF methods on 1) their raw decoding accuracy (percentage correctly classified), to assess which of the methods most consistently classified the EEG data above chance in accordance with the conditions in the experiments 2) their effect sizes, to assess whether the classifiers differed in their ability to detect differences between conditions, and 3) their feature importance weights, to examine the weighting of each of the electrode sites included in the two EEG studies in the classification decisions. We did this for the entire decoding epoch, and further assessed whether those electrodes that show maximal differences between the EEG signals as a function of the experimental manipulations would also be weighted most heavily in the ML classifications. This was done for the N400 effect, which is maximal in the 300 – 500 ms epoch for electrodes over central and parietal sites, and in the 600 – 800 ms epoch where the ERP effect shifted to more anterior sites.

Since the two priming experiments used different tasks, different numbers of participants, different numbers of trials, different electrode configurations and different recording electrodes (active vs. passive), we were able to assess how robust these ML methods were in their decoding of the EEG data. In addition, we not only examined classification accuracy for averaged trials, as is more common, but we also assessed their reliability for classification of single-trial EEG data. To foreshadow the results, SVM generally outperformed classification of the EEG results on measures of raw decoding accuracy, effect size and feature weighting in both visual word priming studies compared to the LDA and the RF methods.

2. Methods

The stimuli and EEG data from Experiment 1, and all machine learning scripts are available on OSF: https://doi.org/10.17605/OSF.IO/V8ACD. The experiment control scripts, data, and univariate analysis scripts for Experiment 2 can be downloaded at https://doi.org/10.18115/D5JW4R. See Kappenman et al. (2021) for additional methods details regarding Experiment 2.

2.1 ERP Methods

2.1.1 Participants

Participants (N=80; 45 in Experiment 1; age range 18-30; 59 female) first provided informed consent. All were native English speakers with normal or corrected-to-normal vision, and all but two participants were right-handed. Both experiments were approved by the UC Davis Institutional Review Board and subjects signed informed consent forms before participating.

2.1.2 Materials

In both experiments, participants read sequences of word pairs for which the second word (the target) could be related or unrelated in meaning to the first word (the prime; see Figure 1).

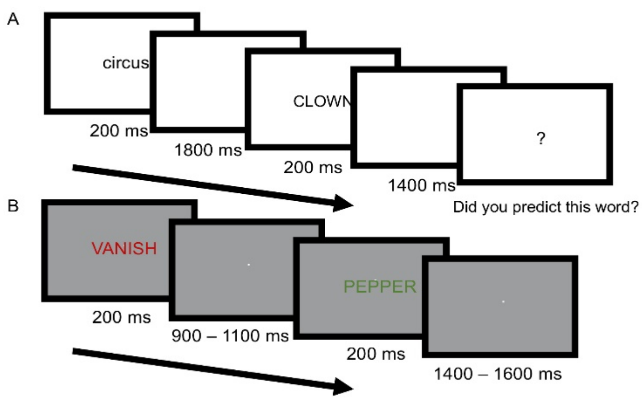

Figure 1. Schematic depiction of a trial in Experiments 1 (A) and 2 (B). See methods for further details.

Experiment 1 consisted of 320 trials in the related condition and 160 in the unrelated condition; Experiment 2 had 60 trials in each condition. In both experiments, the same target words appeared in both related and unrelated conditions, so that they could serve as their own controls. However, within subjects the same target word was never repeated. The mean forward association strength – as assessed by the University of Florida Free Association Norms (Nelson et al., 1998) – was 0.5 in Experiment 1, so that two equally likely target words were predictable, and 0.8 in Experiment 2, to promote accuracy in the semantic-relatedness task (see section 2.1.3). The trials in the related condition in Experiment 1 were subdivided according to each participant's prediction accuracy responses (see procedure below), yielding three conditions: predicted-related, unpredicted-related, and unpredicted-unrelated.

2.1.3 Procedure

In Experiment 1, participants' task was to silently generate a prediction of the target word immediately after presentation of the prime word, and to indicate via a button press after presentation of the target word, whether it matched their prediction. Assessment of prediction accuracy showed that participants on average accurately predicted 50% of the target words in the related condition, which demonstrated that they had adhered to the task. In Experiment 2, participants pressed one of two buttons after each target word to indicate whether it was related or unrelated in meaning to the preceding prime word. Prime and target words were presented for 200 ms in both experiments, with a stimulus onset asynchrony (SOA) of 2000 ms in Experiment 1, and of 1100 – 1300 ms in Experiment 2. In Experiment 1, after the target word was presented, a 1400 ms delay followed and then a question mark appeared in the center of the screen to indicate that participants could make their prediction accuracy response. The experiment then proceeded to the next trial after 1500 ms. In Experiment 2, there was a uniformly distributed 1400-1600 ms inter-trial interval following each target word, during which the participants were asked to make their semantic-relatedness response.

2.1.4 EEG Recording and Processing

All analyses used previously recorded and pre-processed data. We performed traditional univariate ERP analyses on the results both experiments. We will summarize key information about the methods here. For further detail, please see the Supplementary Materials section on ERP Methods.

For Experiment 1, EEG was recorded from 29 tin electrodes in an elastic cap with a sampling rate of 250 Hz (electrodes: FP1, FP2, F3, F4, F7, F8, FC1, FC2, FC3, FC4, FC5, FC6,

C3, C4, CP1, CP2, CP5, CP6, P3, P4, T3, T4, T5, T6, O1, O2, AFz, Fz, Cz, Pz and POz). For Experiment 2, EEG was recorded from 30 active electrodes with a sampling rate of 1024 Hz (electrodes: FP1, F3, F7, FC3, C3, C5, P3, P7, P9, PO7, PO3, O1, Oz, Pz, CPz, FP2, Fz, F4, F8, FC4, FCz, Cz, C4, C6, P4, P8, P10, PO8, PO4, and O2). In both experiments, 20 electrodes were placed according to the 10-20 system (Jasper, 1959). Additional electrodes were included at different scalp locations for experiments 1 and 2. All electrode locations are shown in Figure 5 c and d, respectively.

In both experiments, offline pre-processing analyses were conducted in MATLAB using EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014). To be consistent with Experiment 1, the EEG from Experiment 2 was down sampled to 250 Hz. EEG was re-referenced offline to the average of the right and left mastoids in Experiment 1 and to the average of P9 and P10 electrodes in Experiment 2. In both experiments, a Butterworth bandpass filter was applied with half-amplitude cutoff of 0.05-30 Hz. In Experiment 1, participants with at least 80 trials per condition after artifact rejection were included from the analyses (range: 80 - 231, mean: 152.7). In Experiment 2, participants were included in our analysis when they had a minimum of 30 trials per condition (range: 35 - 60, mean: 50.4) (see supplementary tables S1 and S2 for details on artifact rejection by participant). Artifact rejection led to removal of 9 participants in Experiment 1 (45 remained) and 5 participants in Experiment 2 (35 remained).

2.2 Decoding analyses methods

2.2.1 Classification implementation

All three classifiers were implemented within MATLAB. The linear SVM algorithm was implemented using the *fitcsvm()* function using 'one vs. one' classification. The LDA algorithm was implemented using the *fitcdiscr()* function. The RF algorithm was implemented using the *treebagger()* function. To prevent excessive processing time, the RF was limited to only 10 trees per iteration. This allowed the other parameters – iterations and cross-validation folds – to be held constant across the three machine-learning methods. Other than the function calls for the classifier algorithm, the script for each implementation was identical to ensure that no differences in pre-processing were introduced. For Experiment 1, each method classified 400 time points in the -200-1396 ms target stimulus-locked interval, and in Experiment 2 this was done for 250 time points over the -200-796 ms epoch locked to the target stimulus onset.[1]

Classification was performed for all trials that were included in the ERP analyses after artifact correction for ocular artifacts (using independent component analysis) and rejection of trials with other subject-generated artifacts (e.g., movement). For Experiment 1, classification accuracy of the three ML methods was compared for: 1) accurately predicted-related vs. unpredicted-related to isolate prediction effects, 2) unpredicted-related vs. unpredicted-unrelated to isolate effects of semantic relatedness, and 3) accurately predicted-related vs. unpredicted-unrelated which includes both effects along with any interactions between them. For Experiment 2, method comparison was done for the classification of correctly judged unrelated versus related target words. All classifiers were trained and tested on each individual participants' data. At each time point, the single-sample values from all electrode sites (29 for Experiment 1; 28 for Experiment 2) were used in the ML procedures. Machine learning was performed over 128 iterations – with each iteration repeating the entire process including randomized assignment to cross-validation blocks – using 10-fold cross-validation. Iterating the process 128 times ensured

[1] This was done on the data available online via the ERP Core data set, for which we did not adjust the measured epoch.

sufficient re-sampling so that all trials were included in the random block assignments and spurious results were eliminated. For each iteration, trials were separated into 10 blocks (9 training; 1 testing) and each trial could alternately serve as a training and a test data point. To keep an equal number of trials across conditions, each iteration used a random sampling to reduce the number of trials in one condition to equal the number of trials in the other condition.[2]

2.2.2 Average vs. single trial classification

For the averaged-trial method, we implemented a procedure recommended by Grootswagers et al (2017) which was adapted from the scripts used by Bae and Luck (2019). For example, if a participant's data included 100 trials per condition – 200 total – then those trials were randomly divided into 10 equal sets of 20 trials – 10 per condition. An averaged ERP was calculated for each condition within each of the 10 sets of 20 trials, and this was used by the ML algorithms to classify the data. For the single-trial method, this averaging step was omitted. To assess decoding accuracy, we calculated the percentage of EEG data that was classified above chance for each classifier at each time point across all iterations and cross-validation blocks for each participant. The final accuracy was then averaged across participants. Because of the binary classifications, chance performance was at 50%.

2.2.3 Raw Decoding Accuracy for each individual Classification Method

For each of the three ML methods we used cluster-based permutations to test if their performance was significantly above chance (Bae & Luck, 2019; Maris & Oostenveld, 2007). This generated a null distribution using 10080 random permutations of the existing data labels. This number of permutations is approximate to a .01 alpha value for non-parametric testing

[2] For example, if the predicted-related condition had 101 trials and the unpredicted-related condition had 88 trials, then 88 trials of the 101 within the predicted-related condition would be randomly selected for each iteration.

(Marozzi, 2004). A one-sample *t* test was then performed to compare the individual participants' decoding accuracy at a given time point to chance (separately for each permutated dataset and the actual data). Contiguous sets of timepoints that were significant without correction were considered clusters, and the mass of a given cluster was computed as the sum of the *t* values for that cluster. If the mass of at least one cluster in the actual data was greater than the 95th percentile of the cluster masses in the permutated data, then the decoding accuracy was considered significantly greater-than-chance. This method avoids multiple comparison errors while accounting for autocorrelation in the EEG data. The significance testing was done between 0 – 1396 ms after target onset for Experiment 1 and for 0 – 796 ms for Experiment 2 (because this was the epoch available from the Kappenman et al., 2021 dataset). The *t* tests comparing the decoding accuracy at each time point to chance were one-tailed because performance significantly lower than chance-level has no meaningful interpretation in these analyses.

2.2.4 Comparison of Raw Decoding Accuracy between Classification Methods

To formally test differences between the methods, we used *F* tests from a three-level ANOVA (including decoding results from SVM, LDA, and RF) to perform cluster analyses. This ANOVA including all three decoding methods for direct comparison follows the same steps as described for the performance tests of each of the three individual methods in the previous section. We again used the 10080 permutations, but now a null distribution of the *F* mass was generated and compared to the 95th percentile of the null distribution to assess significance. If significant, this initial analysis was followed by pairwise comparisons of each pair of methods using two-tailed *t* tests. In the pairwise comparisons, a *t* mass larger than the 97.5th percentile or less than the 2.5th percentile of the null distribution indicated a significant difference. These direct comparisons were performed over the same time periods as the decoding-versus-chance analyses (Experiment 1: 0 – 1396 ms, 350 time points; Experiment 2: 0 – 796 ms, 200 time points).

2.2.5 Effect Sizes

Effect sizes were computed using *Cohen's* d_z for each condition in both experiments for all three ML methods, relative to chance level. This quantifies the ability of each classifier to produce statistically significant results while accounting for variation in decoding accuracy across participants as well as the mean. *Cohen's* d_z was computed by subtracting the chance value from the mean percent correct and then dividing by the SD. The classification approach that produced the largest *Cohen's* d_z had the greatest statistical power.

2.2.6 Feature Importance Maps

Calculation of the contribution of each of the electrodes to the classification decision of each of the ML methods was performed. For SVM and LDA, the magnitudes of weight coefficients applied to each electrode site relative to one another were used as a reasonable proxy for how important each site was to the classification. Because SVM and LDA are linear classifiers, the extracted coefficients were multiplied by the covariance of the µV values of the electrode sites that were classified using the training data at each time point to obtain the feature weights (Grootswagers et al., 2017, Haufe et al., 2014). For RF, feature importance measures were directly extracted from the model using the oobPermutedPredictorImportance() function.

Each classifier's feature weights were standardized by separately computing the mean of the absolute weights across all the electrodes for each time point, subtracting that mean from the absolute weight values of each electrode and dividing the resulting weight values by the standard deviation for all the electrode site weights. We used absolute values for the weights because negative and positive weights are equally important to the classifier. Important to note here is that the directionality (+/-) of the feature weights does not necessarily correspond to positive or negative μV values. The standardized values were used to evaluate the relative importance of each electrode site as the number of standard deviations from the mean. A zero value on this measure indicates average importance, while positive values indicate greater than average performance and negative values correspond to less than average importance. Feature importance measures were then plotted for a total epoch of 1600 ms in experiment 1 and for a total epoch of 1200 ms for experiment 2 (See Figure 5a, b). The distribution of the contribution of the feature weights for each of the electrodes is also plotted for two epochs, the N400 epoch (300 – 500 ms) and a later epoch (600 – 800 ms), and compared to the distribution of the ERP effects in those epochs (See Figure 5c and d)

3. Results

The ANOVA for the ERP results of Experiment 1 showed significant N400 effects of predictability and priming with a typical topographic distribution (p's<.0000, for details see Supplementary Materials). Significant effects of priming were reported for Experiment 2 in Kappenman et al., (2021). In the remainder of this results section, we will report on the findings of the decoding analyses.

3.1 Results of the decoding accuracy analyses. Figures 2 and 3 show decoding accuracy at each time point for each decoding method, for averaged and single-trial data, respectively.

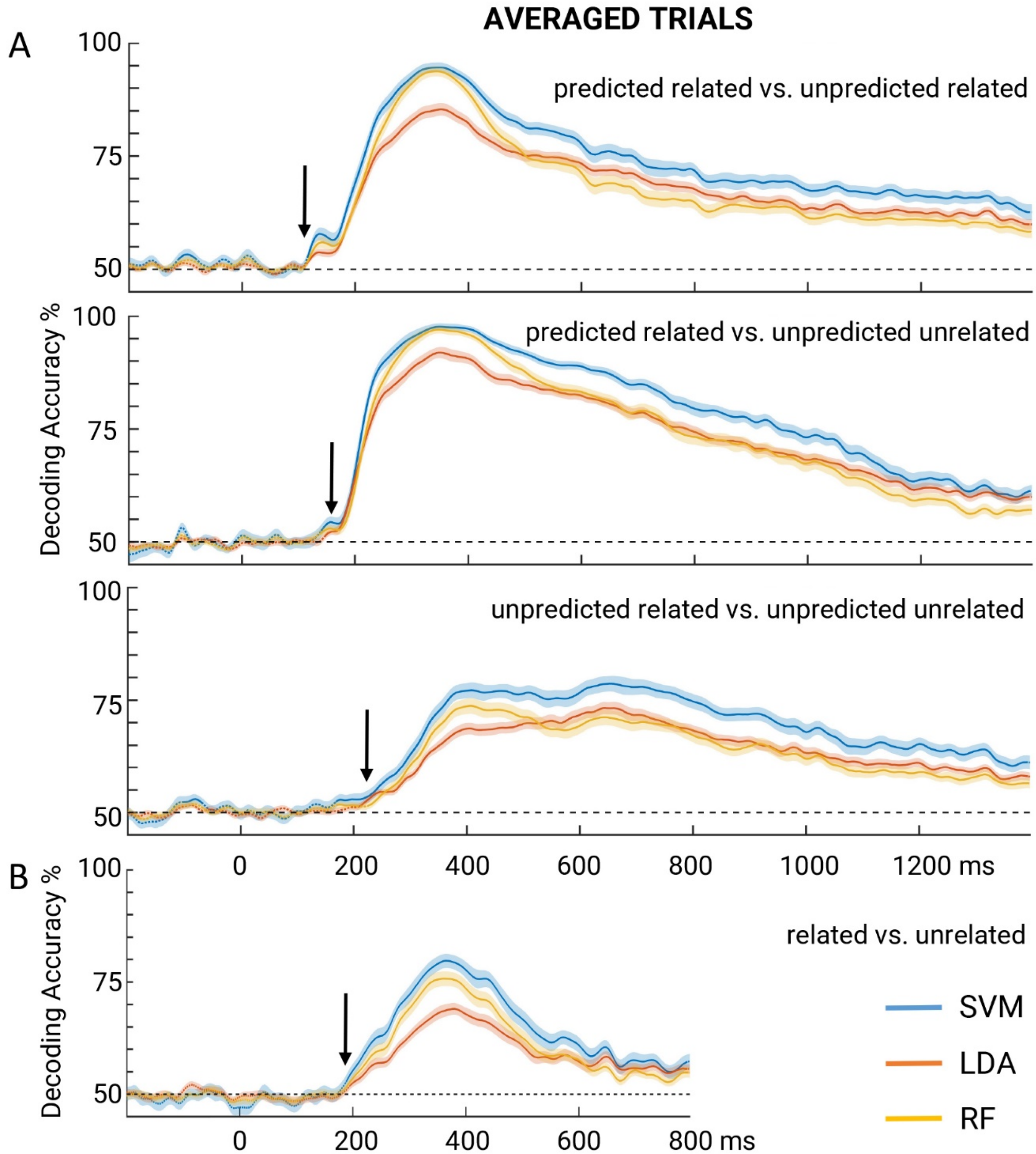

Figure 2. Decoding accuracy (percentage of averaged trials correctly classified) for averaged trial data of the three ML methods. Panel A shows the decoding accuracy for the conditions in Experiment 1 (Brothers et al., 2016), panel B for the conditions in Experiment 2. In Experiment 2, the decoding epochs were shorter than in Experiment 1, to remain consistent with Kappenman et al. (2021). Shading indicates the standard error of decoding accuracy for each method. Decoding accuracy is above chance for all three methods for the entire epoch to the right of the arrows. See Table 1 for exact onset of above chance decoding accuracy of each of the methods.

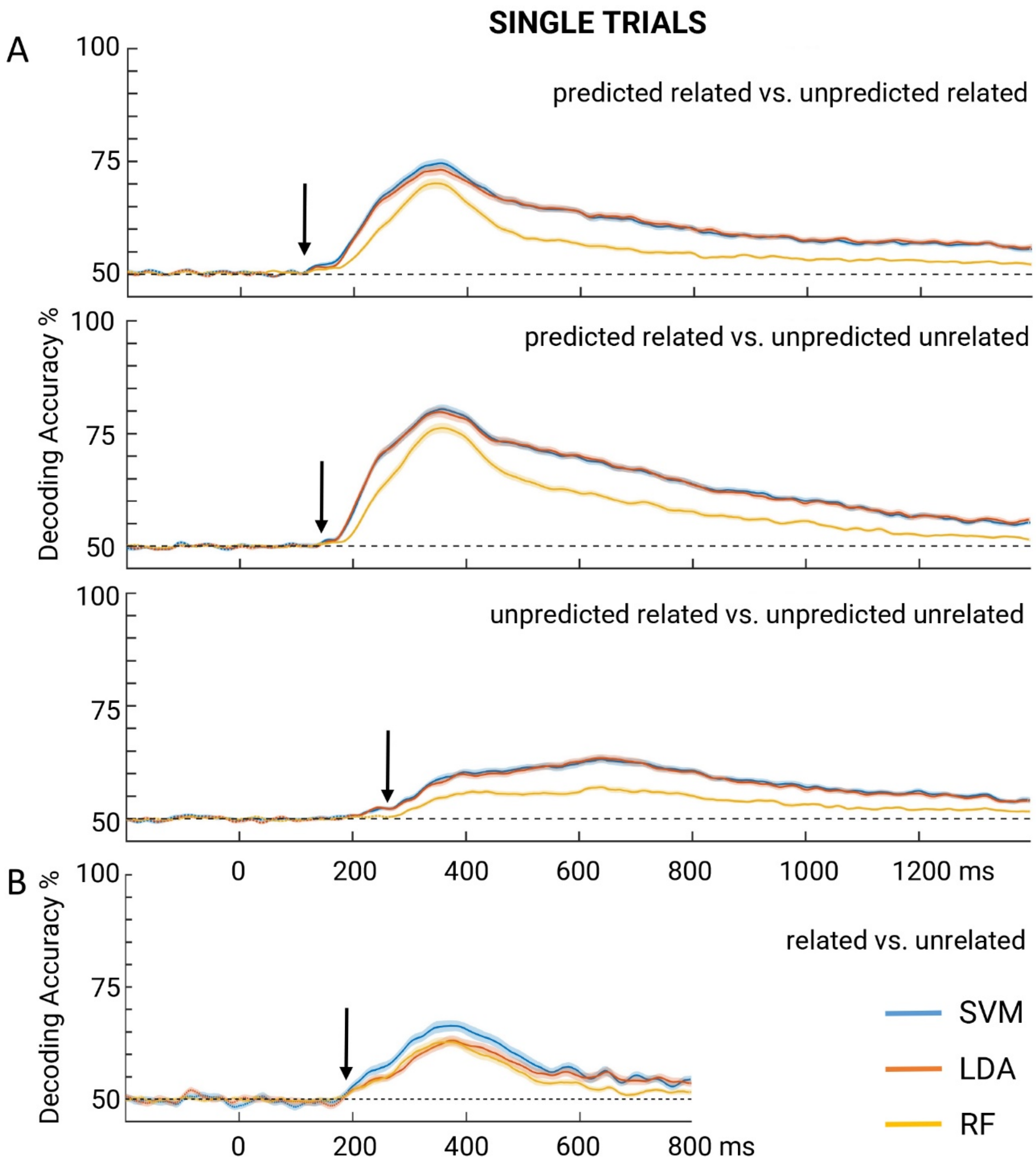

Figure 3. Decoding accuracy (percentage of trials correctly classified) for single trial data of the three ML methods. For all other information about this Figure, see the legend of Figure 2.

The t tests of the decoding accuracy showed that all three ML methods significantly classified above chance (50%) the EEG signal in all conditions for most of the epoch after the target word was presented for both averaged and single trial data. These results are reported in

Table 1 for averaged and single trial analyses under (a) decoding accuracy vs. chance. This table also reports the results of the ANOVA directly comparing the three ML methods (under (b) ANOVA Based Test Method), and the follow up *t* tests comparing SVM to LDA, SVM to RF and RF to LDA (under (c) Pairwise *t* Test method).

Experiment	Conditions	Decoding Method	Significant Cluster Epochs (ms)							
			Averaged Trials				Single Trials			
			a) Decoding Accuracy v. Chance-level	**b)** ANOVA-based Test	**c)** Pairwise *t* Test: LDA	**c)** Pairwise *t* Test: RF	**a)** Decoding Accuracy v. Chance-level	**b)** ANOVA-based Test	**c)** Pairwise *t* Test: LDA	**c)** Pairwise *t* Test: RF
Brothers et al. (2016)	PR v. UR	SVM	116 - 1396	224 - 1396	224 - 1396[#]	224 - 1396[#]	116 - 1396	172 – 1396	220 - 408[#]	172 - 1396[#]
		LDA	116 - 1396			228 - 460[X]	124 - 1396			172 - 1396+
		RF	116 - 1396				44 - 76 116 - 1396			
	PR v. UU	SVM	148 - 1396	220 - 1396	220 - 1140[#]	220 - 1396[#]	144 - 1396	180 – 1396	none	180 - 1396[#]
		LDA	148 - 1396			224 - 512[x] 1248 - 1396[+]	152 - 1396			180 - 1396[+]
		RF	136 - 1396				140 - 1396			
	UR v. UU	SVM	208 - 1396	324 - 1396	324 - 1396[#]	324 - 1396[#]	212 - 1396	236 – 1396	none	236 - 1396[#]
		LDA	212 - 1396			324 - 488[x]	196 - 1396			236 - 1396[+]
		RF	176 - 1396				268 - 1396			
Kappenman et al. (2021)	R v. U	SVM	188 - 796	240 - 476	240 - 476[#]	240 - 476[#]	192-796	264 – 400 628 - 716	264 - 400[#]	264 - 400[#] 628 - 756[#]
		LDA	196 - 796			240 - 476[x]	188-796			628 - 792[x]
		RF	188 - 796				192-796			

Table 1: Onset and offset of epochs where ML methods significantly differed from (a) chance decoding accuracy (percentage of trials or averaged trials correctly classified), (b) each other in an ANOVA (b), and in (c) follow up pairwise *t* tests between each method, for each of the cluster-based permutation tests, for Averaged Trials and Single Trials. PR = Predicted-related, UR = Unpredicted-Related, UU = Unpredicted-Unrelated, R= Related, U= Unrelated. The following symbols indicate significantly higher decoding accuracy of one method over another method in the pairwise *t* tests: # = SVM; + = LDA; x = RF

The results of the follow-up t tests were done for the epochs for which the ANOVA

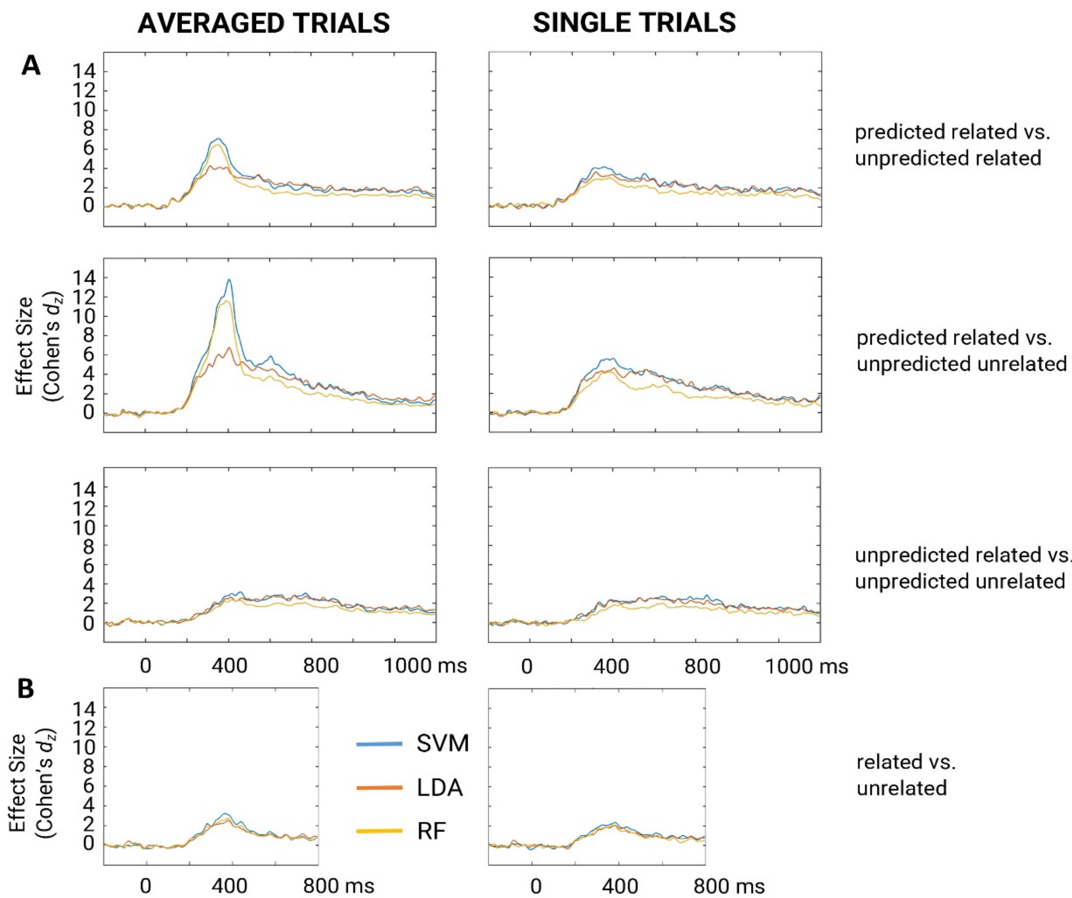

Figure 4. Method decoding accuracy effect sizes (Cohen's d_z) against chance-level, for each condition and at each time point for all three ML methods, for Experiment 1 (Panel A) and Experiment 2 (Panel B). In Experiment 2, the decoding epochs were shorter than in Experiment 1, to remain consistent with the original experiment. Effects sizes are shown with target word onset at 0 ms.

showed significant differences between ML methods for averaged and single trial data. For averaged trial classification, decoding accuracy was significantly greater for SVM than LDA in each condition from both datasets. Additionally, SVM had significantly higher decoding accuracy than RF for most of the epochs and conditions. The pairwise analyses of the single-trial data of Brothers et al. (2016) did not show significant differences in decoding accuracy between SVM and LDA, except for a window in the predicted-related vs. unpredicted-related condition (220 – 408 ms). In contrast, RF consistently had lower decoding accuracy than the other two classifiers for the single-trial data. For Kappenman et al., (2021), SVM outperformed the other classifiers early in the epoch (264 – 400 ms) while RF underperformed relative to the others late in the epoch (628 – 716 ms).

3.2 Results for Effects Size (Cohen's d_z)

Average in Effect Sizes vs. Chance (300 – 500 ms)				
Experiment	**Conditions**	**Method**	*Cohen's d_z*	
			Averaged-Trials	**Single-Trials**
Brothers et al., 2016	PR v. UR	SVM	**5.442**	**4.009**
		LDA	3.847	3.364
		RF	4.501	3.020
	PR v. UU	SVM	**10.092**	**5.717**
		LDA	5.858	4.518
		RF	8.553	4.118
	UR v. UU	SVM	**2.829**	**2.407**
		LDA	2.519	2.260
		RF	2.287	1.810
Kappenman et al., 2021	R v. U	SVM	**2.722**	**2.282**
		LDA	2.202	1.921
		RF	2.383	1.983

Table 2. Averaged method decoding accuracy effect sizes (Cohen's d_z) against chance-level, for each condition within the 300 – 500 ms time window for all three ML methods. The greatest effect size for each comparison of conditions is bolded. PR=Predicted Related; UR=Unpredicted Related; UU=Unpredicted Related, R=Related, U=Unrelated.

Figure 4 shows the results of the analyses of effect size for each of the ML methods for each of the conditions in both experiments. Each of the methods had large effect sizes (>1) for most conditions. However, SVM showed the greatest effect sizes in *all* conditions indicating that SVM will provide the greatest statistical power among the three methods (for effect sizes in univariate ERP analyses, see table S3 in the Supplementary Materials). We also calculated effect sizes for each of the ML models averaged over the time points in the 300 – 500 ms epoch for which the N400 is maximal (see Table 2).

3.3 Feature importance

Feature importance maps for both experiments across the whole target epoch and for selected times points are shown in Figure 5. For the averaged-trial data, all three classifiers placed more relative weight on centro-parietal electrode sites during the epoch typically associated with the N400 ERP component (300 – 500 ms). This pattern is in line with the typical topographical distribution of

the N400 (see Swaab et al., 2012). RF placed more weight on occipital and temporal sites than did SVM or LDA. After 600 ms, all three classifiers shifted more relative weight to frontal electrodes, in line with expectations. For the single-trial data, all three classifiers showed similar weight patterns for all conditions in Experiment 1. In contrast, Experiment 2 showed that RF weighted features differently from the other classifiers during the 600 – 800 ms time window. This coincides with a substantial decrease in the RF decoding accuracy relative to SVM and LDA reported earlier.

In summary, feature importance from all three classifiers matched well to both expectations and to the ERP scalp topographies in Experiment 1. In Experiment 2, RF weights did not match well to the ERP scalp topographies during the 600 – 800 ms time frame, suggesting that improper weighting may have resulted in the drop off in decoding accuracy.

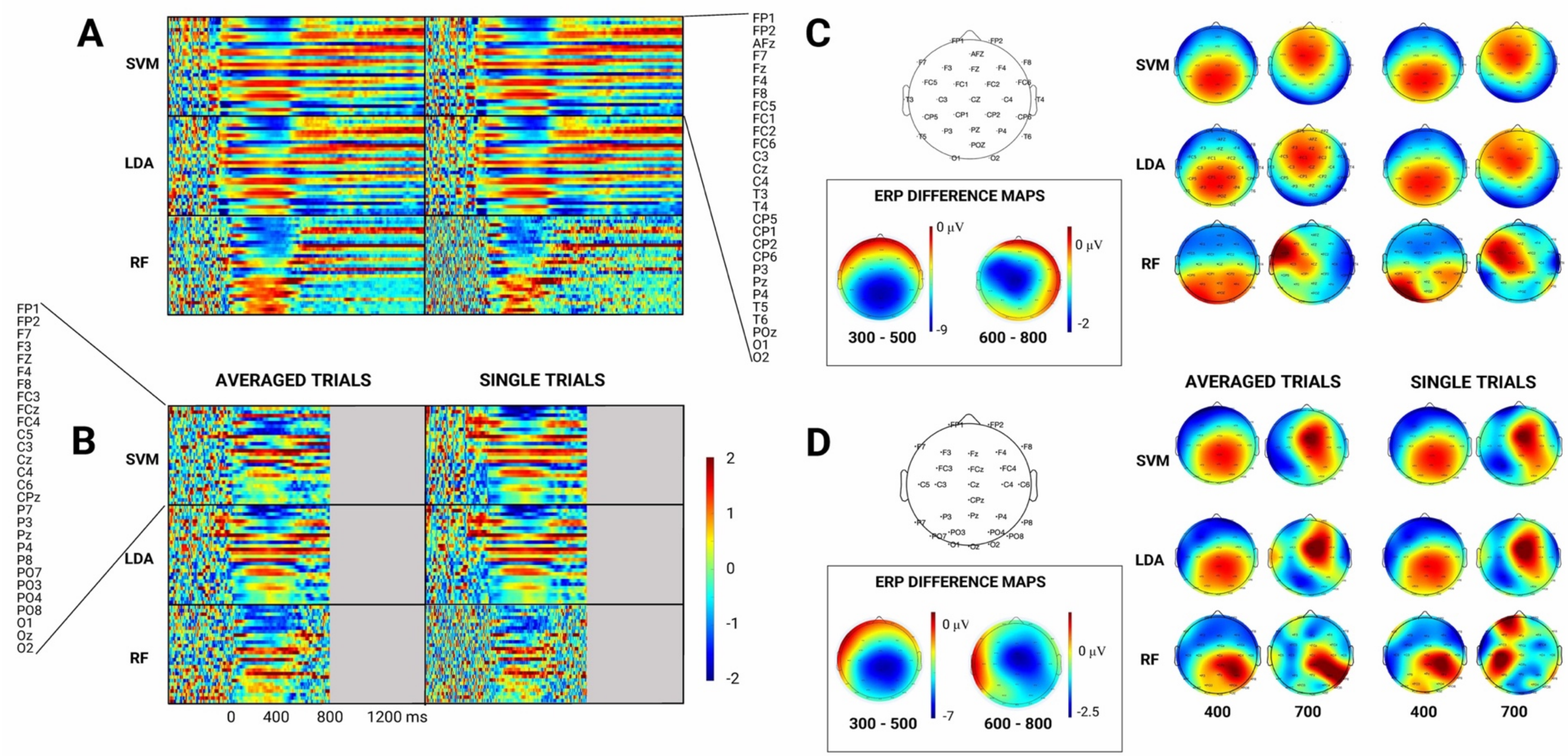

Figure 5. Feature importance maps for all three ML methods. The maps show the normalized weighting of each electrode site in the decoding of the target words in Experiment 1 (panel A) for the predicted related vs. unpredicted related condition, and in Experiment 2 (panel B) for the target words in the unrelated vs. related condition. Six boxes of feature maps are shown for each experiment, for SVM, LDA and RF, respectively, on the left for the averaged trials, on the right for the single trails. Each box shows data for all electrode sites over the entire epoch from anterior to posterior sites, along the vertical axis. For Panel A, the electrodes included are indicated on the right, and for panel B on the left. Values are in standard deviations from mean weights for a given time point. While the figure legend is capped at +/-2, actual weights can exceed these limits. Panels C (predicted related vs. unpredicted related) and D (unrelated vs. related) show feature weights from select time points (400 ms and 700 ms) along with difference topography maps of the ERPs for the conditions being decoded (panel C: unpredicted-related minus predicted-related averaged separately over 300 – 500 ms and 600 – 1000 ms; panel D: unrelated minus related averaged separately over 300 – 500 ms and 600 – 800 ms). Values for ERP difference maps are in µV. Feature maps for the other conditions of Experiment 1 are shown in the Supplementary Materials (Figure S3).

4. Discussion

The goal of this study was to identify which of three well-established ML classifiers, SVM, LDA or RF, was the best performing classifier when decoding EEG data from two visual two-word priming experiments. We assessed the performance of these three ML methods in classifying the EEG responses to the second word of each pair, the target, in each of the conditions in both experiments. In Experiment 1, we examined ML classification of the EEG to in three conditions: 1) related target words that were predicted by the participant 2) related target words not predicted by the participant and 3) unrelated target words. In experiment 2 this was done for related and unrelated target words. Classification performance was assessed for both averaged and single EEG trial classification by establishing decoding accuracy, effect sizes and feature importance measures for each of the classifiers. This was done in epochs of 400 ms before the presentation of the target words and, after the onset of the target words, in an epoch of 1200 ms for Experiment 1 and of 800 ms for Experiment 2. To assess whether the central and parietal electrode sites for which the N400 prediction and priming effects are most prominent were also most heavily weighted in the classifier decisions of each of the ML methods, we further compared the topographic distribution of the N400 effect to the topographic distribution of the weights assigned by the ML classifiers to the electrode sites in this epoch. We compared this to a later epoch between 600 - 800 ms, where the ERP effects had shifted to more anterior electrode sites.

Overall, SVM was the best performing classifier on all three of our performance measures in both experiments (with one exception, discussed below): relative to LDA and RF, SVM decoded the EEG data to the target words in the different conditions with higher decoding accuracy and larger effect sizes. Moreover, its classification decisions for the time points in the N400 epoch were weighted more heavily to cento-parietal electrode sites, and shifted more anteriorly for the 600-800 ms epoch, consistent with the ERP findings. Because the experiments differed in task, the number of participants, the number of trials per condition, and number, configuration, and kind of electrodes (active vs. passive), we conclude that this is a robust finding.

There was one instance where the LDA and SVM classifiers showed similar performance, namely in the single trial analyses for Experiment 1. This might suggest that the decoding method that performs best when decoding averaged data is not always be the best method for decoding single-trial data (or *vice versa*), as SVM appears to have a greater drop-off in decoding accuracy than does LDA in single trial decoding when compared to decoding of averaged trials. Alternatively, it could be the case in the present study that LDA's decoding accuracy is near its performance ceiling when decoding single trial data, and only receives a mild boost in decoding accuracy from increased signal-to-noise ratio in averaged data. In contrast, SVM appears to get a substantial boost in decoding accuracy when classifying averaged trials relative to decoding single trials. We suggest that this may be related to the different ways in which SVM and LDA classify the data. Whereas SVM classifies based on data points that are near the boundary between two classes, LDA considers the entire distribution of the data points. In addition, the LDA decoding accuracy is much reduced relative to SVM for the single trial classification in Experiment 2. As mentioned before, experiment 2 included considerably less trials per condition than Experiment 1 (i.e., 40 vs. 150). This may indicate that LDA requires greater number of individual training trials than SVM does to optimize its decoding accuracy. With a sufficiently large training data set, LDA performs as well as SVM.

All three ML methods showed overlap between the centro-parietal electrode sites that were most heavily weighted in the classification of the N400 effects in the 300 - 500 ms epoch and, consistent with the ERP data, more anterior sites for the 600 - 800 ms epoch. However, there was one exception (See Figure 5d) in the classification of the single trial data of Experiment 2. In contrast to SVM and LDA, for which the weight maps were very similar to the topographic maps of the ERP results, RF placed greater weight on electrodes that were more anterior and left lateralized than the distribution of the maximal ERP differences in this epoch. This may be due to the smaller number of trials in Experiment 2. It also important to note that the less precise feature weighting for the RF coincides with a rapid drop-off in decoding accuracy for the RF classifier of the single trials in the post 500 ms epoch.

While our results demonstrate that SVM was the best classifier of EEG data in this study, it is important to note that both experiments required binary classifications. SVM classifiers are binary classifiers best suited for categorizing two classes – separating two classes from one another or separating out one class from many. SVM can be modified to handle multiple classes (Bae & Luck, 2018), but LDA and RF can handle multiple classes without modification. Thus, future studies are needed to examine the performance of these three ML methods when more than two variables are manipulated in an experiment.

It is also important to emphasize here that the goal of the present study was to compare the ML classification methods performance in cognitive neuroscience studies using EEG data, and this was reflected in the way we assessed their performance. We recognize that ML classifiers can be a powerful tool to uncover spatio-temporal features of EEG signals that cannot be detected in univariate EEG analyses, which could be used to generate new hypothesis for both basic and more applied scientific studies.

The results of our study show that SVM is an excellent and robust choice for decoding EEG data in cognitive neuroscience experiments, with high decoding accuracy, good effect sizes and valid feature importance weights. We further suggest that EEG decoding provides an important tool that can complement the EEG/ERP method, because it will allow researchers to examine a variety of questions about the nature of the information that is contained in the EEG/ERP signal – such as whether and when the semantic features of upcoming words are pre-activated (Heikel et al., 2018) or how phonemic representations are sustained over time (Gwilliams et al., 2020). Questions such as these are difficult to answer with univariate methods alone.

5. Conclusion

In this technical note, we formally compared three classification methods - SVM, LDA, and RF - frequently used for decoding EEG signals. Our goal was to examine the performance and utility of these three ML methods for cognitive neuroscience studies, and we used two visual word priming experiments as a test case in the present study. We demonstrated that SVM is a highly reliable and valid choice for classification of binary data in cognitive neuroscience EEG experiments. SVM showed the highest decoding accuracy and the greatest sensitivity for detecting differences between experimental conditions. Additionally, SVM showed reliable feature weight patterns for testing a priori assumptions about electrode importance. This finding further paves the way for future studies of the content of cognitive computations in EEG signals that can advance basic neuroscience. Because SVM can classify single trial EEG data with a limited number of training trials (i.e., 40), it can also provide a powerful tool to examine individual differences in neurotypical and neuro-diverse populations.

Declarations of interest: none

CRediT authorship contribution statement
Timothy Trammel: Conceptualization of decoding study, Methodology, Machine Learning scripts, Writing – original draft, Writing – review & editing. **Natalia Khodayari:** Methodology, Writing – original draft, Writing – review & editing. **Steven J. Luck:** Methodology, Writing – review & editing, Supervision. **Matthew J. Traxler:** Writing – review & editing, Supervision. **Tamara Y. Swaab:** Writing- original draft, Writing - review & editing, Supervision.

Acknowledgements

TT was supported by a SMART Scholarship, a Department of Defense Program, and by the United States Navy.

Supplementary Materials

1. ERP Methods

1.1. EEG Recording and Analyses

In addition to our decoding analyses, we performed univariate analyses of the EEG data from Experiment 1 (Brothers et al., 2016) and Experiment 2 (Kappenman et al., 2021) on the same EEG data that were used for the three ML classifications. The goal of these ERP analyses was to test for significant differences between conditions that the three ML classifiers should be able to detect.

1.1.1 EEG Recording.

In Experiment 1, in addition to the electrodes for EEG recording described in the main paper, horizontal and vertical eye-movements were recorded via bi-polar montages, placed lateral to the outer canthi of the left and the right eye and the above the supra- and below the subordinal ridges of the right eye, respectively. The left mastoid was included as an active channel for later off-line re-referencing. Horizontal and vertical EOG electrodes were used to capture oculo-motor activity, such as blinks and saccades. All electrode impedances were kept below 5 kΩ. For details of Experiment 2, please see Kappenman et al., (2021). Data are available at: https://doi.org/10.18115/D5JW4R.

In both experiments, offline pre-processing analyses were conducted in MATLAB using the EEGLAB toolbox (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014). Independent component analysis (ICA) was performed to isolate and remove ocular-motor artifacts. Single-trial waveforms were then screened for amplifier drift, muscle artifacts and eye movements, and any epochs containing these artifacts were rejected prior to analysis. Table S1 and S2 show the number of trials rejected per subject and electrode site for Experiment 1 and 2, respectively. ERPs were calculated by averaging individual EEG epochs, time-locked to the onset of the target words. In Experiment 1 this was done for an epoch of 1600 ms and in Experiment 2 for an epoch of 1000 ms, in both experiments with a 200 ms pre-stimulus baseline.

Participant	Predicted-Related			Unpredicted-Related			Unpredicted-Unrelated		
	Accepted Trials	Rejected Trials	Reject %	Accepted Trials	Rejected Trials	Reject %	Accepted Trials	Rejected Trials	Reject %
1	167	5	2.9%	142	6	4.1%	150	3	2.0%
2	47	110	70.1%	72	91	55.8%	61	99	61.9%
3	136	3	2.2%	168	4	2.3%	148	9	5.7%
4	162	0	0.0%	157	1	0.6%	158	1	0.6%
5	134	11	7.6%	159	16	9.1%	139	20	12.6%
6	41	110	72.8%	44	125	74.0%	41	118	74.2%
7	157	0	0.0%	163	0	0.0%	160	0	0.0%
8	124	9	6.8%	176	11	5.9%	153	6	3.8%
9	66	15	18.5%	204	35	14.6%	136	21	13.4%
10	134	5	3.6%	178	3	1.7%	154	5	3.1%
11	207	4	1.9%	107	1	0.9%	157	1	0.6%

12	31	103	76.9%	62	124	66.7%	33	127	79.4%
13	145	0	0.0%	175	0	0.0%	160	0	0.0%
14	138	2	1.4%	178	2	1.1%	158	2	1.3%
15	72	9	11.1%	190	48	20.2%	119	33	21.7%
16	169	5	2.9%	142	4	2.7%	155	3	1.9%
17	147	4	2.6%	157	11	6.5%	151	9	5.6%
18	161	0	0.0%	154	0	0.0%	157	0	0.0%
19	146	0	0.0%	161	0	0.0%	155	0	0.0%
20	152	0	0.0%	168	0	0.0%	159	0	0.0%
21	223	1	0.4%	96	0	0.0%	156	1	0.6%
22	118	35	22.9%	104	57	35.4%	96	60	38.5%
23	117	9	7.1%	186	8	4.1%	153	7	4.4%
24	144	3	2.0%	158	15	8.7%	152	7	4.4%
25	127	1	0.8%	190	2	1.0%	153	4	2.5%
26	147	0	0.0%	172	1	0.6%	160	0	0.0%
27	167	3	1.8%	147	3	2.0%	150	10	6.3%
28	132	0	0.0%	187	0	0.0%	159	1	0.6%
29	177	4	2.2%	131	8	5.8%	154	6	3.8%
30	148	7	4.5%	157	8	4.8%	157	3	1.9%
31	160	1	0.6%	156	3	1.9%	158	2	1.3%
32	187	0	0.0%	133	0	0.0%	160	0	0.0%
33	147	2	1.3%	170	1	0.6%	159	1	0.6%
34	135	32	19.2%	129	24	15.7%	121	37	23.4%
35	181	10	5.2%	120	9	7.0%	153	7	4.4%
36	7	0	0.0%	290	23	7.3%	2	0	0.0%
37	142	4	2.7%	167	7	4.0%	150	10	6.3%
38	178	0	0.0%	142	0	0.0%	159	1	0.6%
39	32	55	63.2%	67	166	71.2%	48	112	70.0%
40	192	1	0.5%	125	0	0.0%	159	0	0.0%
41	231	6	2.5%	80	3	3.6%	150	5	3.2%
42	74	119	61.7%	50	77	60.6%	52	108	67.5%
43	191	0	0.0%	129	0	0.0%	159	0	0.0%
44	201	9	4.3%	109	1	0.9%	155	5	3.1%
45	232	22	8.7%	61	5	7.6%	155	3	1.9%
46	116	3	2.5%	198	3	1.5%	155	1	0.6%
47	274	0	0.0%	46	0	0.0%	158	0	0.0%
48	180	0	0.0%	140	0	0.0%	160	0	0.0%
49	126	51	28.8%	83	60	42.0%	112	45	28.7%
50	187	0	0.0%	132	0	0.0%	160	0	0.0%

51	58	36	38.3%	109	98	47.3%	68	83	55.0%
52	197	0	0.0%	123	0	0.0%	157	0	0.0%
53	163	0	0.0%	157	0	0.0%	160	0	0.0%
54	157	3	1.9%	158	2	1.3%	160	0	0.0%
55	151	0	0.0%	168	0	0.0%	158	2	1.3%
56	140	1	0.7%	152	27	15.1%	144	16	10.0%

Table S1: Shows accepted and rejected trials after artifact rejection for each of the original 56 participants in Experiment 1. After artifact rejection, participants with fewer than 80 trials were excluded from analyses (indicated in bold).

1.1.2 ERP analyses.

In both experiments, we measured the mean amplitude of the N400 in the typical 300-500 ms epoch following target onset. This was done for each of the three conditions in Experiment 1 (predicted-related, unpredicted-related, unpredicted-unrelated) and for the two conditions of

Experiment 2 (related vs. unrelated). For each experiment, we performed two repeated-measures analysis of variance (ANOVAs) with Greenhouse-Geisser corrections. In Experiment 1, we performed a 3x5 (Condition x Midline) ANOVA with a 5-level factor of Anteriority at Midline electrode sites (AFZ, FZ, CZ, PZ, POZ) and a 3x3x2 (Condition x Anteriority x Hemisphere) ANOVA with a 3-level factor of Anteriority at Lateral electrode sites – Frontal (FP1/2, F7/8, F3/4), Central (FC5/6, FC1/2, C3/4, CP1/2, CP5/6), and Posterior (T5/6, P3/4, O1/2) – and a 2-level factor of Hemisphere – Left (FP1, F3, F7,FC1, FC5, C3, T3, CP1, CP5, P3, T5, O1) and Right (FP2, F4, F8, FC2, FC6, C4, T4, CP2, CP6, P4, T6, O2). In Experiment 2 we performed a 3x6 (Condition x Midline) ANOVA with 6-level factor of Anteriority at Midline electrode sites (FZ, FCZ, CZ, CPZ, PZ, OZ) and a 3x3x2 (Condition, Anteriority x Hemisphere) ANOVA with 3-level factor of Anteriority – Frontal (FP1/2, F7/8, F3/4), Central (C3/4, FC3/4, C5/6), and Posterior (P3/4, O1/2, P7/8, PO7/8, PO3/4) – and 2-level factor of Hemisphere – Left (FP1, F3, F7, FC3, C3, C5, P3, P7, PO7, PO3, O1) and Right (FP2, F4, F8, FC4, C4, C6, P4, P8, PO8, PO4, O2). Due differences in the epochs across the two experiments, we did not perform analyses during later windows. However, we did generate scalp topography maps for the differences between conditions (Experiment 1: unpredicted-related minus predicted-related, unpredicted-unrelated minus predicted related, unpredicted-related minus unpredicted-related; Experiment 2: unrelated minus related) for 600 – 800 ms to allow for comparisons with the decoding feature weights (Figure 5 and S3).

Participant	Related			Unrelated		
	Accepted Trials	Rejected Trials	Reject %	Accepted Trials	Rejected Trials	Reject %
1	46	7	13.2%	46	11	19.3%
2	49	8	14.0%	45	7	13.5%
3	51	9	15.0%	54	5	8.5%
4	46	11	19.3%	50	6	10.7%
5	42	16	27.6%	47	8	14.5%
6	56	1	1.8%	54	0	0.0%
7	56	1	1.8%	58	2	3.3%
8	51	5	8.9%	56	3	5.1%
9	13	40	75.5%	15	21	58.3%
10	45	11	19.6%	38	13	25.5%
11	57	2	3.4%	52	3	5.5%
12	43	12	21.8%	50	9	15.3%
13	40	18	31.0%	47	8	14.5%
14	34	25	42.4%	27	24	47.1%
15	51	4	7.3%	49	7	12.5%
16	49	10	16.9%	47	11	19.0%
17	51	7	12.1%	51	6	10.5%
18	53	4	7.0%	53	3	5.4%
19	56	3	5.1%	60	0	0.0%
20	46	9	16.4%	40	14	25.9%
21	55	2	3.5%	55	4	6.8%
22	53	1	1.9%	53	3	5.4%
23	50	9	15.3%	47	11	19.0%
24	57	2	3.4%	52	4	7.1%
25	37	19	33.9%	46	9	16.4%
26	55	5	8.3%	54	4	6.9%
27#	41	1	2.4%	50	4	7.4%
28	52	4	7.1%	55	2	3.5%
29	44	14	24.1%	42	16	27.6%
30	29	30	50.8%	37	22	37.3%
31	49	11	18.3%	49	8	14.0%
32	56	0	0.0%	53	0	0.0%
33	59	0	0.0%	57	1	1.7%
34	53	4	7.0%	56	3	5.1%
35	35	12	25.5%	46	13	22.0%
36	56	1	1.8%	55	1	1.8%
37	55	3	5.2%	56	3	5.1%
38	40	17	29.8%	47	10	17.5%
39	59	0	0.0%	58	1	1.7%
40	6	48	88.9%	4	49	92.5%

Table S2: Shows accepted and rejected trials after artifact rejection for each of the original 56 participants in Experiment 1. Excluded participants are indicated in bold. One participant was excluded for poor task performance (#). Other exclusions were due to participants having fewer than 30 trials per condition after artifact rejection.

1.2 ERP results

Figure S1 illustrates mean amplitude (μV) difference waves across conditions for electrode Pz for Experiment 1 (panel A) and Experiment 2 (panel B). In Experiment 1, we found

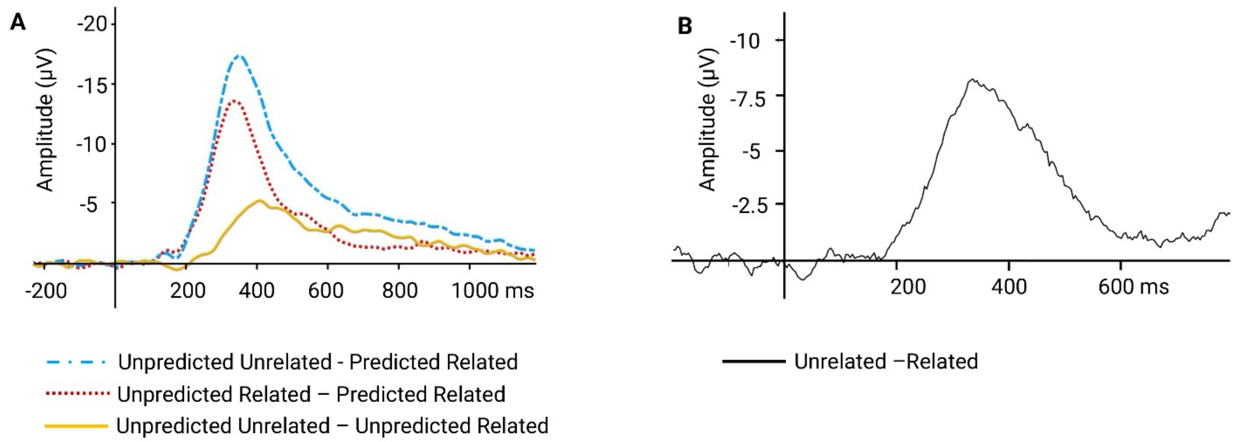

Figure S1: Difference waveforms showing N400 effects for electrode Pz for the comparison of all conditions in Experiment 1 (panel A) and Experiment 2 (panel B).

a significant main effect of Condition ($F_{GG}(2, 88) = 463$, $p < 0.0001$). Subsequent pairwise comparisons showed significant N400 effects when comparing predicted related vs unpredicted-related ($t(44) = -20.1$, $p < 0.0001$) and vs unpredicted-unrelated ($t(44) = -24.7$, $p < 0.0001$). Unpredicted-related was also significantly different from unpredicted-unrelated ($t(44) = -12.8$, $p < 0.0001$). Topographic distribution analyses show a significant main effect of Condition (Midline: $F_{GG}(2, 88) = 403.4$, $p < 0.0001$); Lateral: $F_{GG}(2, 88) = 326.3$, $p < 0.0001$)), and significant interactions for Condition x Midline ($F_{GG}(8, 352) = 111.8$, $p < 0.0001$)), Condition x Hemisphere ($F_{GG}(2, 88) = 30.1$, $p < 0.0001$)), Condition x Anteriority ($F_{GG}(4, 176) = 59.1$, $p < 0.0001$)), and Condition x Hemisphere x Anteriority ($F_{GG}(4, 176) = 9.6$, $p < 0.0001$)), consistent with the centro parietal distribution of the N400 effect. Topographic maps in Figure 5 of the main article show that differences between conditions were greatest across central-parietal and right-hemispheric electrode sites.

In Experiment 2, the N400 effect of semantic priming was significant ($t(34) = 12.2$, $p < 0.0001$). Topographic distribution analyses show a significant main effect of Condition (Midline: $F(1, 34) = 146.6$, $p < 0.0001$); Lateral: $F(1, 34) = 111.8$, $p < 0.0001$)). We found significant interactions for Condition x Midline ($F_{GG}(5, 170) = 23.8$, $p < 0.0001$)), Condition x Hemisphere ($F_{GG}(2, 68) = 7.3$, $p < 0.0001$)), Condition x Anteriority ($F_{GG}(2, 68) = 9.0$, $p < 0.001$)), and Condition x Hemisphere x Anteriority ($F_{GG}(4, 176) = 9.6$, $p < 0.01$)), consistent with the typical topographic distribution of the N400 effect. Topographic maps in Figure 5 of the main article show that the N400 amplitude difference between related and unrelated conditions was greatest across central-parietal and right-hemispheric electrode sites.

We also calculated *Cohen's* d_z effect sizes (Figure S2) for the ERP data using the same methods as were done in our decoding analyses. Average effects sizes were calculated for the 300 – 500 ms post-target epoch and summarized in Table S3. Finally, in Figure S3 we show the distribution of the feature weights and topographic distribution of the ERP effects for the remaining conditions that are not shown in Figure 5 of the main article. The electrodes that were most important in the classification decisions of the classifiers (highest weights) were similar to the electrode sites where the ERP effects were maximal in the 300 - 500 ms (N400) epoch and the 600 - 800 ms epoch.

N400 Effect Sizes (300 – 500 ms)					
Exp.1				**Exp. 2**	
Electrode	PR v. UR	PR v. UU	UR v. UU	Electrode	R v. U
FP1	0.497	0.823	0.3865	FP1	-0.0679
FP2	0.592	1.025	0.4984	FP2	0.1818
F3	1.021	1.452	0.4428	F7	-0.319
F4	1.163	1.813	0.7048	F3	0.468
F7	0.487	0.61	0.1151	FZ	0.7569
F8	0.679	1.176	0.5376	F4	0.825
FC1	**1.357**	**1.986**	**0.6535**	F8	0.4357
FC2	**1.457**	**2.179**	**0.7664**	**FC3**	**0.7462**
FC5	1.122	1.377	0.262	FCZ	1.0939
FC6	1.304	1.963	0.7065	**FC4**	**1.1077**
C3	**1.686**	**2.291**	**0.6574**	**C5**	**0.5488**
C4	**1.855**	**2.655**	**0.8933**	C3	1.1106
T3	1.142	1.179	0.0698	**CZ**	**1.3756**
T4	1.432	2.068	0.7062	C4	1.2962
CP1	**1.944**	**2.776**	**0.9373**	**C6**	**1.1984**
CP2	**2.03**	**2.922**	**1.0007**	**CPZ**	**1.406**
CP5	**1.811**	**2.283**	**0.578**	**P7**	**0.7954**
CP6	**1.959**	**2.704**	**0.8893**	**P3**	**1.2144**
P3	**1.962**	**2.727**	**0.8961**	**PZ**	**1.2393**
P4	**2**	**2.824**	**0.9827**	**P4**	**1.366**
T5	1.576	1.904	0.3972	**P8**	**1.2477**
T6	1.488	2.149	0.7907	PO7	1.1126
O1	1.438	2.087	0.7094	PO3	1.2944
O2	1.455	2.148	0.7985	PO4	1.2673
AFZ	0.836	1.391	0.5743	PO8	1.0787
FZ	1.108	1.72	0.6319	O1	1.0897
CZ	1.736	2.538	0.8809	OZ	1.1692
PZ	**1.961**	**2.835**	**0.9986**	O2	1.2095
POZ	1.769	2.614	0.9499		

Table S3: shows the average effects sizes of the N400 effects in Exps. 1 and 2 for each electrode site. Conditions are abbreviated as follows: PR=Predicted Related, UR=Unpredicted Related, UU= Unpredicted Unrelated, R=Related, U=Unrelated. Bold indicates electrode is within ROI.

2. Supplementary Figures

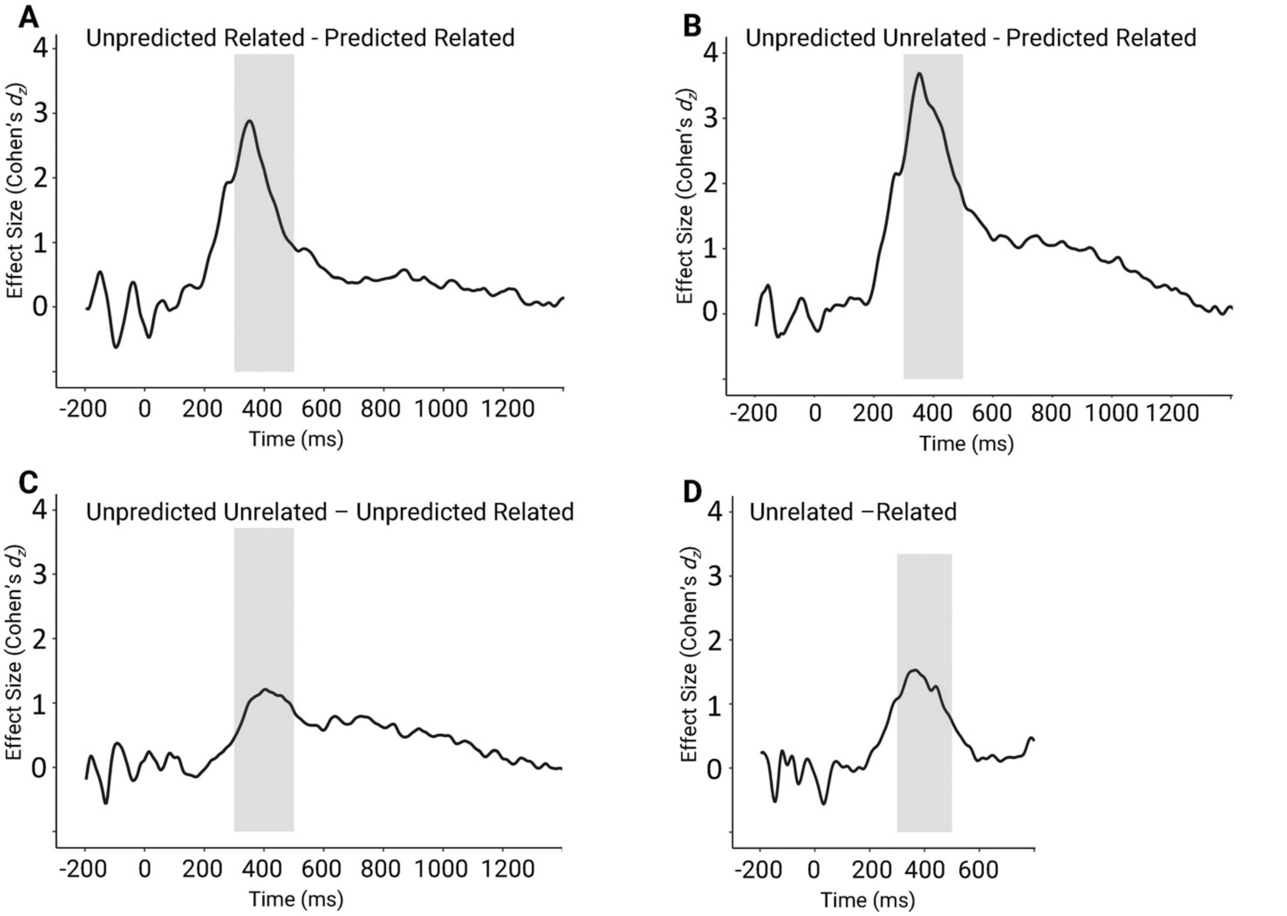

Figure S2: Effect sizes (Cohen's d_z) shown for the ERP effects for electrode Pz for Experiment 1 in panel A: unpredicted-related minus predicted-related, panel B: unpredicted-unrelated minus predicted related, and panel C: unpredicted-related minus unpredicted-related, and for Experiment 2 in panel D: unrelated minus related. Shading indicates the time window (300 – 500 ms) where significant differences between conditions were observed.

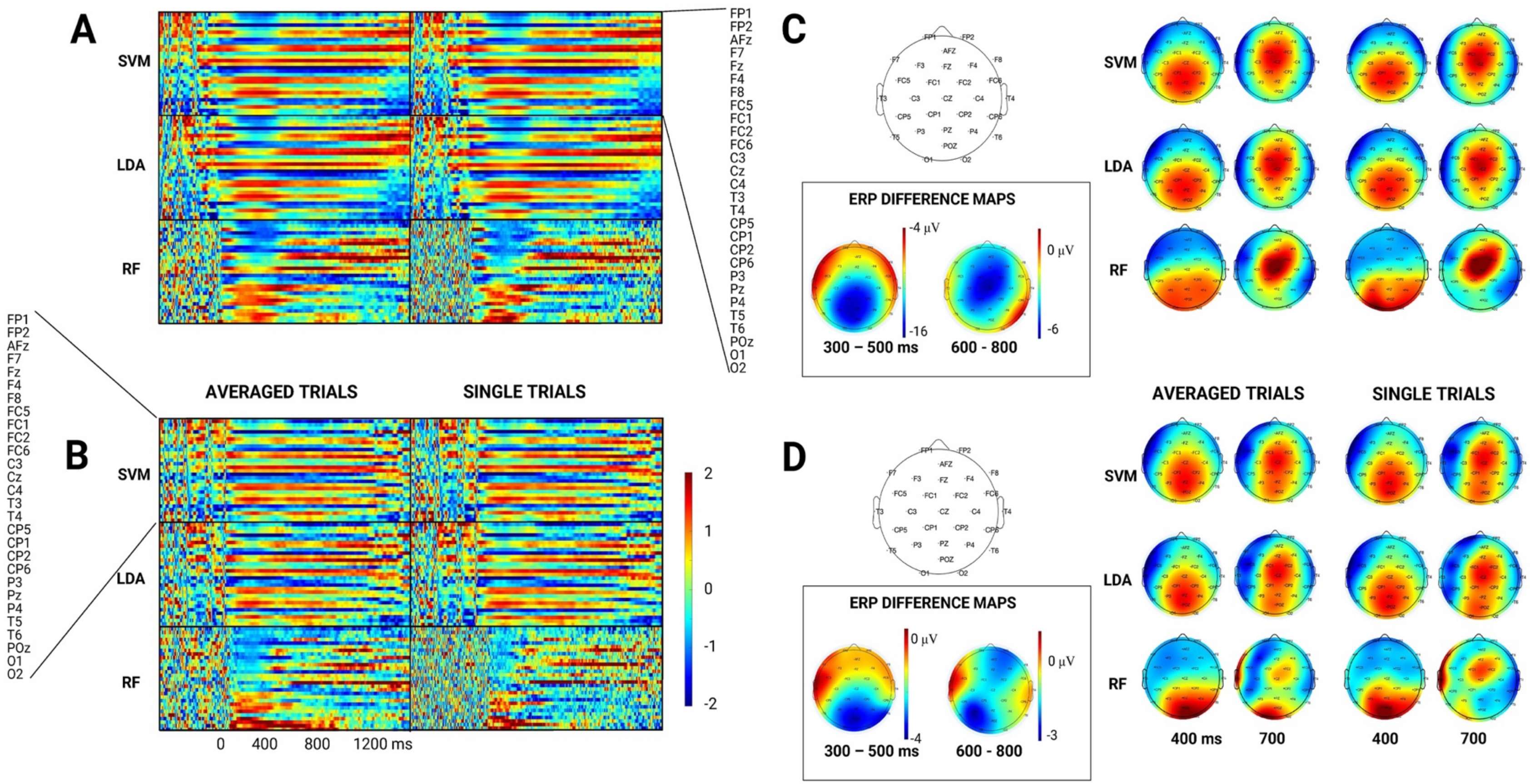

Figure S3: Feature importance maps for all three ML methods for the two remaining conditions of Experiment 1 (see figure 5 in the main text for the feature maps of the third condition in Experiment 1 and for Experiment 2). The maps show the normalized weighting of each electrode site in the decoding of the target words in for the predicted-related vs. unpredicted-unrelated condition (panel A) and for the unpredicted-related vs. unpredicted-unrelated condition (panel B). Six boxes of feature maps are shown for each condition for SVM, LDA and RF, respectively, on the left for the averaged trials, on the right for the single trails. Each box shows data for all electrode sites over the entire epoch from anterior to posterior sites, along the vertical axis. Included electrode sites are shown on the left and right of the figure (same for both panels). Values are in standard deviations from mean weights for a given time point. While the figure legend is capped at +/- 2, actual weights can exceed these limits. Panels C (predicted related vs. unpredicted unrelated) and D (unpredicted-related vs. unpredicted-unrelated) show feature weights from select time points (400 ms and 700 ms) along with difference topography maps of the ERPs for the conditions being decoded (panel C: unpredicted-unrelated minus predicted-related averaged separately over 300 – 500 ms and 600 – 1000 ms; panel D: unpredicted-unrelated minus unpredicted-related averaged separately over 300 – 500 ms and 600 – 800 ms). Values for ERP difference maps are in µV.

References

Bae, G.-Y., & Luck, S. J. (2018). Dissociable Decoding of Spatial Attention and Working Memory from EEG Oscillations and Sustained Potentials. *The Journal of Neuroscience*, *38*(2), 409–422. https://doi.org/10.1523/JNEUROSCI.2860-17.2017

Bae, G.-Y., & Luck, S. J. (2019). *Appropriate Correction for Multiple Comparisons in Decoding of ERP Data: A Re-Analysis of Bae & Luck (2018)* [Preprint]. Neuroscience. https://doi.org/10.1101/672741

Bentin, S., Kutas, M., & Hillyard, S. A. (1993). Electrophysiological evidence for task effects on semantic priming in auditory word processing. *Psychophysiology*, *30*(2), 161-169.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152. https://doi.org/10.1145/130385.130401

Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2015). Graded expectations: Predictive processing and the adjustment of expectations during spoken language comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 607–624. https://doi.org/10.3758/s13415-015-0340-0

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition*, *136*, 135–149. https://doi.org/10.1016/j.cognition.2014.10.017

Brothers, T., Swaab, T. Y., Traxler, M. J. (2016). Isolating visual, orthographic, and semantic pre-activation during lexical processing. Poster presentation at the 2016 Annual meeting of the Cognitive Neuroscience Society.

Chwilla, D. J., Brown, C. M., & Hagoort, P. (1995). The N400 as a function of the level of processing. *Psychophysiology*, *32*(3), 274-285.

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition*, *187*, 10-20.

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

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491-505.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

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. https://doi.org/10.1162/jocn_a_01068

Gwilliams, L., & Davis, M. H. (2020). Extracting language content from speech sounds: An information theoretic approach. In *The Auditory Cognitive Neuroscience of Speech Perception*. https://hal.archives-ouvertes.fr/hal-03013496

Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2020). *Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.04.04.025684
Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage, 87, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067
Hebart, M. N., & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. NeuroImage, 180, 4–18. https://doi.org/10.1016/j.neuroimage.2017.08.005
Heikel, E., Sassenhagen, J., & Fiebach, C. J. (2018). Decoding semantic predictions from EEG prior to word onset. *BioRxiv*, 393066. https://doi.org/10.1101/393066
Holcomb, P. J. (1988). Automatic and attentional processing: An event-related brain potential analysis of semantic priming. *Brain and language*, *35*(1), 66-85.
Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135. https://doi.org/10.1016/j.brainres.2015.02.014
Huettig, F., Audring, J., & Jackendoff, R. (2022). A parallel architecture perspective on pre-activation and prediction in language processing. *Cognition*, *224*, 105050. https://doi.org/10.1016/j.cognition.2022.105050
Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, *86*, 157-171.
Kappenman, E. S., Farrens, J. L., Zhang, W., Stewart, A. X., & Luck, S. J. (2021). ERP CORE: An open resource for human event-related potential research. *NeuroImage*, *225*, 117465. https://doi.org/10.1016/j.neuroimage.2020.117465
Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A Tale of Two Positivities and the N400: Distinct Neural Signatures Are Evoked by Confirmed and Violated Predictions at Different Levels of Representation. *Journal of cognitive neuroscience*, *32*(1), 12–35. https://doi.org/10.1162/jocn_a_01465
Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, 31(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299
Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of cognitive neuroscience*, *25*(3), 484-502.
Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/
Noah, S., Powell, T., Khodayari, N., Olivan, D., Ding, M., & Mangun, G. R. (2020). Neural mechanisms of attentional control for objects: decoding EEG alpha when anticipating faces, scenes, and tools. *Journal of Neuroscience*, *40*(25), 4913-4924.
Marozzi, M. (2004). Some remarks about the number of permutations one should consider to perform a permutation test. Statistica, 64(1), 193–201. https://doi.org/10.6092/issn.1973-2201/32
Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024
McMurray, B., Sarrett, M. E., Chiu, S., Black, A. K., Wang, A., Canale, R., & Aslin, R. N. (2022). Decoding the temporal dynamics of spoken word and nonword processing from EEG. *NeuroImage*, *260*, 119457.

Murphy, A., Bohnet, B., McDonald, R., & Noppeney, U. (2022, May). Decoding Part-of-Speech from Human EEG Signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2201-2210).

Proix, T., Delgado Saa, J., Christen, A., Martin, S., Pasley, B. N., Knight, R. T., ... & Giraud, A. L. (2022). Imagined speech can be decoded from low-and cross-frequency intracranial EEG features. *Nature communications*, 13(1), 48.

Swaab, T. Y., Ledoux, K., Camblin, C. C., & Boudewyn, M. A. (2012). Language-related ERP components. In The Oxford handbook of event-related potential components (pp. 397–439). Oxford University Press.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

Wang, R., Janini, D., & Konkle, T. (2022). Mid-level Feature Differences Support Early Animacy and Object Size Distinctions: Evidence from Electroencephalography Decoding. *Journal of Cognitive Neuroscience*, *34*(9), 1670–1680. https://doi.org/10.1162/jocn_a_01883

Chapter 3

Decoding anticipated semantic and visual word features prior to word onset: Investigating the time course of prediction during word priming

1. Introduction

In our daily interactions, language processing typically appears effortless, masking the fact that it is, in fact, a multifaceted cognitive endeavor, that includes parsing of incoming linguistic stimuli, extracting meaning, integrating contextual information, and generating appropriate responses. In addition, language is often processed under suboptimal conditions, affected by both internal and external sources of noise, including lexical and syntactic ambiguities, lapses in attention, working memory limitations, and noisy environments. Furthermore, the language input is sequential and fleeting, requiring incremental interpretation as the signal unfolds over time. Thus, theories of language processing need to accommodate how humans efficiently process and make sense of language amidst varying conditions and constraints.

Contemporary approaches in psycholinguistics emphasize forward-looking, anticipatory processes to optimize comprehension efficiency amidst internal and external sources of noise and ambiguity (Altmann & Mirković, 2009; DeLong et al., 2021; Elman, 2004; Kuperberg, 2021). Under these approaches, interpretation entails using prior knowledge and experience to derive expectations about how the language input unfolds in the imminent future. The outcome of these anticipatory processes reduces the processing load that "bottom-up" perceptual information imposes within a given context. Such anticipatory, likelihood-driven, processes affect lexical, syntactic, and discourse processes under a variety of theoretical frameworks (e.g., Gibson et al., 2013; Hale, 2011; Jaeger & Snider, 2013; Kuperberg & Jaeger, 2016; Levy, 2008). One such forward-looking framework is predictive coding (Rao & Ballard, 1999), which proposes that higher cortical levels continuously generate top-down predictions of information at lower levels.

With respect to language processing a predictive coding framework suggests that the human brain continuously predicts a hierarchy of representations that span multiple timescales (Kuperberg & Jaeger, 2016). However, it has been challenging to pinpoint the nature and time course of anticipated information during prediction in language processing. The aim of the present study is to examine this by using machine learning decoding and mass univariate ERP analyses in a priming study with a prediction task.

A crucial assumption of predictive coding during language processing is that the brain generates internal predictions about incoming sensory information based on prior experiences at all levels of representation, including sub-lexical visual features (e.g. word length), lexical features (e.g., orthographic neighborhood density, lexical frequency) and semantic features (e.g., concreteness of words). Prediction should modulate processing of these imminent words when they are presented, leading to facilitated processing if the prediction of the imminent word completely or partially coincides with the actual input, or to processing costs if the prediction of the imminent word does not overlap at all with the actual input. Prediction coding models propose that a closer match between anticipated and received information results in smaller prediction errors. When such errors occur, they propagate through the neural network, signaling the need to update internal representations or predictions to better match the incoming information. Over time, minimizing prediction errors allows the brain to refine its internal models of the world, enhancing its ability to predict future language inputs. Additionally, there is evidence suggesting that readers and listeners adapt their anticipatory behavior when faced with a high rate of prediction errors, as demonstrated in studies by Brothers et al. (2017), and Ness & Meltzer-Asscher (2021). This phenomenon, termed 'rational prediction,' indicates that prediction errors influenced by the present context dynamically shape anticipation during real-time language processing.

The aim of the present study was to test two main assumptions of the predictive coding model for language processing: 1) whether and when linguistic features (e.g., concreteness of words) and sub-linguistic features (e.g., word length) are anticipated prior to the sensory input and 2) how anticipation, whether correct or incorrect, during predictive language processing influences the comprehension of incoming words, considering that anticipated information may fully, partially, or not at all align with the predicted word.

Studies of anticipation during prediction have been challenging, because language processing unfolds over time, and predictions are likely to be generated and updated dynamically as new information is encountered. Furthermore, because retrieval of semantic, syntactic, lexical, and visual features of words may occur rapidly and perhaps concurrently, it is methodologically difficult to disentangle if and when these features are anticipated, even with temporally sensitive methods such as eyetracking, event-related potentials (ERPs) and magnetoencephalography (MEG).

Eyetracking is a method used in studies of reading to monitor and analyze the eye movements of participants as they process written text, providing insights into the cognitive processes involved in language comprehension. But studies using eyetracking methods offer limited information about the specific nature of the information that is retrieved prior to fixation of a target word and the time course by which different sources of information might be anticipated. Furthermore, eye movement patterns do not provide direct evidence of predictive coding mechanisms operating in the brain.

Brain activity during language processing as it unfolds in real-time provide can be measured more directly by using ERPs, EEG oscillatory activity, or MEG. Studies using ERPs have been very informative about the brain's sensitivity to predictability of words in context, but they have typically not provided a direct measure of the pre-activation of information prior to the onset of the critical word. Furthermore, ERP studies that have used experimental manipulations that allowed for examination of anticipated information have typically focused on separate individual aspects of the language input, i.e., when an adjective was unexpected given the predicted noun that followed it (e.g., healthy-cake; Boudewyn et al., 2015)), or when prenominal adjectives or determiners were inconsistent with the gender or phonological properties of the predicted noun (DeLong et al., 2005; Grisoni et al., 2017; Szewczyk & Schriefers, 2013; Van Berkum et al., 2005). Although these studies suggest pre-activation of syntactic, phonological, and semantic features of words, they have not examined the time course by which different sources of information are anticipated during on-line language processes.

Several studies using MEG have examined the nature and time course of information anticipated during prediction of imminent predictable words (Dikker & Pylkkänen, 2013; Eisenhauer et al., 2022; Gwilliams et al., 2018). However, Eisenhauer and colleagues used a lexical repetition paradigm, making it difficult to distinguish between activation of lexical-semantic features as a function of the prime, or anticipation of lexical semantic features of the imminent target word. Dikker and colleagues used a picture word matching paradigm, and therefore did not directly test language driven anticipation. Gwilliams et al. focused on the time course of phoneme ambiguity resolution given the context but did not examine anticipation of other linguistic features in the same study.

Analytical techniques of neurophysiological data using representational similarity analyses (RSA) have been used to uncover the nature of anticipated information during language processing, by examining how the similarity structure of neural activity changes based on the predictability of upcoming linguistic input (e.g., Hubbard & Federmeier, 2021; Wang et al., 2018).These studies also suggest that semantic features are anticipated prior to the onset of the imminent predictable word. However, results of other studies using this approach suggest that evidence of activation of semantic features only became available after the initial phonemes of the critical word had been heard (Klimovich-Gray et al., 2019). These divergent results emphasize the importance of examining the time course of the retrieval of various linguistic features in the anticipatory time before the onset of a critical word.

Complementing RSA, machine learning algorithms potentially offer a powerful tool to elucidate the nature and timing of anticipated information in language studies. While both RSA and decoding of EEG signals involve analyzing patterns of neural activity, they differ in their goals and methodologies. RSA focuses on comparing similarity structures of neural representations, whereas decoding aims to directly infer the content of cognitive states or processes from neural data using machine learning algorithms. Previous studies have demonstrated the efficacy of machine learning classification in decoding the content of computations from electrophysiological data collected during working memory (Bae & Luck, 2018, 2019) or attention paradigms(Hong et al., 2020; Nadra et al., 2023; Noah et al., 2020). Trammel et al. (2023, Chapter 2 of this dissertation) have further demonstrated that the SVM machine learning algorithms could classify EEG data from two language priming paradigms with very high accuracy.

In the present study, SVM classification was used to examine whether and when semantic features (concreteness) and sub-lexical visual features (word-length) of target words in a visual two-word priming paradigm could be reliably decoded from the EEG prior to the onset of the target word. Additionally, it was examined whether anticipation of semantic and sub-lexical visual features of the target words modulated the SVM classification of EEG activation of these features when the target word was presented. Finally, mass univariate ERP statistical analyses (e.g., Fields & Kuperberg, 2020) were performed as well -- using the same conditions -- to examine if SVM decoding results could be reliably detected in the averaged ERP signal

The EEG data from the prediction priming study in Chapter 2 were used for decoding in the present study was well (see Figure 1 for paradigm; full details of the paradigm are in methods section below). Briefly, in this study, one third of the target stimuli was unrelated in meaning to the prime (e.g., pillow – clown), and the remaining stimuli were related in meaning (e.g., circus – clown). The association strength in the related condition was manipulated so that two target words were about equally predictable (e.g., circus – clown, vs. circus– acrobat). The task of the participants was to actively predict the target words following the prime, and to indicate with a button press whether the received word was the same as the word they anticipated. This ensured that participants were anticipating upcoming target words, at least most of the time. In the related condition, they accurately predicted the target word 50% of the cases (e.g., they predicted clown, and received clown), but in the other 50% of the cases they indicated that they had not predicted the identity of the target word (e.g., perhaps they had predicted acrobat instead of clown). In the unrelated condition, participants could not anticipate the identity of the target word. The ERP results to the target words of this study are shown for electrode Fz in Figure 2A below. These effects were statistically significant (see Chapter 2, supplementary materials). As can be seen in this Figure, relative to the unrelated condition, the N400 to the related words whose identity was not anticipated by the participants was reduced. Because participants were required to actively predict the upcoming word, this facilitatory effect could be attributable to anticipation of similar semantic features, but not the anticipation of sub-lexical features of word length, because participants had indicated that their predicted word was not the same as the target word. The ERPs to the anticipated target words did not show an N400, but instead showed a large positive shift, which is typically found to repeated words in lexical repetition paradigms (Rugg, 1987) and suggests that participants had indeed predicted the identity of the target words.

In the present study, SVM decoding was used to explore the anticipation of both semantic features (concreteness) and sub-lexical word length features. A second aim was to determine when during the anticipatory period preceding the target words, semantic and visual features became decodable. If the identity of the target words was accurately predicted, the SVM classifier should be able to reliably decode concreteness features before the presentation of the target words. Similarly, if sub-lexical visual features of words were anticipated, these features would also be decodable before the presentation of the target words. For target words where the identity was not predicted, SVM might still decode anticipated concreteness features, as the actual and predicted target words could share semantic features. However, in such cases, sub-lexical visual features such as word length should not be decodable before target onset, as the anticipated word and the received target word were not the same. Finally, in the unrelated condition, where potentially anticipated concreteness and word length features are unlikely to have overlapped, we expected these features not to be decodable before the onset of the target words. Given the predictions of hierarchical predictive coding accounts of language processing (Kuperberg & Jaeger, 2016), anticipation of higher-level semantic information should precede lower-level anticipation of visual features of words. Thus, concreteness information about words (a semantic feature) should be decodable earlier than word length (a sub-lexical visual feature).

We predict a similar pattern of results for the mass univariate ERP analyses. Effects of concreteness and word length on ERPs have been found in prior studies using univariate ERP analyses (e.g., see for effects of concreteness; and Hauk et al., 2006 for effects of word length). These findings have been replicated in large scale ERP studies of visual word processing without context that have performed regression analyses on individual trials. Based on these prior studies, we predict that ERP effects of word length would occur 100ms post-stimulus onset (N1) and ERP effects of concreteness around 400ms post stimulus onset (N400). We therefore predict a reduced amplitude N1 to longer vs. shorter words, and a reduced N400 to abstract relative to concrete words to the prime words. If concreteness and word length features are anticipated prior to target onset, then we predict that ERP effects of these features on the target words will be modulated as a function of prediction accuracy and relatedness. We do not have a-priori predictions about the epochs during which we might find evidence of anticipation of concreteness and visual word length prior to the onset of the target words, but again predict that concreteness features should be decodable before word-length features (Kuperberg & Jaeger, 2016).

2. Methods

2.1 Participants

Participants in the ERP experiment (N=56; age range 18-30; 34 female) were all undergraduate students at the University of California, Davis, who received course credit for participation. Prior to participation, all participants signed informed consent forms approved by the Institutional Review Board at UC Davis. All were monolingual native English speakers with normal or corrected-to-normal vision, and all but two participants were right-handed according to assessment with the Edinburgh Handedness Inventory (Oldfield, 1971). None of the participants reported a history of psychiatric or neurological disorders, or head trauma, and they did not use any neuro-active prescription medication. Statistical analyses were performed on the EEG data from 46 participants (see artifact rejection description in EEG recording; see supplemental Tables S1 and S2 for details on trials rejected).

2.2 Materials

Participants read related and unrelated word pairs. These pairs consisted of 320 trials in the related condition (circus – CLOWN) and 160 in the unrelated condition (pillow – CLOWN). The critical target words were the same in both conditions, but stimuli were distributed across lists to avoid repetition of target words within participants. The mean forward association strength was 0.5 (range = 0.4-0.6; Nelson et al., 2004), so that two equally likely target words were predictable (see task description in procedure). Please note that the primes in the unrelated condition had the same forward association strength characteristics (e.g., pillow- bed/ sleep), but in this case were replaced with an unrelated target word. Participants rate of accurate prediction closely aligned with forward association strength (mean= 51.5%, SD= 11.9%,). Concreteness of words in the experiment was used as a measure of semantic activation. Concreteness values were taken from the English Lexicon Project (Balota et al., 2007). Concreteness ratings are between 1 and 5 with 5 being most concrete and 1 being most abstract. The words were separated into high and low concreteness groups based on a median split of the concreteness values for the target words. To measure activation of sub-lexical visual features of the words, they were grouped into long (6 or more letters) and short (4 or less letters) words. Medium length words (5 letters) were omitted to allow some separation between the shorter and longer words.

Figure 1. *Predictive processing paradigm*

Stimuli were randomized and organized into five blocks. Block order was counterbalanced according to Latin Square design to minimize ordering effects. Three lists were used to allow counterbalancing of related and unrelated word pairs such that across participants, each target word was used as both related and unrelated target without repetition of words within participant.

2.3 Procedure

While recording EEG, participants were seated in an electrically shielded, sound-attenuated booth. At the beginning of each trial (see Figure 1), participants were presented with two separated and horizontal fixation lines for 200 ms. which remained on the screen throughout the trial. This was done so that participants could keep their eyes fixated, to prevent eye-movement artefacts. A prime word was then presented for 200 ms in white lower case Calibri text (font size: 65 pt.) between the fixation lines. Participants were asked to actively predict the upcoming target given the prime word during the 1800 ms delay before the onset of the target word. They were encouraged to use the first word that came to mind as their prediction to limit the possibility of switching to a different prediction. The participant practiced this orally during the instructions. Following the delay, a target word was displayed in uppercase letters for 200 ms followed by an additional 1400 ms delay, during which participants only saw the fixation lines.

Participants were then prompted with a blue question mark and were instructed to indicate, via button press, whether they had accurately predicted the target (yes) word or not (no). Participants were encouraged to honestly report whether they had predicted the presented target word. After the participant's response, the experiment automatically proceeded to the next trial following a 1200 ms intertrial interval with the fixation lines present. Stimuli were presented on an CRT monitor with a refresh rate of 60Hz connected to a Dell Optiplex 980 Small Form-Factor Intel Core i7 Quad-Core 2.8GHz Processor).

The paradigm generated three plausible outcomes dependent on whether participants had successfully predicted the identity of the target words, and whether target words were related or unrelated in meaning to the prime: successfully predicted related target words (*predicted*), unsuccessfully predicted related target words (*unpredicted*), and unsuccessfully predicted unrelated target words (*unrelated*). For example, if the participant sees "circus" as the prime word and predicts "clown" when clown is the target word, then that was a *predicted* trial. On the other hand, if the participant sees "circus" as the prime word and predicts "acrobat" when clown is shown as the target word, then that was an *unpredicted* trial. Finally, if the participant sees "pillow" as the prime word and predicts "bed" when clown is shown as the target word, then that was an *unrelated* trial. It is important to emphasize here that we assume that participants predicted the imminent target word for most trials, as this was their task. Thus, prediction accuracy in this paradigm indicates trials on which participants predicted a target word that was subsequently presented. A fourth outcome, successfully predicted unrelated words, is also technically possible, but is extremely unlikely and typically results from participant error (i.e., pressing predicted when they did not predict; <1% of total trials).

Condition specific codes were sent out at the onset of the prime and the onset of the target words. Target words were sorted for each participant based on their yes or no prediction accuracy response. Prior to decoding analyses, additional codes were generated according to prime and target word concreteness values and target word length values. Concreteness (high or low) and word length (short or long) were then also sorted in bins according to prediction accuracy and relatedness.

2.4 EEG Recording

EEG was recorded from 29 tin electrodes, embedded within an elastic cap (Electro-Cap International). Additionally, electrodes were placed on the outer canthi of the left and right eyes to monitor for horizontal eye-movements, and above and below the left eye to monitor for blinks. All electrode impedances were kept below 5 kΩ. The EEG signal was amplified using a Synamps Model 8050 Amplifier (Compumedics Neuroscan) with online AC band-pass 0.05-100 Hz and recorded digitally at a 250 Hz sampling rate. Channels were referenced online to the right mastoid electrode and re-referenced offline to the average of the right and left mastoid channels. Offline pre-processing analyses were conducted using the EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) MATLAB toolboxes. A 0.1 - 10 Hz band-pass filter was applied to reduce high frequency electrical noise in accordance with recommended filtering cutoffs for N400 ERP components (Zhang et al., 2024). Independent component analysis (ICA) was performed to isolate and remove ocular-motor artifacts. Epochs were generated 3600 ms post-prime stimulus onset with a 200 ms pre-prime baseline. The target word appeared 2000 ms after the prime onset. Artifact rejection was performed on these epochs to remove remaining artifacts from the data. Subjects with at least 20 trials per condition after pre-processing were retained. This led to rejection of the data of 8 participants, leaving data from 48 subjects for the analyses.

2.5 Analyses

To investigate the time course of anticipated information during predictive processing, we used a support vector machine (SVM; Boser et al., 1992) classifier to categorize the EEG data from each of our conditions. We chose to use an SVMs as our classifier as they have been shown to have superior performance when classifier linguistic data when compared to similar classifiers (linear discriminant analysis and random forest; Trammel et al., 2023). We first used the classifier to decode the main effects of prediction accuracy and relatedness by comparing *1) predicted vs. unpredicted, 2) predicted vs. unrelated,* and *3) unpredicted vs. unrelated.* This was done for a total epoch of 4000 ms, beginning 400 ms prior to prime onset until 2000 ms post target onset. Word feature decoding was done separately for each of the conditions (*predicted*, *unpredicted*, and *unrelated*) according to whether the target word was concrete or abstract (*concreteness*) and, separately, whether the *length* of the target word (number of letters) was short (4 or fewer letters) or long (6 or more letters). We also decoded the concreteness and word length of the prime words in the same conditions – *predicted*: primes followed by accurately predicted targets, *unpredicted*: primes followed by related targets that were not accurately predicted, and *unrelated*: primes followed by unrelated target words – to compare to our target word decoding and ensure that we were not merely decoding the correlation of prime and target word features. For the prime words we additionally tested reliable decoding of concreteness and word length independent of whether subsequent target words were accurately predicted or semantically related. This was done to identify the earliest epoch during which these effects could be reliably decoded. We will consider these earliest epochs as our estimate of effects of concreteness and word-length relatively uninfluenced by anticipatory predictions of these features for the target words. Finally, we performed mass univariate ERP analyses (e.g., Fields &

Kuperberg, 2020) using the same conditions to confirm that these differences are detectable with alternative methods.

2.5.1 Decoding Analyses

The SVM algorithm was implemented within MATLAB using the fitcsvm() function. Due to the 250 Hz sampling rate, the original time course is 1000 time points: one every 4 ms. The method classified 200 time points, one every 20 ms, across the -400 – 3596ms time-locked to the onset of the prime. We did this without down sampling the original 250 Hz data. All 29 scalp electrode sites were used to create the feature set – the set of input variables which the SVM classifies – for the classifier at each time point. For each time point classified, we utilized the electrode site voltages for that time point as well as the two time points preceding it and the two time points that followed. For example, when decoding the time point 400 ms after stimulus onset, all 29 electrode site voltages were used from 392 ms, 396 ms, 400 ms, 404 ms, and 408 ms providing a total of 145 features. Using this technique allowed us to functionally "down sample" the data from the original 250 Hz (one time point every 4 ms) to 50 Hz (one time point every 20 ms) to conserve computing time without sacrificing the dynamics of the data between time points. Machine-learning was performed across 24 iterations using 10-fold cross-validation. In each iteration, trials were separated into 10 blocks (9 training; 1 testing) and were alternated to allow each trial to serve as both a training and a test data point. To avoid spurious above-chance decoding accuracy (Carrasco et al., 2023), we kept the number of trials constant across conditions. This was done by using a random sampling of the larger of the two conditions equal to the number of trials in the smaller condition for each iteration. For each interval, trials were then averaged. To assess decoding accuracy, we calculated the proportion correctly classified for each classifier at each interval across all iterations for each subject and then averaged the final accuracy across subjects. Chance-level is indicated by a decoding accuracy of 50% correct. This entire process was done each of our conditions.

To determine if the SVM decoding accuracy was significantly above-chance, we used a cluster-based permutation method (Bae & Luck, 2018, 2019; Maris & Oostenveld, 2007) which we had successfully used for language stimuli in a prior method comparison study (Trammel et al., 2023). Cluster based analyses generated a null distribution using 10080 random permutations of the existing data labels. We performed *t* tests ($\alpha = 0.05$) at each time point for both the null distribution and the actual data. Contiguous, significant intervals were organized into clusters. To mitigate spurious findings, significant clusters consisting of a single time point were removed as orphan clusters. The *t* test results for the remaining significant clusters were summed. The summed *t* test results for each cluster in the data were compared to the 95th percentile of the summed *t* test results of the generated null distribution. If the *t* test sum of the data clusters was greater than the 95th percentile of the null distribution, then the decoding accuracy of that cluster was significantly greater-than-chance. This method avoided the need for multiple comparison correction and accounted for autocorrelation in the EEG data (Bae & Luck, 2019). The significance testing was limited to the 0 – 3596 ms time frame as the pre-stimulus baseline period (-400 – 0 ms) should not contain any effects. One-tailed t-tests were used because performance significantly lower than chance-level has no meaningful interpretation within this study. The null hypothesis of the permutation-cluster analysis is that within the examined epoch, there are no clusters of significantly above-chance decoding accuracy. Therefore, we were not explicitly testing the timing of these clusters. To address this issue, we followed up the initial testing with additional tests limiting the epoch to the time frames of each significant cluster from the first analysis. This process confirmed the onset and offset of the observed significant clusters.

2.5.2 ERP Analyses

For our ERP analyses, we implemented a cluster-based permutation analysis (Maris & Oostenveld, 2007) that is similar to that used for our decoding analyses. This was done as we did not have *a priori* assumptions of where effects of pre-activation of concreteness and length features might be found in the interval between prime and target, and this technique avoids multiple comparison problems. For each condition (*predicted*, *related*, and *unrelated*), we generated ERP waveforms for each participant of either the concrete vs. abstract words or long vs. short words. Then, we generated a null distribution of the differences in concreteness waveforms or word length wave forms using 10080 random permutations. We performed two-tailed *t* tests ($\alpha = 0.01$) at each scalp electrode site (29 electrodes) and each time point within the 0 – 3596 epoch (900 time points). We then generated spatio-temporal clusters of contiguous, significantly different, neighboring electrode sites across contiguous time points. Electrode sites were considered neighbors if they were within 4 cm of each other. Orphan clusters – those consisting of electrode sites or fewer than 7 contiguous time points (<28 ms) – were removed to mitigate spurious effects. The *t* test scores for these clusters were summed. If summed *t* tests of clusters from the data were less than the 0.5^{th} percentile of the null distribution or greater than the 99.5^{th} percentile of the null distribution, then that cluster was deemed significantly different. This method can generate very large spatio-temporal clusters; therefore, we used the stricter alpha criterion of 0.01 for our *t* tests.

3. Results

We performed EEG decoding analyses using SVM classifiers and mass univariate ERP analyses using cluster permutation testing and report the results below. Within all decoding analyses, significant time clusters are those for which their summed *t* test scores exceeded the 95th percentile of the null distribution generated by the one-tailed permutation cluster *t* test. For all mass univariate analyses, significant time clusters are those for which their summed *t* test scores exceeded the 99.5th percentile or were lower than the 0.5th percentile of the of the null distribution generated by the two-tailed permutation cluster *t* test. It is important to note that because cluster permutation testing is not designed to answer questions of latency, we worked

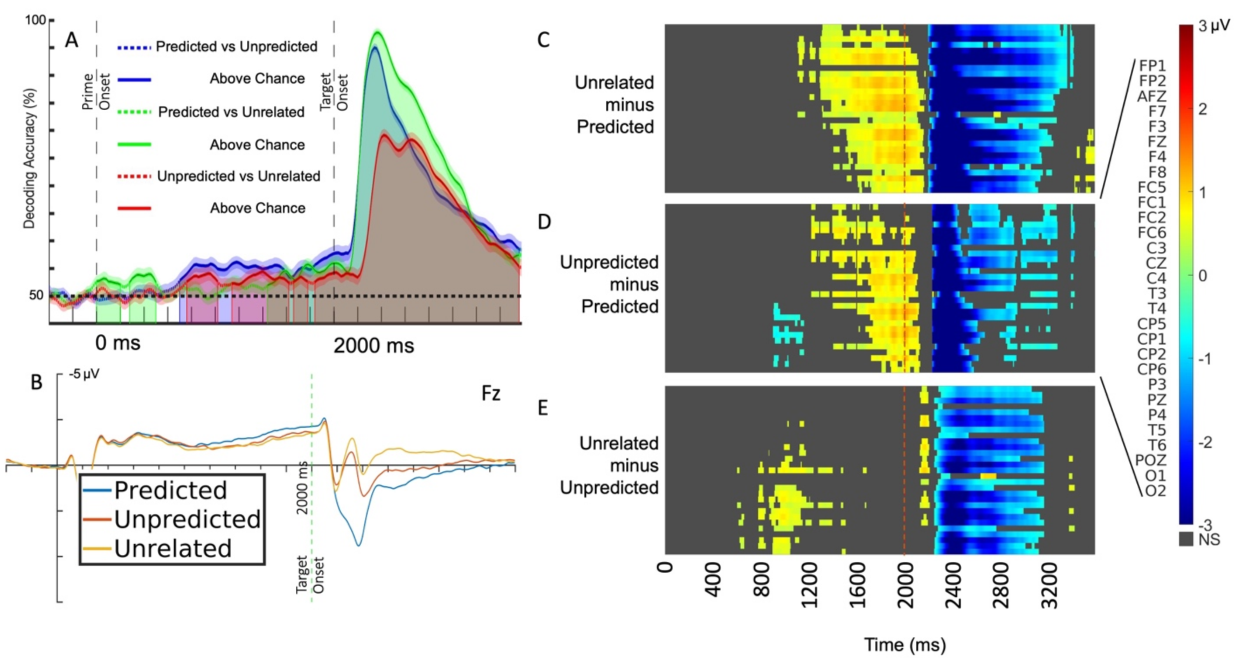

Figure 2. Shows the results from analyses of the main prediction and relatedness conditions: predicted, unpredicted, and unrelated. A) SVM decoding accuracy for three classification conditions: predicted vs. unpredicted (blue), predicted vs. unrelated (green), and unpredicted vs. unrelated (red). Prime (0 ms) and target (2000 ms) word onsets are indicated by vertical dashed lines (grey). Solid lines with shading under the curve highlight clusters of significantly above chance-level (50%; cluster t test sums exceeded 95th percentile of null distribution) decoding accuracy. B) ERP results for electrode Fz (see supplemental figure S1 for all electrode plots) waveform for the three main conditions: predicted (blue), unpredicted (red), and unrelated (yellow). The epoch is time-locked to prime onset and target onset is indicated by a vertical dashed line (green). ERP difference waves for the mass-univariate analyses are shown for: C) unrelated minus predicted, D) unpredicted minus predicted, and E) unrelated minus unpredicted. Target onset is indicated by the vertical dashed line (red). Depicted cluster t test sums were less than the 0.5th percentile of the null distribution or greater than the 99.5th percentile of the null distribution. Within these plots, non-significant time points are masked in grey (see supplemental figure S2 for unmasked plots).

around this limitation by performing follow-up cluster testing on all significant cluster epochs separately. This confirms that there is a significant cluster for both analyses for the epoch shown in all figures and tables.

3.1 Results of main prediction and relatedness effects

Table 1. Main Effects - Significantly above chance decoding accuracy

Condition	Cluster	Times (ms) Begin	End
Predicted vs Unpredicted	1	700	3560
Predicted vs Unrelated	1	0	200
	2	280	500
	3	1440	3560
Unpredicted vs Unrelated	1	760	1020
	2	1140	1620
	3	1660	1780
	4	1840	3560

Table 1. Summary of clusters where significantly above chance (50%; cluster t test sums exceeded 95th percentile of null distribution) decoding accuracy for the main effects of prediction and relatedness. Target onset is 2000 ms.

Figure 2 illustrates the results from our analyses of main prediction and relatedness effects. SVMs reliably classified (significantly above-chance; 50%) the ERP results for the target words: *predicted vs. unpredicted, predicted vs. unrelated,* and *unpredicted vs. unrelated* (Figure 2A) after target onset. As in Chapter 2, reliable decoding was found for these comparisons post target (see Table 1 for exact epochs) onset. In addition, all three conditions were decoded at above-chance levels prior to target word onset as well. The reliable decoding for *predicted vs. unpredicted* classification spanned a single continuous period from*:* 700-3560ms. In the *predicted vs. unrelated* reliably decoding was found from prime word onset (0 ms) through the end of the trial epoch with a few time periods where decoding accuracy dropped pre-target onset. Finally, *unpredicted vs. unrelated* conditions were decoded at above-chance levels prior to target word onset as well with reliable decoding first beginning at 760 ms. This pattern is consistent with anticipation of information about the target words that were predicted or unpredicted but related. Figure 2B shows the averaged ERP effects for electrode Fz, and the pattern of results shows facilitatory effects of relatedness, with the greatest facilitatory effect for related target words that were accurately predicted.

Figures 2C– 2E show the results of the mass-univariate permutation cluster analyses (Figures 2C – 2E)). These analyses confirm our observed N400 effects, and they echo the decoding results. Exact epochs and member electrodes are reported in Table 2. The mass-univariate analysis showed significantly different clusters prior to target onset when comparing the *predicted* and *unpredicted* conditions (earliest onset: 900 ms), when comparing the *predicted* and *unrelated* conditions (earliest onset: 1108 ms), and when comparing the *unpredicted* and *unrelated* conditions (earliest onset: 600 ms). Additionally, the periods of the most robust differences for all conditions overlap substantially with the significant pre-target decoding clusters. Interestingly, in this anticipatory period, more negative ERPs were found for the related predicted and related unpredicted than the unrelated condition, the opposite pattern than what is found for these same conditions to the target words. This pattern of results is entirely consistent with "costs" of retrieval and maintenance of information about the upcoming target words prior to their presentation, and reduced prediction error/ greater facilitation for target words that were (partially) predicted.

Table 2. Main Prediction and Relatedness Effects - Significantly different ERP cluster times

Condition	Cluster	Difference (μV)		Time (ms)		Electrodes
		Min	Max	Begin	End	
Unpredicted	1	-0.8	-0.4	900	1156	CP1 CP2 CP5 CP6 P3 P4 T6 O1 PZ POZ
minus	2	0.4	0.8	1220	1284	FP1 F3 F7 FC1 FC5 T3 AFZ FZ
Predicted	3	0.5	0.7	1296	1368	F3 F7 FC1 FC5 T3
	4	0.3	0.9	1396	1592	FP1 FP2 F3 F4 F7 FC1 FC2 FC5 C3 T3 CP5 T5 AFZ FZ
	5	0.3	0.9	1604	1716	FP1 F3 F4 F7 FC1 FC2 FC5 FC6 C3 T3 CP1 CP2 CP5 P3 T5 AFZ FZ CZ PZ
	6	0.4	1.3	1720	2128	FP1 F3 F4 F7 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
	7	-12.5	0.6	2232	3276	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
	8	-1.2	-0.4	3392	3420	FP1 FP2 F3 F7 FC1 FC5 C3 T3 CP5 AFZ FZ
	9	-0.6	-0.4	3592	3592	F7 FC5
Unrelated	1	0.3	0.6	1108	1168	F3 F4 F7 FC1 FC5 T3 AFZ FZ
minus	2	0.3	0.6	1204	1256	F3 F7 FC1 FC5 C3 FZ
Predicted	3	0.3	1.2	1288	2176	FP1 FP2 F3 F4 F7 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
	4	-15.9	0.6	2200	3360	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
	5	-0.9	-0.5	3388	3416	FP1 FP2 F3 F4 F7 F8 FC2 FC5 AFZ FZ
	6	0.4	0.6	3424	3464	P3 O1 O2 PZ POZ
	7	0.4	0.7	3516	3596	CP1 CP2 CP6 P3 P4 O1 O2 PZ POZ
Unrelated	1	0.3	0.5	600	656	C4 CP6 P4 O1 PZ POZ
minus	2	0.3	0.7	776	936	FC6 C4 CP1 CP2 CP5 CP6 P3 P4 T5 O1 O2 CZ PZ POZ
Unpredicted	3	0.3	0.9	940	1236	F4 FC2 FC6 C3 C4 CP1 CP2 CP5 CP6 P3 P4 T5 O1 O2 CZ PZ POZ
	4	0.4	0.6	1276	1396	C4 CP6 P4
	5	0.4	0.6	1468	1504	C4 CP6 P4 O2
	6	0.4	0.6	1652	1680	C4 CP6
	7	0.3	1.0	2124	2208	FP1 FP2 F4 FC1 FC2 FC6 C3 C4 T4 CP1 CP2 CP6 P4 AFZ FZ CZ PZ
	8	-4.7	1.0	2248	3164	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
	9	0.3	0.6	3376	3420	C3 T3 CP5 P3 T5

Table 2. Summary of all significant ERP clusters (cluster t test sums less than the 0.5th percentile of the null distribution or greater than the 99.5th percentile of the null distribution) found in the main prediction and semantic relationship effects analyses. Target word onset is 2000 ms.

3.2 Results of the feature decoding analysis

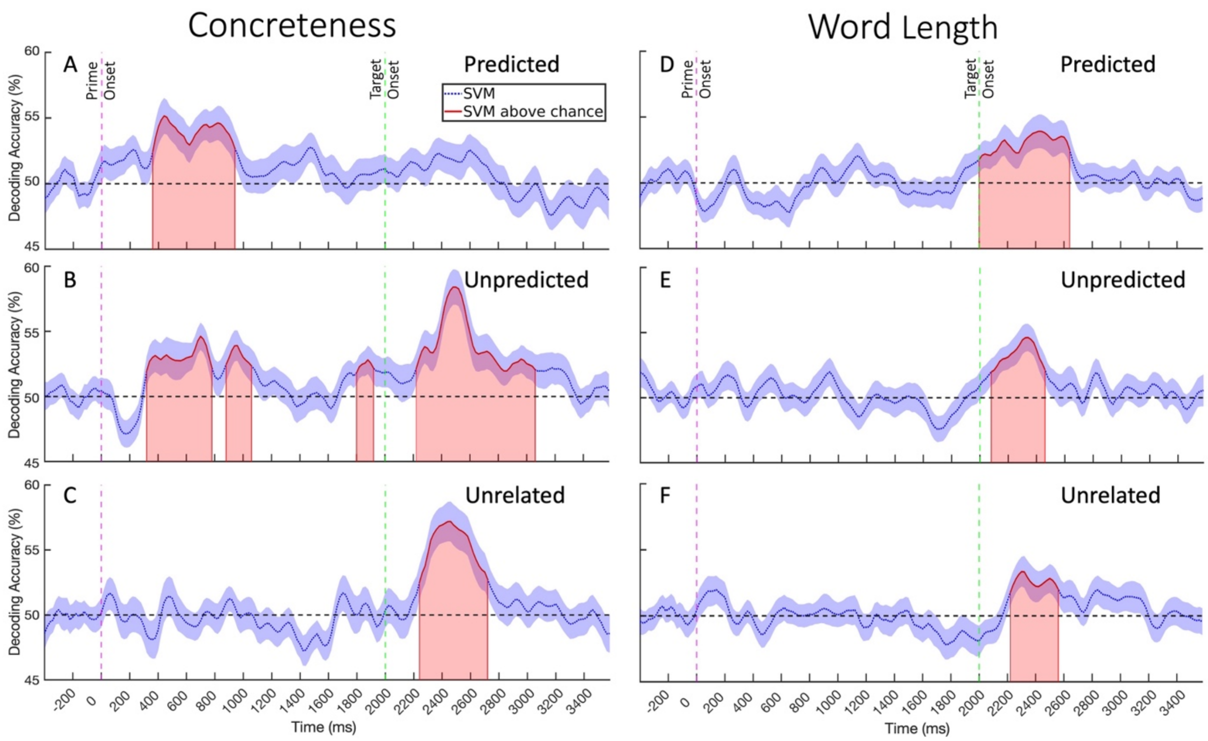

Figure 3*. Shows decoding accuracy of target word concreteness (A – C) and word length (D – F) classifications over the entire -400 – 3600 ms epoch. Prime word onsets at 0 ms (magenta) and target word onsets at 2000 ms (green). Solid red lines indicate decoding accuracy that is significantly above-chance level (50%; cluster t test sums exceeded 95th percentile of null distribution); highlighted by red shading under the curve.*

As illustrated in Figures 3 and 4, SVM reliably classified EEG data for concreteness and word-length. For the targets, SVMs were trained to classify concreteness and word-length based on the values of these features for the target words, but for the primes SVMs were trained to classify these features based on values for the prime words. Decoding of the features was examined for target words that were binned according to prediction accuracy and relatedness yielding three conditions: *predicted*, *unpredicted*, and *unrelated*. The same target word conditions were used to examine decoding of features for the primes. The exact epochs during which SVM reliably classified concreteness and length features are shown in Table 3. These epochs were found to have significantly above-chance decoding accuracy in the cluster-permutation analyses (exceeded 95th percentile of null distribution). As illustrated in Figure 3, significant clusters of above-chance decoding of concreteness features for the target words were found in epochs prior to target word onset for the accurately predicted targets and for related target words that were not accurately predicted. These significant clusters were found in a period that concentrated between 1640 to 1040ms prior to target word onset (between 360 – 940 ms in

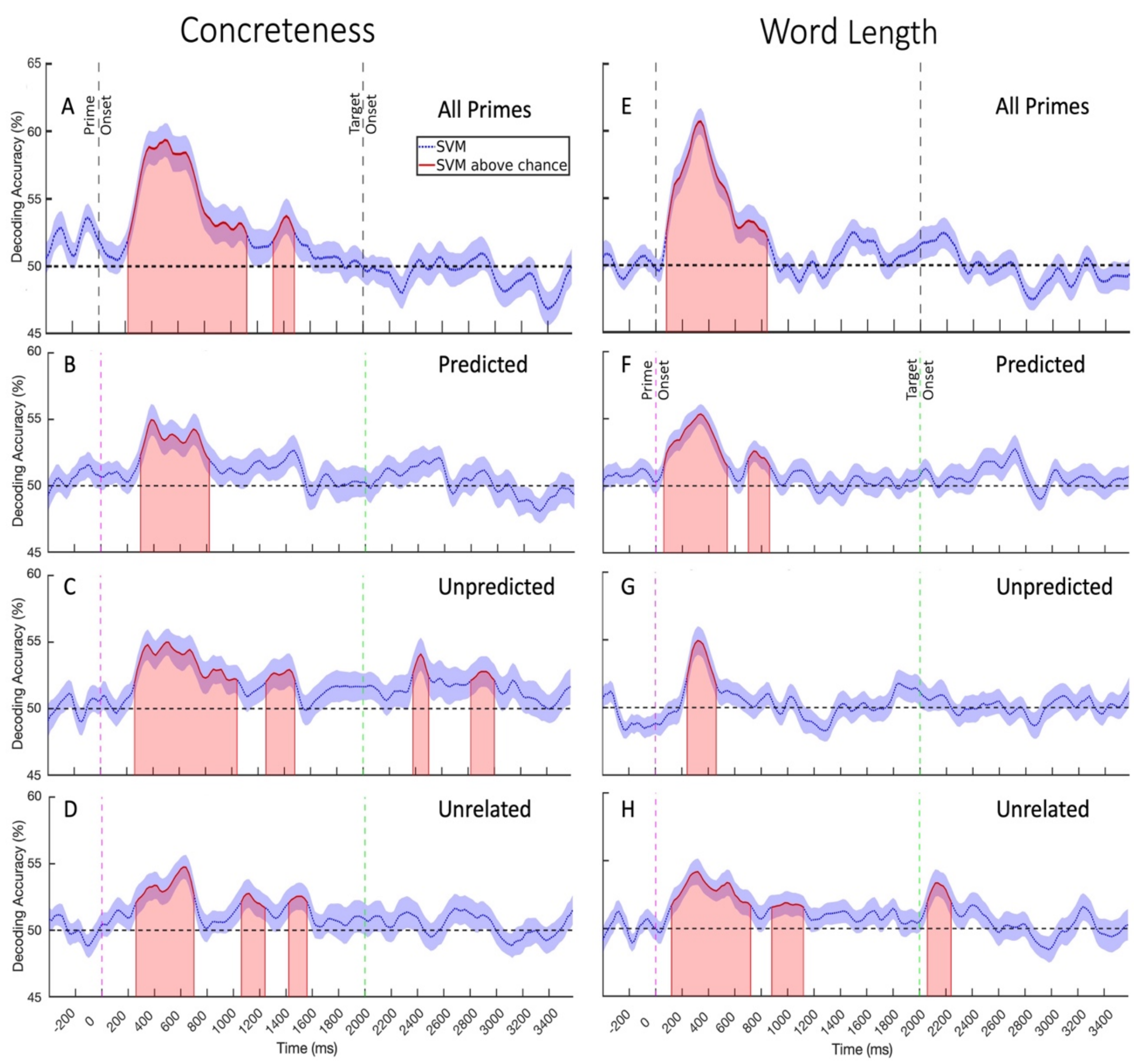

Figure 4. Shows decoding accuracy of prime word concreteness (A – D) and word length (E – H) classifications over the entire -400 – 3600 ms epoch. Prime word onsets at 0 ms (magenta) and target word onsets at 2000 ms (green). Solid red lines indicate decoding accuracy that is significantly above-chance level (50%; cluster t test sums exceeded 95th percentile of null distribution); highlighted by red shading under the curve.

Figure 3). In contrast, for unrelated target words, we did not find any evidence of significant decoding of concreteness features prior to target word onset. After target word onset,

Table 3. Features - Significantly above chance decoding cluster times

Decoded Word	Feature	Condition	Cluster	Times (ms)	
				Begin	End
Target	Concreteness	Predicted	1	360	940
		Unpredicted	1	320	780
			2	880	1060
			3	1800	1920
			4	2220	3060
		Unrelated	1	2240	2720
	Length	Predicted	1	2000	2640
		Unpredicted	1	2080	2460
		Unrelated	1	2220	2560
Prime	Concreteness	Predicted	1	300	820
		Unpredicted	1	260	1040
			2	1260	1480
			3	2380	2500
			4	2820	3000
		Unrelated	1	260	700
			2	1060	1240
			3	1420	1560
	Length	Predicted	1	60	540
			2	700	860
		Unpredicted	1	240	460
		Unrelated	1	120	720
			2	880	1120
			3	2060	2240

Table 3. Summary of all significantly above-chance (50%; cluster t test sums exceeded 95th percentile of null distribution) decoding cluster start and stop times. Target word onset is 2000 ms.

concreteness was significantly decodable for target words that were related but not accurately predicted and for unrelated target words in a period that spanned 220 - 1060ms relative to target onset (2220 - 3060ms in Figure 3). In contrast, for the accurately predicted target words we did not find any evidence for reliable decoding of concreteness features. For word length, clusters of significantly above-chance decoding were exclusively found after or at target word onset in all three conditions. *Predicted* target words, reliable decoding accuracy began at target word onset (2000 ms in Figure 3). Related *unpredicted* words were decodable early as well (80 ms post-target; 2080 ms in Figure 3). Finally, *unrelated* target words were reliably decodable 220 ms after target onset (2220 ms in Figure 3).

These decoding findings suggest that accurate prediction of the target words resulted in anticipation of both semantic and sub-lexical visual features. We infer pre-activation of word-length features because reliable decoding of these features was found at the onset of the target word before any information about word-length was available. For related target words that were not accurately predicted, there was evidence of anticipation of semantic features as well, presumably because there was overlap of semantic features between the predicted and received target words. For example, after reading the prime word circus, participants may have predicted acrobat instead of clown, because it is approximately equally associated to "circus". However, because the semantic features in this case are related but not identical, the exact features of the target words still need to be retrieved when it is read, evident from reliable decoding of concreteness after target word onset. Reliable decoding of word-length features for the related but unpredicted targets was found 80 ms after target word onset, consistent with ERP findings (Dufau et al., 2015) and MEG findings showing word length effects in ventral occipito-temporal cortex around 100ms (Wydell et al., 2003; see also Hsu et al., 2011, for MEG evidence of character complexity in Chinese around 100ms). Finally, in the unrelated condition, the predicted target words never overlapped with the actual target word, and the semantic and visual features of the target words could only be retrieved after it was presented. Interestingly, visual features in the unrelated condition were not decodable until 200ms after target word onset. We suggest that this delay could be due to the initial disconfirmation of the prediction after presentation of the target word, leading to delayed visual word recognition.

Significant above chance decoding for concreteness and visual word forms was found for the primes, starting at 80ms for word length and 220ms for concreteness, indicating the earliest reliable classification of these features likely uncontaminated by anticipation of these features for

the target words. Figure 4 illustrates clusters of reliable, significantly above-chance decoding for concreteness and word length of the prime words. Exact epochs are reported in Table 3. When target words were predicted, prime word concreteness was reliably decoded in a period spanning 300 – 820 ms after prime onset. When target words were unpredicted, both when related and unrelated, there was a much longer overall period during which prime word concreteness was reliably decoded. Also, we found above-chance decoding accuracy after target onset for unpredicted related words. We will address possible explanations for these findings in the discussion. Word length was reliably decoded for clusters that concentrated in the period between 80 - 860 ms after the prime onset. Additionally, when the target word was *unrelated*, prime word length was decodable 60 – 240 ms post target (*2060 – 2240 ms in Figure 4*). This suggests that prediction failure leads to partial reconsideration of the prime in working memory and is consistent with our findings of delayed decoding of target word length features.
Although no direct statistical analyses were done to compare the periods of reliable decoding between prime and target words, the overall decoding patterns appear to suggest that target word decoding was truly classifying target word features. This was evident from the earlier onset of reliable decoding of concreteness features for the primes in the first decoding analyses (220ms post prime for prime concreteness vs 340 ms post prime for target concreteness), and in the second analysis of the primes that was done contingent on the subsequent prediction accuracy

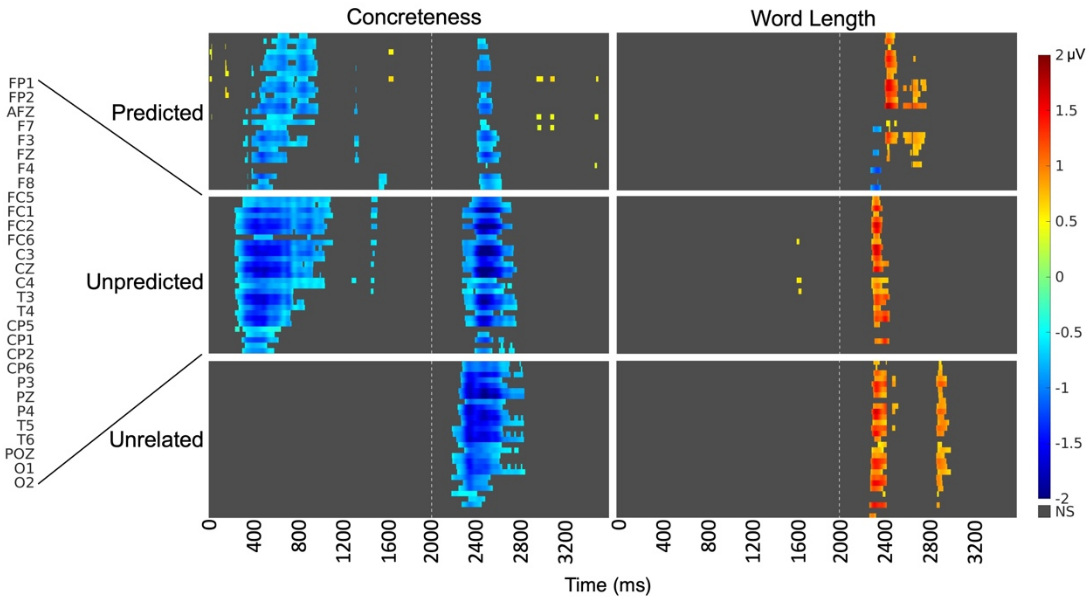

Figure 5. Shows difference wave ERP results (μV) from target concreteness (concrete minus abstract) and word length (long minus short) for each condition. Depicted cluster t test sums were less than the 0.5th percentile of the null distribution or greater than the 99.5th percentile of the null distribution. Darkened areas indicate no significant clusters of differences detected within that spatio-temporal range (see supplemental figure S3 for unmasked plots). The dashed white line indicates target word onset (2000 ms).

and relatedness of the targets. In these analyses, reliable decoding of concreteness features was found for the primes in a period between 300 - 820 ms after prime onset, reliable decoding of this feature for target words prior to target word onset spanned a period between 1640 to 960 ms pre-target (340 – 960 ms after prime onset). This suggests that pre-activation of concreteness features for the target appeared later and lasted longer than activation of these features following the prime

3.3 Results of mass-univariate ERP analysis of the features

Figures 5 and 6 show the ERP effects of concreteness and length features for the target and the prime words, respectively. Exact epochs and member electrodes are reported in Table 4. These effects were computed using ERP permutation cluster analyses (Maris & Oostenveld, 2007). As can be seen in Figure 5, the results of the cluster permutation *t* tests showed significant effects of concreteness for predicted target words and for related target words that were not predicted prior to target word onset. These effects are primarily concentrated 1852 – 888 ms prior to target onset (148 – 1112 ms in Figure 5). No effect of concreteness was found for unpredicted target words in the unrelated condition prior to target word onset. Given that participants were always attempting to predict the target words, these results indicate that semantic features of the target word were activated prior to the target word onset when successfully predicted or related to the prime. Effects of concreteness for the unrelated target words were only found in the 180 – 840 ms epoch after target onset (2180 – 2840 ms in Figure 5). There was also evidence for effects of concreteness after target onset for related target words – both predicted and unpredicted. For unpredicted related target words, this provides further support for the idea that when the target was related but not identical to the prediction, there was still concreteness information that needed to be processed. In the case of predicted words, this effect appeared to be greatly reduced. This suggests that anticipation of concreteness features of words does not prevent of these features during target word processing, but greatly reduced the need to do so.

In all three conditions, significant differences between long and short target words occurred almost exclusively after target onset, concentrated between 272 and 1000 ms (2272 – 3000 ms in Figure 5). This effect of word length is much is later than was found in previous studies and later than we observed in the decoding analysis. Perhaps the effects of word length in the ERP results were partially masked by the huge N400 prediction effects that followed them. features. The one unexpected pre-target difference found was a small significant cluster in the unpredicted condition (1616 – 1656 ms). Due to the small size and timing of this cluster, it is difficult to determine if this is a real effect which the classifier missed, or a spurious effect detected by the mass univariate analyses. For now, we prefer to err on the side of the latter.

As Figure 6 illustrates, significant differences between concrete and abstract prime words for all three conditions are concentrated primarily 240 – 1000 ms after prime word onset. However, similar to the decoding analyses, we observe some effects post-target which will be expanded upon in the discussion. Results of word length for the primes were found for all three conditions and for all primes without separating by conditions, but these word length effect occur later than prior studies (e.g., Dufeau et al, 2015). In the unpredicted condition the effects of word length of the prime were observed across the epoch from prime to target word onset, which is also an unexpected finding. Furthermore, reliable effects of word length were observed after target onset for the predicted target word and for the unrelated target words, but in the latter case the ERPs were more negative to the long words, which is contrary to the effects of word length found in prior studies (Dufeau et al, 2015). The results of this analysis of the effects of word

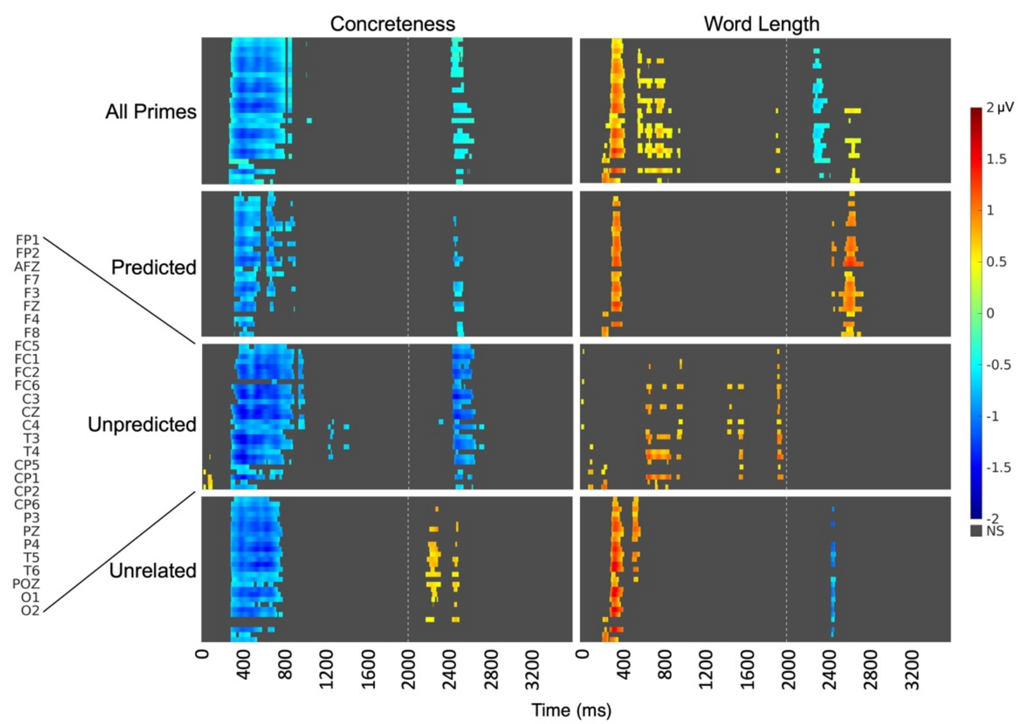

Figure 6. Shows difference wave ERP results (μV) from prime concreteness (concrete minus abstract) and word length (long minus short) for each condition. Depicted cluster t test sums were less than the 0.5th percentile of the null distribution or greater than the 99.5th percentile of the null distribution (see supplemental figure S4 for unmasked plots). The dashed white line indicates target word onset (2000 ms).

length of the primes contingent upon accurate prediction and relatedness of the targets is inconsistent with our decoding results.

Summarized in Table 5 is a post-hoc cluster analysis on the paired prime and target difference waves within each the conditions. In addition to significant cluster times and electrode membership, this table reports the μV differences between the target and prime difference waves (target – prime). The direction of this difference indicates which word had a greater feature effect (positive indicates target was greater). Significant clusters which overlapped with significant differences in either of the prior mass univariate analyses are reported in bold. Significant differences between ERP responses to primes and targets are observed during times in which significantly different clusters in their respective ERP cluster analyses were found both for predicted (340 – 808 ms) and unpredicted but related target words (212 – 924 ms). These post-hoc results provide further evidence that our original analyses were not driven by correlations of concreteness and word length features between primes and targets.

Table 4. Features - Significant differences from ERP cluster analyses

Word	Feature	Condition	Cluster	Difference (µV)		Times (ms)		Electrodes
				Min	Max	Begin	End	
Target	Concreteness	Predicted	1	0.3	0.5	8	24	F3 F7 FC5 T3 AFZ
			2	0.4	0.6	148	176	F4 F8 FC2 FC6 AFZ FZ
			3	-0.7	-0.4	308	348	C4 T4 CP2 CP6 P4 O2 PZ POZ
			4	-1.4	-0.4	384	980	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T6 O1 O2 AFZ FZ CZ PZ POZ
			5	-0.9	-0.5	1308	1344	F4 FC2 C4 CP2 CP6 P4 CZ PZ
			6	-0.7	-0.5	1528	1596	O1 O2 POZ
			7	0.4	0.6	1616	1656	F7 FC5
			8	-1.4	-0.4	2408	2628	FP2 F3 F4 FC1 FC2 FC6 C3 C4 T4 CP1 CP2 CP5 CP6 P3 P4 T5 O1 O2 AFZ FZ CZ PZ POZ
			9	0.3	0.5	2944	3000	FC5 T3 CP5
			10	0.4	0.7	3068	3104	FC5 T3 CP5
			11	0.3	0.5	3468	3496	FC5 T3 T5
		Unpredicted	1	-1.8	-0.3	232	1112	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
			2	-0.5	-0.5	1284	1320	T3
			3	-0.8	-0.4	1456	1512	FP1 F7 FC2 FC5 C3 T3 CP5 AFZ FZ
			4	-2.5	-0.4	2276	2768	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T6 O1 O2 AFZ FZ CZ PZ POZ
		Unrelated	1	-0.8	-0.4	2180	2220	C3 CP1 CP5 P3 T5 CZ
			2	-2	-0.4	2224	2740	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 AFZ FZ CZ PZ POZ
			3	-1	-0.5	2744	2800	FP1 F4 FC2 FC6 C4 T4 CP2 CP6 AFZ FZ CZ

			4	-1	-0.4	2804	2840	FP2 F4 FC2 FC6 C4 CP2 CP6 AFZ FZ CZ
	Length	Predicted	1	-1.3	-0.6	2280	2372	CP5 P3 T5 O1 O2 POZ
			2	0.5	1.8	2408	2648	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 T4 CP1 CP2 CP5 CP6 P3 P4 AFZ FZ CZ PZ
			3	0.6	1.2	2660	2780	FC1 FC2 FC6 C3 CP1 CP2 CP6 P3 P4 FZ CZ
		Unpredicted	1	0.5	0.6	1616	1656	FC5 T3 CP5
			2	0.6	1.7	2288	2448	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 T3 T4 CP1 CP2 CP5 CP6 P3 P4 AFZ FZ CZ PZ POZ
		Unrelated	1	0.4	1.7	2272	2424	FP1 FP2 F3 F4 F7 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 O2 AFZ FZ CZ PZ POZ
			2	0.7	1	2476	2528	F3 F7 FC1 FC5 C3
			3	0.5	1.2	2872	3000	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 AFZ FZ CZ PZ POZ
Prime	Concreteness	Predicted	1	-1.4	-0.2	264	808	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
			1	-1.3	-0.4	316	568	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T6 O1 O2 AFZ FZ CZ PZ POZ
			2	-1.2	-0.5	628	908	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P4 AFZ FZ CZ PZ
			2	-0.8	-0.3	836	872	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T4 CP2 CP6 P4 AFZ FZ CZ
			3	-0.4	-0.3	1016	1064	FP2 F8 T4
			4	-0.6	-0.3	2416	2640	FP1 FP2 F3 F4 F7 F8 FC1 FC2 C4 T4 CP1 CP2 CP6 P4 O2 AFZ FZ CZ PZ POZ

			3	-0.9	-0.4	2440	2468	F4 FC1 FC2 C3 C4 CP1 CP2 CP6 P4 FZ CZ PZ
			4	-0.9	-0.4	2472	2540	FC1 C3 C4 CP1 CP2 CP6 P4 O1 O2 CZ PZ POZ
		Unpredicted	1	0.4	0.5	12	32	T6 O2
			2	0.5	0.7	56	96	O1 O2 PZ POZ
			3	-1.7	-0.4	276	888	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
			4	-1	-0.5	932	992	FP2 F3 F4 F7 FC1 FC2 FC5 FC6 C3 C4 T3 T4 AFZ FZ
			5	-0.8	-0.6	1220	1272	T3 T4 CP1 CP5 CP6 P3 PZ
			6	-0.7	-0.6	1372	1420	T4 CP6
			7	-0.7	-0.6	2292	2332	T3
			8	-1.5	-0.5	2424	2728	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T6 O2 AFZ FZ CZ PZ POZ
		Unrelated	1	-1.5	-0.4	284	784	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 O1 O2 AFZ FZ CZ PZ POZ
			2	0.4	0.7	2172	2204	FC1 C3 T3 CP5 T5
			3	0.4	0.9	2208	2316	F4 FC1 FC2 FC6 C3 C4 T3 CP1 CP2 CP5 CP6 P3 T5 AFZ CZ
			4	0.5	0.8	2428	2460	FC1 C3 T3 CP1 CP5 P3 T5 CZ
			5	0.4	0.8	2464	2496	F4 FC1 C3 T3 CP1 CP5 P3 T5 FZ CZ
	Length	Predicted	1	0.3	1.4	212	436	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 O1 O2 AFZ FZ CZ PZ POZ
			1	0.6	1.1	212	272	T5 O1 O2
			2	0.5	1.3	304	408	FP2 F3 F4 F8 FC1 FC2 FC5 FC6 C3 C4 T4 CP1 CP2 CP5 CP6 P3 P4 AFZ FZ CZ PZ POZ
			2	0.3	0.6	556	636	F3 F4 F7 F8 FC1 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 P3 AFZ PZ POZ
			3	0.3	0.6	640	720	F3 FC1 FC5 C3 C4 CP1 CP2 CP5 CP6 P3 P4 FZ CZ PZ POZ

			4	0.3	0.9	724	892	F3 FC1 FC5 C3 C4 CP1 CP2 CP5 P3 P4 O2 CZ PZ POZ
			5	0.4	0.5	940	968	CP1 P4 PZ POZ
			6	0.4	0.6	1900	1936	C4 CP2 PZ POZ
			7	-0.7	-0.3	2256	2424	F3 FC1 FC2 FC5 FC6 C3 C4 CP1 CP2 CP5 CP6 P3 P4 T5 O1 AFZ FZ CZ PZ
			3	0.5	1.3	2440	2744	FP1 F3 F4 FC1 FC2 FC6 C3 C4 T4 CP1 CP2 CP5 CP6 P3 P4 T5 O1 O2 AFZ FZ CZ PZ POZ
			8	0.3	0.6	2564	2624	C4 T4 CP6 P4 O2
			9	0.3	0.6	2628	2720	C4 CP6 P3 P4 O1 O2 PZ POZ
		Unpredicted	1	0.4	0.5	12	36	FP2 F8 FC6 T4
			2	0.5	0.9	80	124	CP6 P4 T6 O2
			3	0.5	1.2	208	260	T5 T6 O1 O2
			4	0.5	1.2	636	880	F3 FC1 FC5 C3 CP1 CP2 CP5 P3 P4 O1 O2 CZ PZ POZ
			5	0.5	0.9	936	1000	F3 F7 FC5 C3 T3 CP1 CP5 PZ POZ
			6	0.5	0.9	1428	1584	FC5 C3 T3 CP1 CP5 P3 T5 PZ POZ
			7	0.5	1.2	1908	1968	FP2 F3 F4 FC1 FC2 FC5 C3 T3 CP1 CP2 CP5 P3 T5 CZ PZ POZ
		Unrelated	1	0.5	1.7	216	424	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 O1 O2 AFZ FZ CZ PZ POZ
			2	0.6	1.1	504	584	FP1 FP2 F3 F4 F7 F8 FC5 FC6 C3 C4 T4 AFZ FZ
			3	-1.2	-0.5	2432	2476	FC1 FC2 FC6 C3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 O1 AFZ FZ CZ PZ POZ

Table 4. Summary of all significantly different (cluster t test sums less than the 0.5th percentile of the null distribution or greater than the 99.5th percentile of the null distribution) ERP cluster start times, stop times, and member electrodes for the word feature conditions. For concreteness, a positive feature difference (µV) indicates that abstract words had greater amplitude than concrete words. For word length, a positive feature difference indicates long words had greater amplitude than short words. Target word onset is 2000 ms.

Table 5. Features - Significant differences from ERP cluster analyses between target and prime words

Feature	Condition	Cluster	Target Difference (μV)		Prime Difference (μV)		Target - Prime (μV)		Times (ms)		Electrodes
			Min	Max	Min	Max	Min	Max	Begin	End	
Concreteness	Predicted	1	-0.2	0.2	-0.3	0.1	-0.5	-0.3	**24**	**56**	CP1 CP2 CP6 P3 P4 T6 O1 O2 PZ POZ
		2	-0.4	0	-0.4	0	-0.6	-0.3	**340**	**420**	FP1 FP2 F3 F8 FC1 FC2 FC5 C3 CP1 AFZ FZ CZ
		3	-0.2	0.3	0	0.2	0.4	0.7	**520**	**576**	T5 O1 O2 POZ
		4	0.4	0.4	0.7	0.8	0.4	0.5	**796**	**808**	F7 FC5
		5	-0.2	0.6	-0.2	0.5	0.4	0.5	1080	1132	O1 O2 POZ
		6	-0.3	0.9	-0.3	0.7	0.4	0.7	**1296**	**1396**	FC1 FC2 FC6 C4 T4 CP1 CP2 CP5 CP6 P3 P4 T5 O1 O2 CZ PZ POZ
		7	0	0.5	-0.3	0.1	0.4	0.6	**1444**	**1584**	O1 O2 POZ
		8	-0.3	0.3	-0.4	0.3	0.3	0.6	**2476**	**2492**	FC6 C4 CP1 CP2 CP6 P3 P4 T6 O1 O2 PZ POZ
		9	-0.1	0.1	-0.1	0.3	0.4	0.6	**2520**	**2536**	FC6 C4 CP2 P3 P4 O1 O2 PZ POZ
		10	-0.4	0.1	-0.8	-0.1	0.4	0.5	2700	2716	T6 O2
		11	0.2	0.3	0	0.3	0.3	0.4	**3180**	**3192**	O1 O2 POZ
	Unpredicted	1	0	0.3	0	0.3	0.3	0.6	**212**	**260**	FP1 F4 F7 F8 FC1 FC2 FC6 AFZ FZ CZ
		2	0	0.3	0	0.2	0.4	0.6	**288**	**316**	F4 F8 FC6 AFZ FZ
		3	-0.3	0.1	-0.5	0	0.3	0.6	**464**	**492**	F3 F4 FC1 FC2 FC6 C3 CP1 CP2 CP5 P3 AFZ FZ CZ
		4	0.6	1.5	0.6	1.7	-0.8	-0.4	**744**	**788**	P4 T6 O2 PZ POZ
		5	1.2	1.6	1.1	1.3	-0.7	-0.4	**844**	**864**	PZ POZ
		6	0.7	1.7	0.5	1.4	0.3	0.7	**888**	**924**	FP1 FP2 F3 F4 F8 FC1 FC2 FC5 FC6 AFZ FZ
		7	-0.3	1.2	-0.2	1.1	-1.1	-0.4	**1068**	**1196**	CP1 P3 P4 T5 T6 O1 O2 PZ POZ
		8	-0.4	0.4	-0.1	0.5	-0.8	-0.4	**1320**	**1388**	CP6 P3 P4 T6 O2 PZ POZ
		9	-0.4	0.1	0.3	0.9	-1.1	-0.4	**1504**	**1568**	P3 P4 T6 O1 O2 PZ POZ
		10	0.2	0.5	0.4	0.6	0.6	0.7	1932	1956	FC1 FZ CZ
		11	0.1	0.5	-0.3	0	0.7	0.9	**2288**	**2300**	FC1 AFZ FZ CZ
		12	-0.2	0.5	-0.2	0.3	0.4	0.8	**2388**	**2400**	C4 CP1 CP2 P4 T6 O2 CZ PZ
		13	-0.2	0.7	-0.5	0.5	0.4	1.2	**2420**	**2672**	FP1 F3 FC1 FC2 FC5 FC6 C3 C4 CP1 CP2 CP6 P4 T6 O1 O2 AFZ FZ CZ PZ
		14	0.3	0.5	0.2	0.3	0.5	0.7	3260	3276	CP1 CZ

	Unrelated	1	-0.4	0.3	-0.5	1	-1.9	-0.4	284	764	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T6 O1 O2 AFZ FZ CZ PZ POZ
		2	0	0.3	-0.1	0.7	0.5	0.9	964	1008	CP5 P3 T5 O1
		3	-0.1	0.1	-0.5	-0.1	0.5	0.8	1460	1520	C3 CP5 P3 T5 O1
		4	-0.1	0.1	-0.3	0	0.6	0.8	1536	1560	F3 FC1 FC5 C3 T3 CP5 P3 T5
		5	-0.1	0.1	-0.3	0	0.6	0.9	1572	1608	FC5 C3 T3 CP5 P3 T5 O1
		6	0.2	0.3	-0.2	-0.1	0.6	0.8	1688	1716	CP5 T5
		7	0	0.2	-0.5	-0.4	0.4	0.7	**2144**	**2164**	FC5 C3 T3 CP5 T5
		8	-0.3	1.9	-0.9	0.3	0.4	2.7	**2168**	**2860**	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 AFZ FZ CZ PZ POZ
		9	1.3	1.5	-0.2	-0.1	0.6	0.7	2932	2964	FC6 C4
		10	0.7	1	0	0.1	0.6	0.7	3056	3068	C4 CP6
		11	-0.1	0.4	-0.5	-0.2	0.5	0.7	3180	3212	CP5 P3 T5 O1
Length	Predicted	1	0	0.2	0	0.1	-0.6	-0.4	32	44	F4 F8 FC2 FC6
		2	-0.1	0	-0.3	-0.2	0.4	0.5	136	148	FC5 C3 CP5
		3	-0.6	0.2	-0.4	0.3	0.4	1.4	208	308	T3 CP5 P3 P4 T5 O1 O2 PZ POZ
		4	-0.2	0.3	-0.3	0.5	0.4	1.3	336	424	F3 F4 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 AFZ FZ CZ PZ POZ
		5	-0.4	-0.1	-0.4	0.2	-0.9	-0.6	460	488	T6 O2
		6	-0.1	0.3	-0.7	-0.4	-0.7	-0.5	676	708	O1 O2
		7	-0.6	-0.2	-0.5	-0.4	-0.5	-0.5	1056	1072	T4 CP6
		8	-0.2	0	0	0.5	-0.8	-0.5	1320	1356	FP1 F3 F7 FC1 FC2 C3 AFZ FZ
		9	0.1	0.3	0.2	0.4	-0.8	-0.6	**2156**	**2204**	FP1 FP2 F7
		10	-0.2	0.1	0.3	0.6	-0.9	-0.6	**2412**	**2456**	FP1 FP2 F3 F7 FC1 FC5 C3
	Unpredicted	1	-0.1	0.1	-0.3	0.3	0.3	0.7	76	120	T6 O2
		2	0.1	0.2	-0.2	-0.1	-0.5	-0.5	156	168	CP1 P3
		3	-0.2	0.3	-0.2	0.1	0.5	1.1	220	312	F4 F8 FC2 FC5 FC6 C3 C4 T4 CP1 CP2 CP5 CP6 P3 P4 T6 O1 O2 PZ POZ
		4	-0.6	0.2	-0.4	0.3	0.5	1.3	316	420	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T6 O2 AFZ FZ CZ PZ POZ
		5	-0.3	0	-1.1	0.2	0.5	0.9	652	692	CP1 CP2 P3 P4 O1 O2 CZ PZ POZ
		6	-0.4	0.5	-1.1	0.1	0.6	1.1	744	876	FC5 C3 CP1 CP2 P4 O1 O2 PZ POZ

		7	-0.4	0	-0.7	-0.2	0.4	0.7	1096	1144	FC2 C4 CP2 CP6
		8	0	0.1	-0.8	-0.3	0.5	0.8	1196	1208	F3 F7 FC5 C3 CP5
		9	-0.1	0.2	-1.1	-0.3	0.4	0.9	1228	1256	F3 F7 FC5 C3 T3 CP1 CP2 CP5 PZ POZ
		10	-0.2	0	-0.7	-0.3	0.4	0.9	1276	1372	F3 F7 FC5 C3 T3
		11	-0.8	0.1	-1.2	0	-1.7	-0.4	**2284**	**2464**	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 T3 T4 CP1 CP2 CP5 CP6 P3 T5 O1 AFZ FZ CZ PZ POZ
		12	-0.4	0.4	-0.1	0.2	0.5	0.9	3092	3148	P4 O1 O2 POZ
	Unrelated	1	-0.2	0.1	-0.2	0.2	-1.2	-0.4	56	120	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T4 CP1 CP2 CP5 CP6 P3 P4 O1 O2 AFZ FZ CZ PZ POZ
		2	-0.1	0.2	-0.1	0.1	0.4	1.5	212	252	CP5 P3 P4 T5 O1 O2 PZ POZ
		3	-0.5	0.3	-0.3	0.2	0.5	2	272	420	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
		4	-0.3	0.1	-0.2	0.2	0.6	1.1	548	580	FP1 FP2 F4 F8 FC6 C4 AFZ FZ
		5	0	0.1	-0.6	-0.5	0.9	1	1580	1612	C4
		6	-0.8	0.4	-0.7	0.6	-2.4	-0.6	**2268**	**2516**	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
		7	-1.7	-0.8	0.3	0.6	-1.2	-0.7	**2704**	**2776**	F3 FC1 FC5 C3 CP1 CP5
		8	-1.3	-0.3	0.1	1	-1.2	-0.6	**2780**	**2848**	F3 F7 FC1 FC5 C3 T3 CP1 CP2 CP5 P3 T5 CZ PZ POZ
		9	-0.9	0.1	-0.3	0.7	-1.3	-0.6	**2880**	**3004**	FP2 F3 F7 F8 FC1 FC2 FC5 FC6 C3 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 AFZ FZ CZ PZ

Table 5. Summary of all significantly different ERP cluster start times, stop times, and member electrodes for comparison between target and prime word feature effects. Clusters which overlap with significant clusters from the original target or prime word analyses are in ***bold.*** *The max and min difference values (Target – Prime) indicates direction of the difference; positive values indicate that target word had a larger difference between the features and negative values indicates that the prime word had larger differences. Target word onset is 2000 ms.*

4. Discussion

The present study aimed to investigate the time course of top-down pre-activations during predictive processing using a combination of machine-learning EEG decoding and mass univariate ERP cluster analyses. Our prediction task allowed us to distinguish between trials for which participants predicted the exact target word from those in which the participants were unsuccessful at predicting a related target word or encountered a completely unrelated word. This allowed us to identify whether specific linguistic and visual features were anticipated, and if successful prediction resulted in a reduction or absence of evidence of prediction error. In agreement with our predictions, when words were successfully predicted, concreteness and word length features were pre-activated, and our ERP analyses suggest a reduction in additional processing after the target word was encountered. It is unclear whether word length pre-activation resulted in a reduction of prediction error because reliable decoding began at target onset and continued well into the period during which the expected prediction error measure, the N250 ERP component, would have been observed. However, these results provide compelling evidence for predictive coding accounts of language comprehension (e.g., Kuperberg & Jaeger, 2016) during visual word recognition.

Even though we did not experimentally manipulate concreteness and sub-lexical word length, the findings still showed reliable decoding and identified significant differences in mass univariate ERP analysis which is promising for studies that use more naturalistic stimuli. However, to gain a more detailed view of the temporal dynamics of pre-activation of difference linguistic and sub-linguistic features, explicit manipulation may prove beneficial.

Another interesting observation was evidence from decoding of extended activation of semantic features when target words were not accurately predicted – whether it is related or not. One limitation of the paradigm, as implemented here, is that we do not know what exactly happens during an unsuccessful prediction. One possibility is a prediction that is different from the target, but another possibility is that the participant failed to make a prediction at all. It is not unreasonable to assume the latter happens with some frequency. This could result in the prime word remaining active in working memory longer as the participant continues to think of a prediction. Future studies could further explore this by making predictions more explicit.

5. Conclusion

In summary, the present experiment utilized a combination of EEG decoding and mass univariate ERP analysis to uncover the time course of the anticipation of semantic – such as concreteness – and sub-lexical – such as word length – information, and that successfully anticipated information largely prevents the need for further processing as indicated by the absence or reduction of prediction error. Our findings are consistent with predictive coding theories of language, showing that higher level semantic features are anticipated prior to low level visual features during word recognition, and that accurately anticipated features reduce prediction error (Kuperberg & Jaeger, 2016). In future studies we will examine anticipation of other features that are relevant to word recognition, such as visual complexity of the words, orthographic neighborhood, phonological neighborhood, and lexical frequency.

Appendix

The appendix contains supplemental tables and figures that could not be included within the main paper. Figure S1 shows ERP waveforms for all electrode sites. Figures S2 – S4 depict the unmasked ERP difference waves for all the conditions of the study: Figure S2 shows the main effects waveforms, Figure S3 shows target word waveforms, and Figure S4 shows the prime word waveforms. Tables S1 and S2 show the artifact rejection summaries for word length and concreteness, respectively.

Subject	Table S1. Length Artifact Rejection											
	Target											
	Predicted				Unpredicted				Unrelated			
	Acc. Long	Rej. Long	Acc. Short	Rej. Short	Acc. Long	Rej. Long	Acc. Short	Rej. Short	Acc. Long	Rej. Long	Acc. Short	Rej. Short
1	54	2	61	2	38	2	71	3	44	2	59	2
2	49	1	61	1	45	1	73	2	48	0	65	0
3	37	0	53	2	53	1	80	2	46	0	63	1
4	46	0	62	0	50	0	75	0	47	0	65	0
5	45	1	57	1	50	0	79	0	48	0	63	1
6	46	1	61	0	54	0	68	1	42	1	70	0
7	55	0	63	0	46	0	68	0	43	0	71	0
8	25	3	37	1	46	2	70	0	33	0	59	1
9	26	0	35	0	74	1	96	0	43	0	68	1
10	39	0	39	1	41	0	61	1	38	2	53	1
11	56	1	90	0	33	0	46	0	52	0	65	1

12	35	2	51	2	53	1	83	0	52	1	66	0
13	39	0	66	0	52	0	70	0	53	0	66	0
14	46	0	64	0	45	0	71	0	53	0	66	0
15	14	4	36	6	55	15	74	16	42	7	52	10
16	48	0	77	0	47	1	60	0	48	0	63	0
17	44	0	68	0	51	0	68	1	48	0	65	0
18	52	0	60	0	40	0	77	0	47	0	64	0
19	44	0	59	0	46	0	74	0	47	0	61	0
20	40	0	70	0	56	0	67	0	48	0	64	0
21	75	0	83	1	26	0	47	0	41	0	68	2
22	49	2	54	5	47	1	63	6	42	1	65	4
23	32	0	36	0	49	0	62	0	34	0	54	0
24	41	0	67	0	60	0	63	1	43	0	70	1
25	45	0	56	0	56	0	75	0	42	0	70	1
26	38	0	67	0	53	0	69	0	53	0	66	0
27	44	1	72	2	45	1	60	2	53	0	65	1
28	40	0	53	0	50	0	83	0	53	0	65	1
29	44	0	85	1	47	0	50	0	53	0	66	0
30	47	3	64	1	39	2	69	2	53	0	66	0
31	49	0	61	2	47	0	74	0	47	1	64	1
32	56	0	76	0	40	0	61	0	46	0	65	0
33	45	0	64	0	51	0	73	0	48	0	65	0
34	47	2	69	2	43	4	63	3	45	3	63	1
35	51	1	82	0	44	0	55	0	48	0	65	0
36	97	1	119	9	3	0	2	1	42	1	63	7

37	45	5	53	3	50	1	73	2	36	7	67	4
38	54	0	70	0	47	0	61	0	43	0	70	1
39	21	4	36	3	66	9	79	13	35	8	64	7
40	67	0	78	0	32	0	52	0	42	0	71	0
41	54	0	80	0	21	0	21	1	40	1	52	1
42	28	23	48	42	24	16	28	18	30	23	37	29
43	61	0	78	0	30	0	58	0	53	0	66	0
44	59	1	88	2	31	0	46	0	51	2	61	5
45	70	7	99	7	13	1	29	1	52	1	65	0
46	33	1	49	0	57	0	87	0	51	1	65	0
47	37	0	57	0	5	0	8	0	20	0	35	0
48	57	0	76	0	34	0	60	0	53	0	66	0
49	58	0	68	2	41	2	60	1	41	0	70	0
50	52	0	86	0	38	0	49	0	53	0	66	0
51	19	9	34	5	35	22	50	37	28	21	39	23
52	60	0	81	0	41	0	50	0	42	0	70	0
53	52	0	70	0	43	0	67	0	48	0	65	0
54	53	1	59	0	47	0	72	0	43	0	71	0
55	45	0	61	0	46	0	72	1	52	1	65	1
56	50	0	57	1	51	0	73	0	43	0	70	1
Subject	Prime											
	Predicted				Unpredicted				Unrelated			
	Acc. Long	Rej. Long	Acc. Short	Rej. Short	Acc. Long	Rej. Long	Acc. Short	Rej. Short	Acc. Long	Rej. Long	Acc. Short	Rej. Short

1	81	5	48	2	79	2	30	3	78	3	40	1
2	85	0	42	1	81	2	40	1	87	0	42	0
3	66	1	38	1	94	2	41	1	81	4	39	2
4	82	0	45	0	86	0	39	0	86	0	41	0
5	76	1	46	1	91	0	37	0	85	1	42	0
6	76	2	43	0	92	0	43	1	84	0	37	1
7	77	0	45	0	93	0	43	0	85	0	38	0
8	45	2	21	5	83	3	48	2	67	1	29	0
9	42	0	21	0	126	2	66	0	81	1	37	0
10	71	1	43	0	95	1	41	0	86	1	39	3
11	111	1	57	0	60	0	23	0	81	1	46	0
12	67	3	37	0	101	1	42	1	82	1	46	0
13	75	0	43	0	97	0	37	0	83	0	46	0
14	73	0	39	0	98	0	41	0	82	0	46	0
15	39	8	17	4	102	21	40	14	64	16	37	7
16	94	0	48	0	74	0	36	0	86	0	42	0
17	80	1	40	0	86	0	43	1	87	0	42	0
18	87	0	43	0	76	0	39	0	84	0	42	0
19	72	0	46	0	88	0	35	0	82	0	42	0
20	79	0	39	0	89	0	45	0	86	0	42	0
21	119	1	56	1	50	0	31	0	82	2	37	0
22	80	4	38	2	79	2	43	4	80	4	36	1
23	67	0	39	0	103	0	49	0	85	0	38	0
24	76	0	39	0	94	0	48	1	84	0	38	0
25	72	0	34	0	98	0	53	1	83	0	35	1

26	76	0	39	0	96	0	41	0	83	0	46	0
27	86	2	44	0	84	0	35	1	82	1	44	2
28	62	0	44	0	109	0	36	0	82	1	46	0
29	99	0	51	1	73	0	28	0	83	0	46	0
30	80	1	46	1	88	3	32	1	81	2	46	0
31	77	2	43	1	89	0	40	0	86	1	42	0
32	100	0	46	0	68	0	38	0	86	0	41	0
33	88	0	33	0	80	0	51	0	87	0	42	0
34	71	6	54	0	83	8	28	2	83	3	37	4
35	100	1	53	0	67	0	31	0	87	0	42	0
36	154	12	86	1	4	0	1	0	82	3	34	4
37	73	6	38	4	86	5	45	1	76	9	36	2
38	86	0	57	0	84	0	31	0	84	1	38	0
39	40	4	19	3	108	18	57	8	73	12	34	4
40	99	0	57	0	71	0	28	0	85	0	38	0
41	121	1	66	0	49	1	14	0	81	0	42	2
42	58	41	26	28	45	28	15	11	51	32	19	27
43	93	0	51	0	79	0	29	0	82	0	46	0
44	117	3	47	1	52	0	32	0	80	3	44	2
45	129	10	63	2	32	1	13	2	80	2	43	2
46	66	1	28	0	105	0	52	0	80	0	46	0
47	91	0	47	0	11	0	8	0	52	0	21	0
48	97	0	49	0	75	0	31	0	83	0	46	0
49	96	1	51	1	71	2	35	1	84	0	38	0
50	99	0	45	0	72	0	34	0	83	0	46	0

51	38	10	18	7	76	40	26	19	50	29	25	17
52	100	0	54	0	70	0	34	0	85	0	36	0
53	85	0	46	0	83	0	37	0	87	0	42	0
54	79	0	43	0	91	0	45	0	84	1	38	0
55	78	0	43	0	92	0	37	0	82	1	45	1
56	76	1	37	0	93	0	51	0	85	0	38	0

Table S1. Shows the artifact rejection summary for all word length conditions.

	Table S2. Concreteness Artifact Rejection											
	Target											
	Predicted				Unpredicted				Unrelated			
Subject	Acc. High	Rej. High	Acc. Low	Rej. Low	Acc. High	Rej. High	Acc. Low	Rej. Low	Acc. High	Rej. High	Acc. Low	Rej. Low
1	82	3	82	5	72	2	68	4	73	3	74	3
2	72	1	83	1	86	1	74	2	80	0	80	0
3	63	2	74	0	90	1	79	2	76	2	74	4
4	82	0	80	0	78	0	80	0	80	0	78	0
5	73	1	70	1	86	0	89	0	80	0	78	1
6	73	1	75	1	84	1	84	0	79	1	79	0
7	73	0	84	0	87	0	76	0	80	0	80	0
8	39	3	45	4	86	4	72	2	60	0	65	1
9	37	0	44	0	122	1	114	1	77	1	78	0
10	50	0	64	1	72	1	64	0	65	0	60	4
11	97	0	113	1	63	0	45	0	80	0	77	1

12	60	4	70	0	93	3	90	0	80	0	79	1
13	71	0	74	0	89	0	86	0	80	0	80	0
14	73	0	67	0	87	0	92	0	79	0	80	0
15	26	7	41	6	106	20	87	19	61	15	64	10
16	91	0	83	0	68	1	77	0	79	0	79	0
17	72	0	78	1	88	0	79	1	80	0	80	0
18	79	0	80	0	78	0	76	0	79	0	78	0
19	83	0	63	0	71	0	90	0	78	0	77	0
20	73	0	79	0	87	0	81	0	80	0	79	0
21	101	1	121	1	58	0	38	0	76	1	79	1
22	68	3	77	5	79	6	72	4	75	2	73	5
23	51	0	55	0	76	0	76	0	60	0	63	0
24	72	0	75	0	88	0	84	1	79	0	79	1
25	73	0	55	0	86	1	105	0	79	0	76	1
26	67	0	80	0	93	0	80	0	80	0	80	0
27	75	3	92	0	81	1	66	2	80	0	77	3
28	71	0	61	0	89	0	98	0	79	1	80	0
29	82	1	98	0	77	0	62	0	80	0	80	0
30	73	0	78	4	83	4	76	2	78	2	80	0
31	78	2	79	2	80	0	79	0	80	0	78	2
32	93	0	94	0	67	0	66	0	79	0	79	0
33	74	0	75	0	86	0	85	0	80	0	80	0
34	85	4	74	4	67	4	75	7	76	3	75	4
35	87	1	103	0	72	0	57	0	80	0	80	0
36	145	9	153	6	5	1	1	0	75	4	72	7

37	68	6	67	5	83	3	83	5	72	8	74	6
38	89	0	89	0	71	0	71	0	79	1	80	0
39	39	6	41	1	99	15	102	16	72	8	67	13
40	93	0	100	0	66	0	58	0	80	0	79	0
41	90	0	97	1	31	1	32	0	62	2	61	0
42	48	41	54	50	45	26	34	22	43	37	45	35
43	91	0	100	0	69	0	60	0	79	0	80	0
44	90	0	116	4	70	0	40	0	75	5	77	3
45	114	12	123	5	31	3	31	1	75	4	78	1
46	58	1	60	0	100	1	100	0	76	1	79	0
47	74	0	64	0	10	0	9	0	35	0	38	0
48	81	0	99	0	79	0	61	0	80	0	80	0
49	85	2	89	1	72	1	68	2	79	0	78	0
50	94	0	93	0	66	0	65	0	80	0	80	0
51	36	10	36	11	59	43	68	36	44	29	50	28
52	95	0	102	0	65	0	58	0	79	0	78	0
53	79	0	84	0	81	0	75	0	80	0	80	0
54	78	0	81	1	82	0	78	0	80	0	79	1
55	73	0	77	0	86	0	81	1	80	0	77	3
56	71	0	68	2	89	0	90	0	80	0	79	1
	Prime											
	Predicted				Unpredicted				Unrelated			
Subject	Acc. High	Rej. High	Acc. Low	Rej. Low	Acc. High	Rej. High	Acc. Low	Rej. Low	Acc. High	Rej. High	Acc. Low	Rej. Low

1	72	1	92	7	49	2	91	4	66	3	81	3
2	56	1	99	1	67	1	93	2	74	0	86	0
3	57	1	80	1	63	2	106	1	69	2	81	4
4	70	0	92	0	55	0	103	0	73	0	85	0
5	53	1	90	1	71	0	104	0	74	0	84	1
6	59	0	89	2	72	0	96	1	67	0	91	1
7	65	0	92	0	67	0	96	0	67	0	93	0
8	31	3	53	4	71	2	87	4	54	0	71	1
9	31	0	50	0	99	2	137	0	67	0	88	1
10	42	0	72	1	49	1	87	0	56	1	69	3
11	91	1	119	0	49	0	59	0	56	0	101	1
12	55	2	75	2	81	3	102	0	58	0	101	1
13	66	0	79	0	75	0	100	0	58	0	102	0
14	71	0	69	0	70	0	109	0	58	0	101	0
15	25	9	42	4	88	16	105	23	47	6	78	19
16	71	0	103	0	53	1	92	0	73	0	85	0
17	56	0	94	1	69	0	98	1	74	0	86	0
18	60	0	99	0	63	0	91	0	73	0	84	0
19	63	0	83	0	58	0	103	0	71	0	84	0
20	60	0	92	0	65	0	103	0	74	0	85	0
21	86	1	136	1	45	0	51	0	65	1	90	1
22	53	4	92	4	72	2	79	8	61	2	87	5
23	44	0	62	0	58	0	94	0	47	0	76	0
24	62	0	85	0	70	0	102	1	66	1	92	0
25	60	0	68	0	71	1	120	0	64	0	91	1

26	58	0	89	0	83	0	90	0	58	0	102	0
27	73	1	94	2	65	2	82	1	56	2	101	1
28	65	0	67	0	75	0	112	0	57	1	102	0
29	65	1	115	0	75	0	64	0	58	0	102	0
30	67	0	84	4	71	3	88	3	58	0	100	2
31	58	0	99	4	67	0	92	0	73	1	85	1
32	74	0	113	0	51	0	82	0	73	0	85	0
33	57	0	92	0	68	0	103	0	74	0	86	0
34	68	2	91	6	53	2	89	9	67	5	84	2
35	74	0	116	1	51	0	78	0	74	0	86	0
36	123	4	175	11	4	1	2	0	61	5	86	6
37	49	6	86	5	74	3	92	5	61	6	85	8
38	73	0	105	0	59	0	83	0	66	1	93	0
39	30	4	50	3	88	9	113	22	54	13	85	8
40	75	0	118	0	57	0	67	0	66	0	93	0
41	78	0	109	1	25	1	38	0	43	1	80	1
42	45	41	57	50	38	17	41	31	33	25	55	47
43	80	0	111	0	61	0	68	0	58	0	101	0
44	78	1	128	3	62	0	48	0	56	2	96	6
45	97	10	140	7	30	4	32	0	56	1	97	4
46	52	1	66	0	88	0	112	1	58	0	97	1
47	60	0	78	0	8	0	11	0	27	0	46	0
48	76	0	104	0	65	0	75	0	58	0	102	0
49	73	1	101	2	57	1	83	2	66	0	91	0
50	81	0	106	0	60	0	71	0	58	0	102	0

51	34	8	38	13	58	34	69	45	36	17	58	40
52	82	0	115	0	50	0	73	0	65	0	92	0
53	64	0	99	0	61	0	95	0	74	0	86	0
54	68	1	91	0	63	0	97	0	67	0	92	1
55	67	0	83	0	73	0	94	1	57	1	100	2
56	57	0	82	2	75	0	104	0	67	0	92	1

Table S2. Shows the artifact rejection summary for all concreteness conditions.

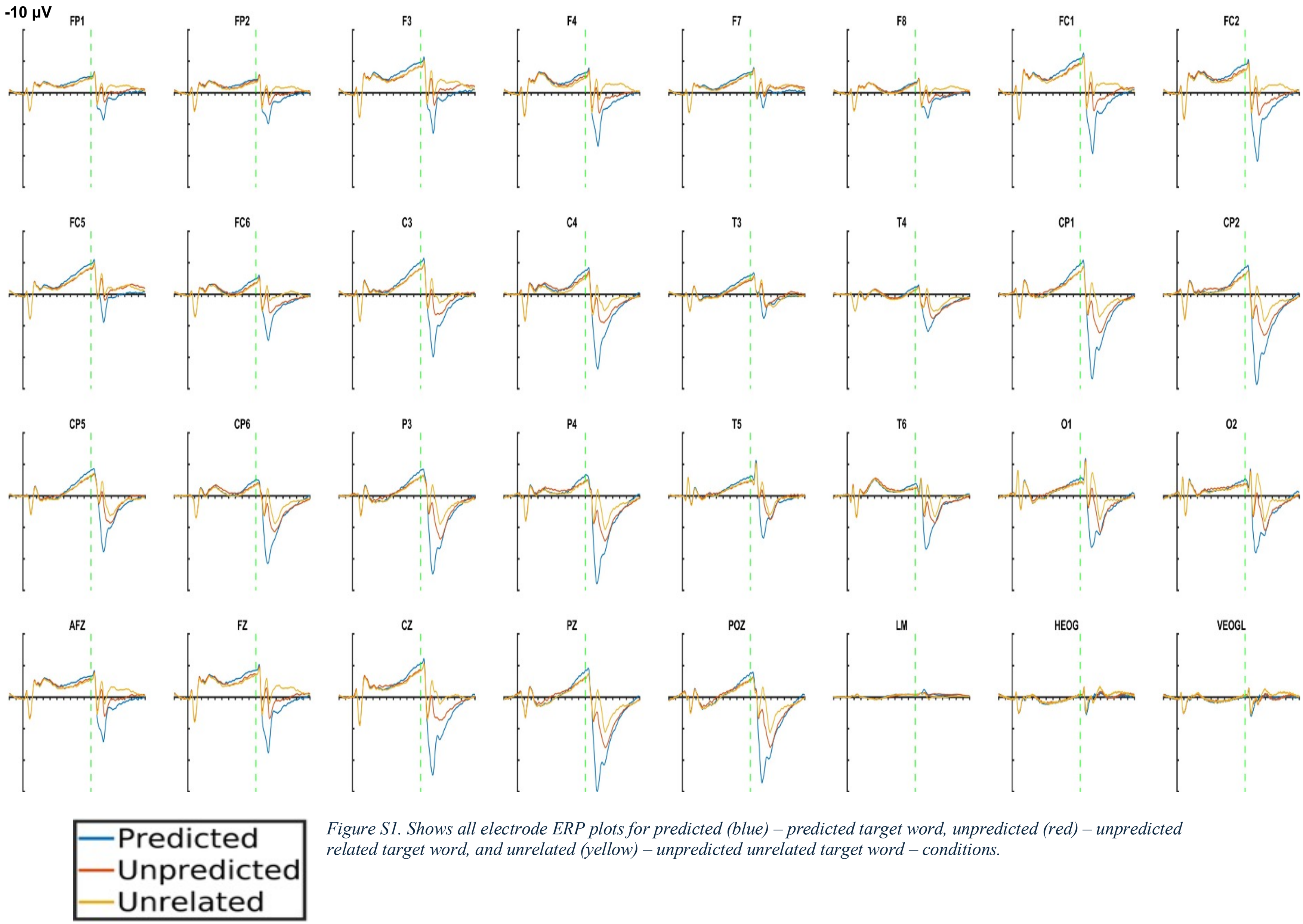

Figure S1. Shows all electrode ERP plots for predicted (blue) – predicted target word, unpredicted (red) – unpredicted related target word, and unrelated (yellow) – unpredicted unrelated target word – conditions.

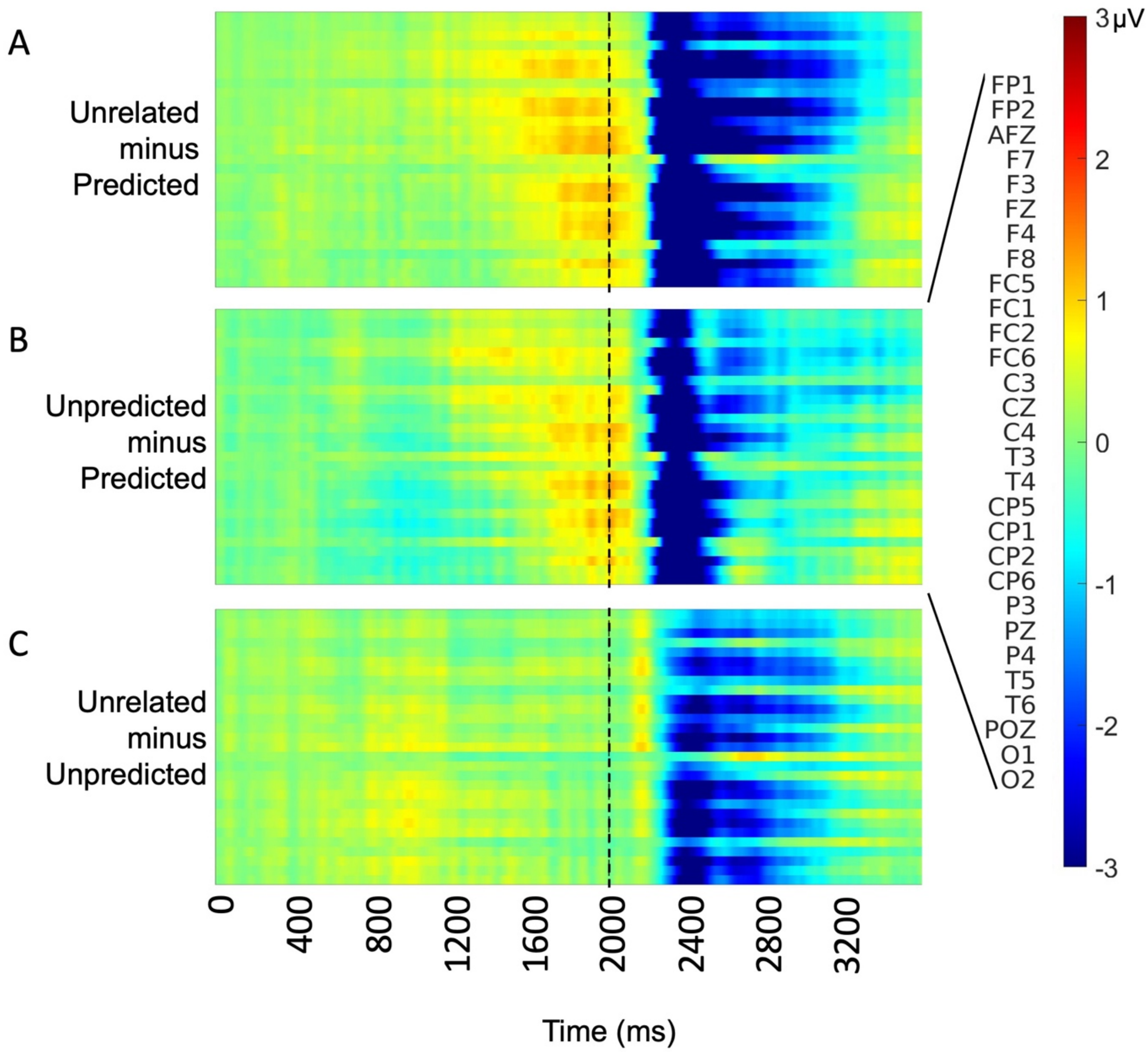

Figure S2. Shows the unmasked difference waves for main effects of target words. A) is unrelated minus predicted target words, B) is unpredicted minus predicted target words, and C) is unrelated minus unpredicted target words. Target word onset is 2000 ms.

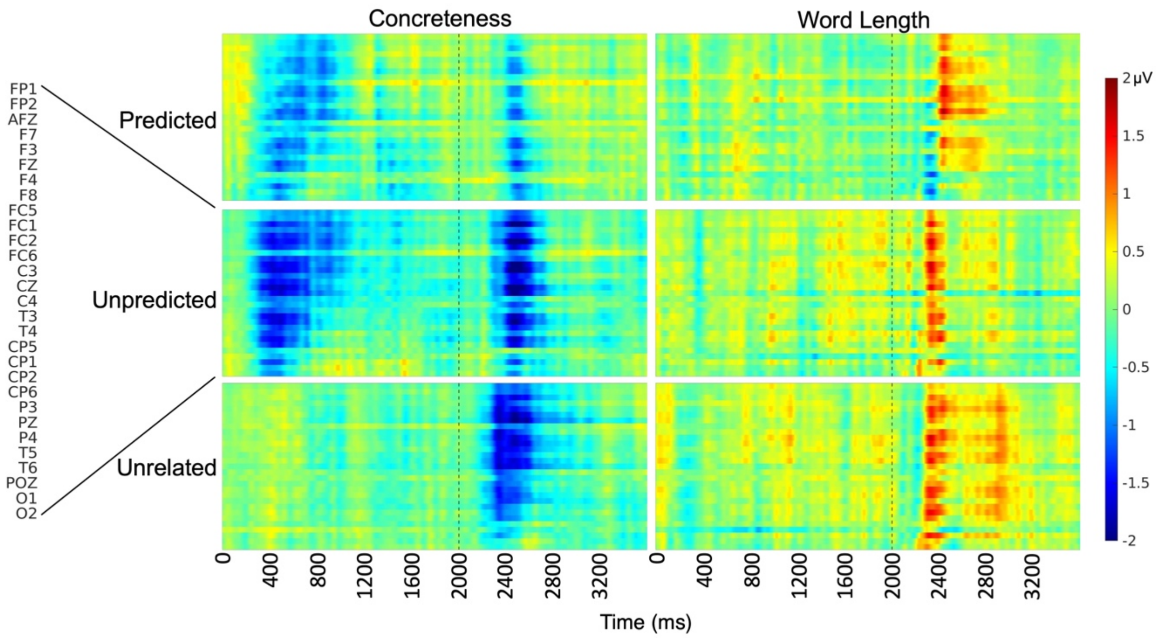

Figure S3. Shows the unmasked difference waves for feature effects of predicted target words, unpredicted – unpredicted related target word, and unrelated – unpredicted unrelated target word –. Target word onset is 2000 ms.

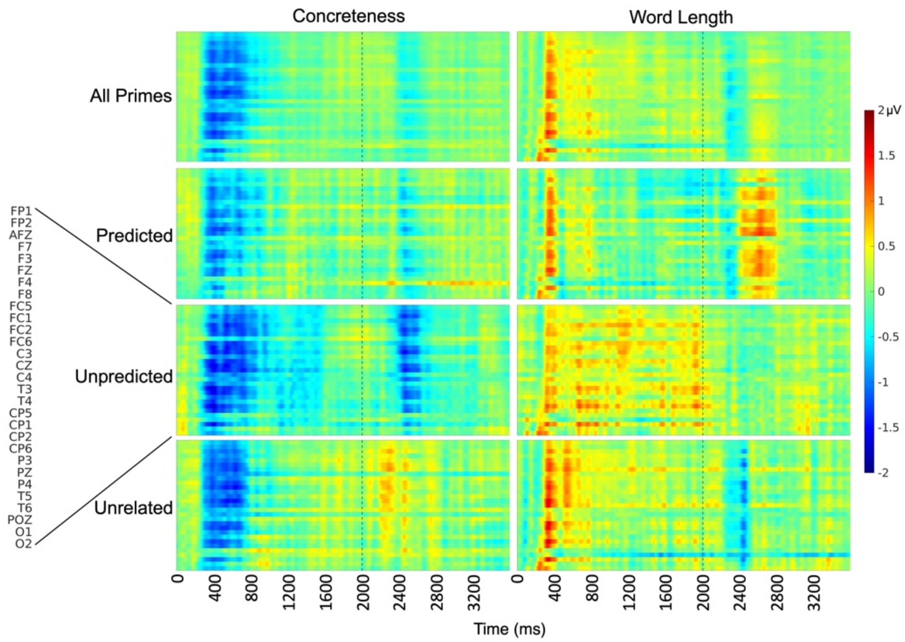

Figure S4. Shows the unmasked difference waves for feature effects of prime words. Conditions are relative to the target word: predicted – predicted target word, unpredicted – unpredicted related target word, and unrelated – unpredicted unrelated target word. Target word onset is 2000 ms.

References

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and Prediction in Human Sentence Processing. *Cognitive Science*, *33*(4), 583–609. https://doi.org/10.1111/j.1551-6709.2009.01022.x

Bae, G.-Y., & Luck, S. J. (2018). Dissociable Decoding of Spatial Attention and Working Memory from EEG Oscillations and Sustained Potentials. *The Journal of Neuroscience*, *38*(2), 409–422. https://doi.org/10.1523/JNEUROSCI.2860-17.2017

Bae, G.-Y., & Luck, S. J. (2019). *Appropriate Correction for Multiple Comparisons in Decoding of ERP Data: A Re-Analysis of Bae & Luck (2018)* [Preprint]. Neuroscience. https://doi.org/10.1101/672741

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152. https://doi.org/10.1145/130385.130401

Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2015). Graded expectations: Predictive processing and the adjustment of expectations during spoken language comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 607–624. https://doi.org/10.3758/s13415-015-0340-0

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language*, *93*, 203–216. https://doi.org/10.1016/j.jml.2016.10.002

Carrasco, C. D., Bahle, B., Simmons, A. M., & Luck, S. J. (2023). *Using Multivariate Pattern Analysis to Increase Effect Sizes for Event-Related Potential Analyses* [Preprint]. Neuroscience. https://doi.org/10.1101/2023.11.07.566051

DeLong, K. A., Chan, W., & Kutas, M. (2021). Testing limits: ERP evidence for word form preactivation during speeded sentence reading. *Psychophysiology*, *58*(2), e13720. https://doi.org/10.1111/psyp.13720

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), Article 8. https://doi.org/10.1038/nn1504

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis.

Journal of Neuroscience Methods, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009
Dikker, S., & Pylkkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation. *Brain and Language*, *127*(1), 55–64. https://doi.org/10.1016/j.bandl.2012.08.004
Eisenhauer, S., Gagl, B., & Fiebach, C. J. (2022). Predictive pre-activation of orthographic and lexical-semantic representations facilitates visual word recognition. *Psychophysiology*, *59*(3), e13970. https://doi.org/10.1111/psyp.13970
Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, *8*(7), 301–306. https://doi.org/10.1016/j.tics.2004.05.003
Fields, E. C., & Kuperberg, G. R. (2020). Having your cake and eating it too: Flexibility and power with mass univariate statistics for ERP data. *Psychophysiology*, *57*(2), e13468. https://doi.org/10.1111/psyp.13468
Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056. https://doi.org/10.1073/pnas.1216438110
Grisoni, L., Miller, T. M., & Pulvermüller, F. (2017). Neural Correlates of Semantic Prediction and Resolution in Sentence Processing. *Journal of Neuroscience*, *37*(18), 4848–4858. https://doi.org/10.1523/JNEUROSCI.2800-16.2017
Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In Spoken Word Recognition, the Future Predicts the Past. *The Journal of Neuroscience*, *38*(35), 7585–7599. https://doi.org/10.1523/JNEUROSCI.0065-18.2018
Hale, J. T. (2011). What a Rational Parser Would Do. *Cognitive Science*, *35*(3), 399–443. https://doi.org/10.1111/j.1551-6709.2010.01145.x
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. https://doi.org/10.1016/j.neuroimage.2005.11.048
Hong, X., Bo, K., Meyyappan, S., Tong, S., & Ding, M. (2020). Decoding attention control and selection in visual spatial attention. *Human Brain Mapping*, *41*(14), 3900–3921. https://doi.org/10.1002/hbm.25094
Hsu, C.-H., Lee, C.-Y., & Marantz, A. (2011). Effects of visual complexity and sublexical information in the occipitotemporal cortex in the reading of Chinese phonograms: A single-trial analysis with MEG. *Brain and Language*, *117*(1), 1–11. https://doi.org/10.1016/j.bandl.2010.10.002
Hubbard, R. J., & Federmeier, K. D. (2021). Representational Pattern Similarity of Electrical Brain Activity Reveals Rapid and Specific Prediction during

Language Comprehension. *Cerebral Cortex*, *31*(9), 4300–4313. https://doi.org/10.1093/cercor/bhab087

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*(1), 57–83. https://doi.org/10.1016/j.cognition.2012.10.013

Klimovich-Gray, A., Tyler, L. K., Randall, B., Kocagoncu, E., Devereux, B., & Marslen-Wilson, W. D. (2019). Balancing Prediction and Sensory Input in Speech Comprehension: The Spatiotemporal Dynamics of Word Recognition in Context. *Journal of Neuroscience*, *39*(3), 519–527. https://doi.org/10.1523/JNEUROSCI.3573-17.2018

Kuperberg, G. R. (2021). Tea With Milk? A Hierarchical Generative Framework of Sequential Event Comprehension. *Topics in Cognitive Science*, *13*(1), 256–298. https://doi.org/10.1111/tops.12518

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00213

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Nadra, J. G., Bengson, J. J., Morales, A. B., & Mangun, G. R. (2023). Attention without Constraint: Alpha Lateralization in Uncued Willed Attention. *eNeuro*, *10*(6). https://doi.org/10.1523/ENEURO.0258-22.2023

Ness, T., & Meltzer-Asscher, A. (2021). Rational Adaptation in Lexical Prediction: The Influence of Prediction Strength. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.622873

Noah, S., Powell, T., Khodayari, N., Olivan, D., Ding, M., & Mangun, G. R. (2020). Neural Mechanisms of Attentional Control for Objects: Decoding EEG Alpha When Anticipating Faces, Scenes,and Tools. *Journal of Neuroscience*, *40*(25), 4913–4924. https://doi.org/10.1523/JNEUROSCI.2685-19.2020

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), Article 1. https://doi.org/10.1038/4580

Rugg, M. D. (1987). Dissociation of semantic priming, word and non-word repetition effects by event-related potentials. *The Quarterly Journal of Experimental Psychology Section A*, *39*(1), 123–148. https://doi.org/10.1080/02724988743000060

Swaab, T. Y., Baynes, K., & Knight, R. T. (2002). Separable effects of priming and imageability on word processing: An ERP study. *Cognitive Brain Research*, *15*(1), 99–103. https://doi.org/10.1016/S0926-6410(02)00219-7

Szewczyk, J. M., & Schriefers, H. (2013). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and Language*, *68*(4), 297–314. https://doi.org/10.1016/j.jml.2012.12.002

Trammel, T., Khodayari, N., Luck, S. J., Traxler, M. J., & Swaab, T. Y. (2023). Decoding semantic relatedness and prediction from EEG: A classification method comparison. *NeuroImage*, *277*, 120268. https://doi.org/10.1016/j.neuroimage.2023.120268

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

Wang, L., Kuperberg, G., & Jensen, O. (2018). Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. *eLife*, *7*, e39061. https://doi.org/10.7554/eLife.39061

Wydell, T. N., Vuorinen, T., Helenius, P., & Salmelin, R. (2003). Neural Correlates of Letter-String Length and Lexicality during Reading in a Regular Orthography. *Journal of Cognitive Neuroscience*, *15*(7), 1052–1062. https://doi.org/10.1162/089892903770007434

Zhang, G., Garrett, D. R., & Luck, S. J. (2024). Optimal filters for ERP research II: Recommended settings for seven common ERP components. *Psychophysiology*, e14530. https://doi.org/10.1111/psyp.14530

Chapter 4

Decoding anticipated semantic and visual word features prior to word onset in bilinguals

1. Introduction

As discussed in previous chapters, increasingly predictive processing is considered an important mechanism to facilitate human language comprehension. To date, most research on prediction during language processing has focused on native speakers of one language. However, as more than 50% of the world speaks at least two languages (Grosjean, 2019) there is a need to better understand how bilinguals predict within both their L1 and their second language (L2). There has been a great deal of debate on whether and how bilinguals predict in L2. The aim of the present study was to examine prediction and anticipation of information during visual word recognition in L2 of Spanish-English bilinguals.

After prediction became a popular topic within mainstream psycholinguistics, research began looking for evidence of prediction within L2 processing (Kaan, 2023). Early studies found no evidence of anticipation during prediction in L2 (Grüter et al., 2012; Lew-Williams & Fernald, 2010; Martin et al., 2013, see Kaan, 2023 for overview), suggesting that bilinguals had reduced prediction ability (Grüter et al., 2014). More recent studies have shown evidence suggesting that bilinguals predict in their L2 in similar fashion to how they predict in their L1 (Chun & Kaan, 2019; Dijkgraaf et al., 2019; Hopp, 2013; Kaan, 2014), though perhaps with delays in timing relative to L1. Currently, results of studies in bilingual language users who are proficient in both languages suggests that bilinguals predict higher level semantic features (Dijkgraaf et al., 2017), and lower level lexical features including morphology (Hopp, 2013; Kaan, 2014), and orthography (Casaponsa et al., 2015; Hoversten et al., 2017). However, since these studies often have focused on the processing of the predictable word itself, not much is known about if and when semantic and lexical features are anticipated prior to the onset of this word. Thus, much is still unknown about the architecture of predictive processing in bilinguals.

As discussed in Chapters 1 and 3, predictive coding theory (Rao & Ballard, 1999) is a potentially unifying model of prediction during perceptual and cognitive processing that has been adapted to fit within language processing (Kuperberg, 2021; Kuperberg & Jaeger, 2016). In brief, predictive coding models suggest that higher levels of the cortical hierarchy make continuous predictions about upcoming input from lower levels by using prior beliefs and the current context. At each level, prediction error is calculated between the top-down prediction and the bottom-up input. This prediction error is passed back up to higher levels to update beliefs. Thus, predictive coding accounts have two major assumptions: 1) that higher level features of predicted upcoming content should be retrieved earlier during anticipation than lower-level features because higher level information must be predicted before lower level, and 2) that some representation of prediction error should be present after the bottom-up input is presented when predictions are inaccurate.

Prediction in bilingual language users may be different from monolinguals, because many findings suggest that bilinguals always activate both languages and that language selection occurs at a later stage, as suggested by the BIA+ model (Dijkstra & Heuven, 2002). BIA+ predicts that language selection (L1 or L2) occurs too late to influence semantic processing. However, many studies in support of nonselective access to both languages have studied word processing with minimal context. When the context is extended to sentences, there is evidence that of early selection of the context relevant language, such that bilinguals only activate the context relevant meaning of interlingual homographs such as pie (which means foot in Spanish) (Hoversten et al., 2015).

But even if bilinguals are able to select the context appropriate language early on, they are still facing increased entropy relative to monolinguals, as they have knowledge of two languages (Gullifer & Titone, 2020). Entropy is a concept adapted from information theory (Shannon, 1948) that is used to quantify uncertainty. Along with its mathematical equivalent surprisal, entropy has been proposed as a mechanism in language research to represent the uncertainty during language processing, particularly in Bayesian language models (e.g., Gibson et al., 2013; Hale, 2011; Jaeger & Snider, 2013; Kuperberg & Jaeger, 2016; Levy, 2008). An increase in entropy in bilinguals could impact predictive processing by making prediction error more likely due to the increased number of options available.

In the present study we will use ERPs to examine anticipation and prediction of imminent words in Spanish English bilinguals. Two ERP components that have been strongly linked to predictive effects during language processing: N400 – a negatively deflecting, centro-parietal ERP component that is maximal 300 – 500 ms post stimulus onset – and N250 – a negatively deflecting, fronto-central ERP component that is maximal 200 – 350 ms post-target onset. As was discussed in previous chapters, the N400 can be viewed as an index of prediction error within a predictive processing framework (e.g., see Federmeier, 2022 and Kuperberg, 2021 for views on this). The N250 is sensitive to processing of orthography, and is larger in amplitude when there is a mismatch between orthography of words, as has been shown in masked repetition priming when the target word differs from the masked prime, and in sentence reading studies when a predicted word differs from the actually presented word (Holcomb & Grainger, 2007; Brothers et al., 2015). The sensitivity to repetition priming of the N250 means that it could also be viewed as a measure of prediction error, but at the level of the complexity of orthographic visual features rather than at the semantic level. In monolinguals, effects of sub-lexical visual features during word recognition, such as complexity and word length, have been shown to modulate the amplitude of the N1: shorter words elicit larger N1 amplitude than longer words (Dufau, Grainger, Midgley & Holcomb, 2015). Studies of bilingual language using ERPs have typically found reduced and delayed N400 responses relative to monolinguals (Frenck-Mestre et al., 2014; Kaan, 2023; Newman et al., 2012). Casaponsa and colleagues (2015) examined modulations of the N250 and N400 in Spanish Basque bilinguals in a masked priming paradigm and found that Spanish-Basque bilinguals use orthographic features as important cues to select the appropriate language. To our knowledge, there are no studies that have found evidence of modulation of N1 amplitude as a function of word-length in Spanish-English bilinguals.

While these ERPs studies suggest that semantic, orthographic, and word-length information may be predicted, they have not examined the time course of retrieval of anticipated information before a predicted stimulus is encountered. As shown in Chapter 3 of this dissertation, full trial epoch EEG decoding accompanied by mass univariate ERP permutation cluster analysis has proven promising in monolinguals to examine this time course of anticipated information and may prove quite useful in elucidating time course of prediction prior to the onset of a predicted word as well as elucidating the effects of incorrect anticipation on prediction error in bilinguals.

The present study aimed to examine prediction during visual word processing in L2 (English) of Spanish-English bilinguals. Given results of prior studies, we predicted that bilinguals would be able to select the context appropriate language and tested the two crucial assumption of the predictive coding theory 1) bilinguals anticipate semantic features before sub-lexical features during visual word processing in a priming, and 2) failed prediction will be evident from increased prediction errors. Pre-activation of semantic features are expected to load on the N400, of orthographic features on the N250 and potentially the N1. In the present study, we used the same prediction priming paradigm as reported in Chapters 2 and 3, and the same analytical methods as reported in Chapter 3. As a brief reminder, during this priming paradigm with a prediction task, participants read prime and target words in sequence with a delay between the onsets of the two words. Participants are instructed to try to predict the target word before it appears. They then self-report their prediction accuracy at the end of the trial. Two-thirds of the word pairs are related, and the remainder is unrelated. Based on the behavioral responses about prediction accuracy and the relatedness of the target stimuli, ERP results in this study are sorted as follows: predicted – participants successfully predict the related target word, unpredicted – participants unsuccessfully predict the related target word, and unrelated – participants could not predict the target word because it was unrelated. Given prior findings about L2 early language selection, predictive processing, and the assumptions of the predictive coding theory, we predict that Spanish English bilinguals will predict target words in their L2. We anticipate that we will find electrophysiological evidence of prediction errors for target words that were unrelated and for target words that were related but not predicted. The size of the prediction error will depend on whether bilinguals anticipate semantic and sub-lexical orthographic and visual features in their L2. If these features are anticipated, then we should be able to decode them in the EEG signal prior to the onset of the target words. Additionally, successful anticipation of related target words should lead to a reduction in prediction error for the target words, which would lead to a reduction of the amplitude of the N400, the N250 and the N1. When bilinguals do not successfully predict a related target word, we predict that there will still be evidence in the EEG signal prior to the target word onset of semantic features, and reduction of the N400 amplitude relative to the unrelated condition, but not of visual features because the predicted word and the target word are likely related, but not identical. However, the pre-activation during these two conditions may be delayed relative to what has been seen in native speakers, consistent with delayed N400 responses observed post-target in bilingual speakers (Frenck-Mestre et al., 2014; Kaan, 2023; Newman et al., 2012) Finally, when bilinguals encounter an unrelated word, there should be no evidence of pre-activation of target features prior to the target word onset, but evidence of a large prediction error post-target onset (larger N400, N250 and N1 in unrelated relative to predicted and unpredicted related conditions).

2. Methods

The materials, experimental procedures, EEG recording procedures, EEG data processing procedures, and data analyses for this study are identical to those used in the study described in chapter 3. Here, we will focus on the methods were specific to this study in Spanish English bilinguals.

Figure 1. Shows the predictive priming paradigm. This is the same paradigm used in Chapter 3.

2.1 Participants

Spanish-English bilinguals (Spanish L1, *n*=30; 20 Female; age=19.2 SD=.95) participated in the study. All participants were undergraduate students at the University of California, Davis and provided written informed consent at the beginning of every recording session. Participants received research credit required for their courses as compensation for their participation. All participants were right-handed. One participant was rejected due to not having sufficient trials per condition after artifact rejection (see supplemental Table S1 for summary of artifact rejection). Statistical analyses were performed on the remaining 29 participants.

2.2 Language Proficiency Measures

Language	Spanish (L1)	English (L2)
Years of Use	18.37 (2.87)	15 (3.79)
Mode of Acquisition	Home	School
*Proficiency (LHQ)	.80 (.13)	.87 (.11)
*Dominance (LHQ)	.48 (.10)	.64 (.12)
Immersion (LHQ)	.73 (.15)	.70 (.16)
Multilingual Language Diversity (MLD) Score	1.04 (.21)	
LexTALE Score	–	.73 (26.18)

*Table 1. Results from the participants' language proficiency assessments. Means are reported along with SD in parentheses. Significance ($p < .01$) indicated by *.*

Participants completed three assessments to gauge their proficiency and use of Spanish and English, the LexTale, the Language History Questionnaire and the Multilingual Language Diversity Score. The results of these assessments are reported in Table 1. The LexTALE test is a validated measure of language proficiency (Lemhöfer and Broersma, 2012) which the participants completed for English. Participants scores on the LexTALE were in the ranges of 60-80-% or 80-100%, which corresponds with "upper intermediate" and "upper & lower advanced/proficient users" respectively, according to the Common European Framework (CEF), which is the Council of Europe's pedagogical guidelines for assessment of language proficiency (Europarat, 2020). These are the highest 2 proficiency levels across the 7 identified by the CEF.

The results on the Language History Questionnaire 3.0, a self-report measure of language usage, proficiency, and age and manner of acquisition (Li, Zhang, Yu, & Zhao, 2019), showed that the Spanish-English participants in this study were highly proficient in both languages, but somewhat more dominant in English (L2), which is typical of Spanish-English bilingual university undergraduates *(*d′ scores on the LHQ3, ($t(29) = 3.06$; $p < .01$. ($p=.0047$); ($t(29) = 5.55$; $p < .0001$*)*.

The Multilingual Language Diversity Score (MLD; Li, Zhang, Yu, & Zhao, 2019; Gullifer and Titone, 2019; DeLuca, Rothman, Bialystok, and Pliatsikas, 2019) is a continuous measure of bilingualism which considers language context and diversity. The MLD score ranges from 0-2 with 0 representing a purely monolingual individual, 1 indicating a balanced bilingual, and 2 representing a multilingual individual that is equally proficient in, at least, four different languages with equal usage. As can be seen in Table 1, the Spanish English participants in the

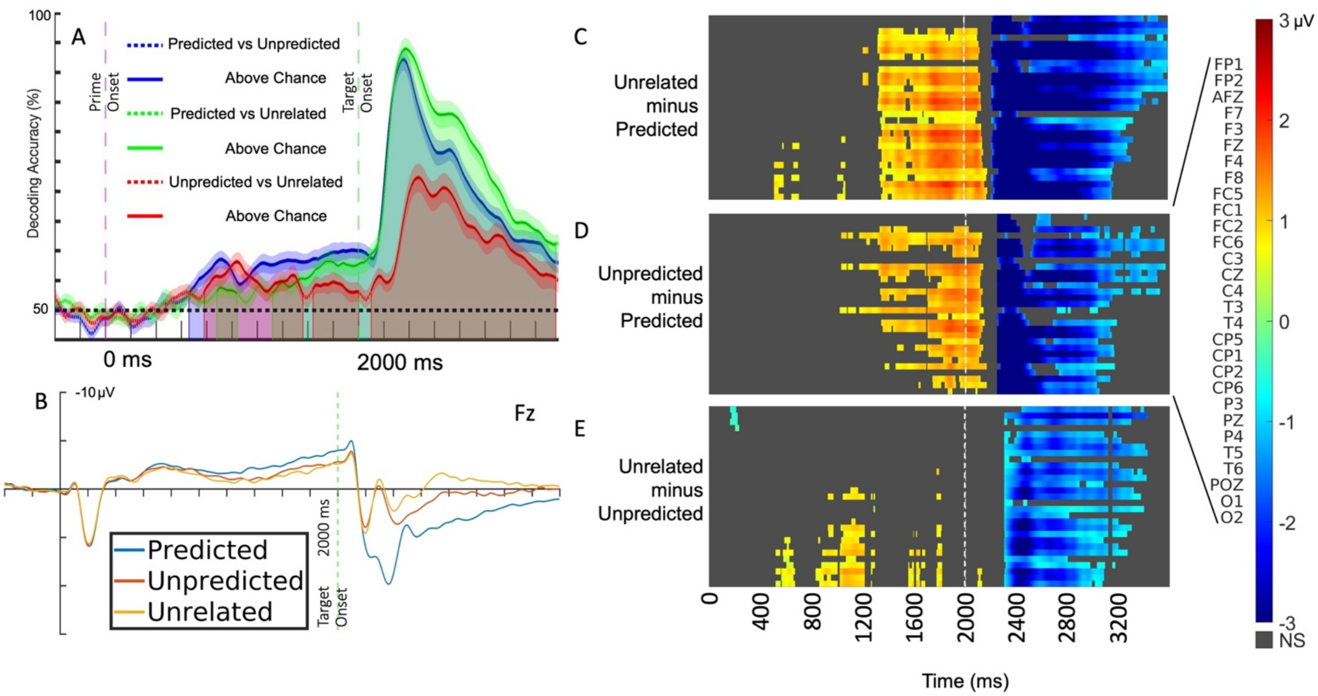

Figure 2. Shows the results from analyses of the main prediction and relatedness conditions: predicted, unpredicted, and unrelated. A) SVM decoding accuracy for three classification conditions: predicted vs. unpredicted (blue), predicted vs. unrelated (green), and unpredicted vs. unrelated (red). Prime (0 ms) and target (2000 ms) word onsets are indicated by vertical dashed lines (grey). Solid lines with shading under the curve highlight clusters of significantly above chance-level (50%; cluster t test sums exceeded 95th percentile of null distribution) decoding accuracy. B) An example ERP (Pz, see supplemental figure S1 for all electrode plots) waveform for the three main conditions: predicted (blue), unpredicted (red), and unrelated (yellow). The epoch is time-locked to prime onset and target onset is indicated by a vertical dashed line (green). ERP differences between conditions as found in the mass univariate analyses are shown for: C) predicted minus unrelated, D) predicted minus unpredicted, and E) unpredicted minus unrelated. Target onset is indicated by the vertical dashed line (white). Within these plots, non-significant time points are masked in grey (see supplemental figure S2 for unmasked plots).

present study had an MLD score of 1.04, indicating that they were equally proficient in both Spanish and English.

2.2 Materials

The materials were identical to those used in Chapter 2 and 3.

3. Results

We performed EEG decoding analyses using SVM classifiers and mass univariate ERP analyses using cluster permutation testing and report the results below. As in Chapter 3, significant time clusters for all decoding analyses are those for which their summed *t* test scores exceeded the 95th percentile of the null distribution generated by the one-tailed permutation cluster *t* test. For all mass univariate analyses, significant time clusters are those for which their summed *t* test scores exceeded the 99.5th percentile or were lower than the 0.5th percentile of the of the null distribution generated by the two-tailed permutation cluster *t* test. It is important to note that because cluster permutation testing is not designed to answer questions of latency, we worked around this limitation by performing follow-up cluster testing on all significant cluster epochs separately. This confirms that there is a significant cluster for both analyses for the epoch shown in all figures and tables.

3.1 Results of main prediction and relatedness effects

In Figure 2, results are shown for the target words when they were accurately predicted, related but unpredicted and unrelated for the decoding analyses, for univariate ERP analyses (electrode Pz) and for mass univariate ERP analyses. In all three of our decoding conditions - predicted vs. unrelated (880 – 1040 ms; 1320 – 3560 ms), predicted vs. unpredicted (660 – 3560 ms), and unpredicted vs. unrelated (780 – 1560 ms; 1640 – 2000 ms; 2100 – 3560 ms) – we observed above chance-level (50%) decoding accuracy in several epochs before and after target word onset (Figure 2A): for predicted vs. unrelated between 880 – 1040 ms and 1320 – 3560 ms, for predicted vs. unpredicted between 660 – 3560 ms, and for unpredicted vs. unrelated between 780 – 1560 ms, 1640 – 2000 ms and 2100 – 3560 ms. Figure 2B shows the ERP results from the univariate ERP analysis for electrode Fz (Figure S1 in the supplementary materials shows ERP results for all electrode sites). As can be seen in this Figure, unrelated target words elicit the greatest amplitude N400 response, related but unpredicted words elicited a reduced N400 relative to the unrelated targets, and accurately predicted words show the greatest reduction in N400 amplitude. These results were confirmed by our mass univariate analysis. The full results of the mass univariate ERP cluster permutation analyses are illustrated in figures 2C – 2E. The exact epochs and member electrodes for significantly different clusters are reported in Table 2. These results indicate large clusters of differences post-target onset for each of the conditions spanning 2000 – 3560 ms (one cluster spans both pre- and post-target onset) when the participant predicted the target word (Figure 2C), 2000 – 3596 ms when the participant failed to predict a related target word (Figure 2D), and 2096 – 3420 (first cluster begin before target onset) when the target word is unrelated to the prime (Figure 2E). We also observed significantly different clusters in all three conditions before the target word onset: 1020 – 2000 ms (one cluster begins pre-target and continues after target onset) for unpredicted minus predicted target words, 880 – 1000 (with one cluster continuing post-target) for unrelated minus predicted target words, and 164 – 1816 ms for unrelated minus unpredicted target words. The findings of the mass univariate analyses echo and support the findings of the decoding analyses. Additionally, the reversal of polarity in the pre-target to post-target results matches prior findings of the semantic prediction potential (SPP, Grisoni et al., 2020, see Chapter 1 for more details).

Table 2. Main Prediction and Relatedness Effects - Significantly different ERP cluster times

Condition	Cluster	Difference (μV)		Time (ms)		Electrodes
		Min	Max	Begin	End	
Unpredicted minus Predicted	1	0.6	1.0	1020	1072	F7 FC5 T3
	2	0.7	0.9	1084	1144	F7 FC5 T3
	3	0.6	1.4	1172	1692	F3 F7 FC1 FC2 FC5 C3 T3 CP1 CP5 P3 T5 O1 AFZ FZ CZ
	4	0.8	1.4	1704	1776	F3 F7 FC1 FC2 FC5 C3 C4 T3 CP1 CP2 CP5 P3 T5 O1 FZ CZ PZ POZ
	5	0.6	1.8	1780	2164	F3 F4 F7 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 O1 O2 AFZ FZ CZ PZ POZ
	6	-12.4	-0.6	2248	3228	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
	7	-1.6	-0.8	3248	3560	F3 F4 F7 FC1 FC2 FC5 C3 AFZ FZ
Unrelated minus Predicted	1	0.6	0.9	512	584	CP6 P4 T6 O1 O2 POZ
	2	0.5	0.8	648	708	CP6 P4 T6 O2
	3	0.5	0.8	1008	1068	CP2 P4 T6 O1 O2 POZ
	4	0.9	1.1	1208	1248	FC1 FC2 FZ
	5	0.4	2.1	1324	2180	F3 F4 F7 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
	6	-14.4	-0.7	2216	3596	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
Unrelated minus Unpredicted	1	-0.6	-0.3	164	232	FP1 FP2 F7 AFZ
	2	0.4	0.9	520	668	P3 P4 T5 T6 O1 O2 PZ POZ
	3	0.5	1.1	832	924	P3 P4 T6 O1 O2 POZ
	4	0.6	1.1	928	1000	C4 CP2 CP6 P4 T6 O1 O2 POZ
	5	0.5	1.3	1020	1212	C4 CP1 CP2 CP6 P3 P4 T5 T6 O1 O2 CZ PZ POZ
	6	0.3	0.8	1264	1292	C4 T4 CP6 P4 T6 O2 POZ
	7	0.5	1.0	1556	1596	CP6 P4 T6 O1 O2 POZ
	8	0.5	1.1	1616	1684	CP6 P4 T6 O1 O2 POZ
	9	0.5	1.0	1776	1816	FC2 C4 CP2 CP6 P4 T6 O1 O2 PZ POZ
	10	0.6	0.9	2096	2128	O1 O2 POZ
	11	-3.9	-0.5	2304	3112	FP1 FP2 F3 F4 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
	12	-2.7	-0.5	3148	3420	FP1 FP2 F3 F4 F8 FC1 FC2 FC6 C4 T4 CP1 CP2 CP6 P4 T6 AFZ FZ CZ PZ

Table 2. Summary of all significant ERP clusters found in the main prediction and semantic relationship effects analyses. Target word onset is 2000 ms.

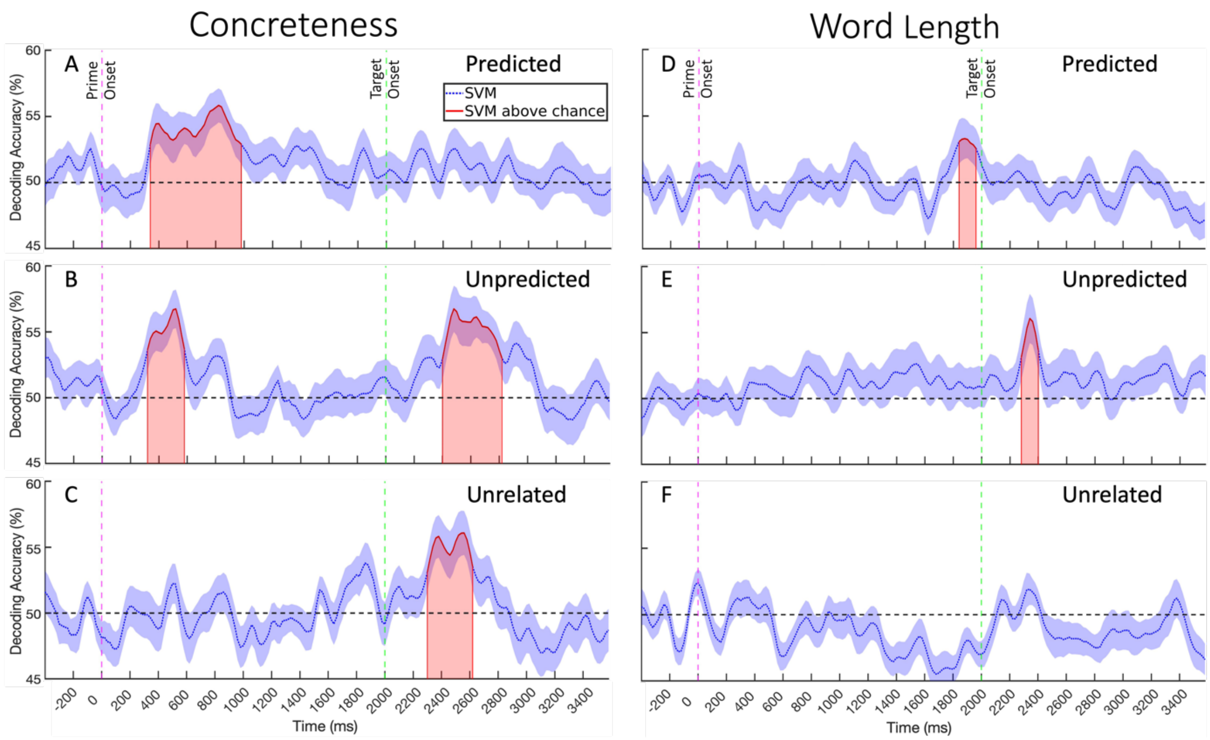

Figure 3. *Shows decoding accuracy of target word concreteness (A – C) and word length (D – F) classifications over the entire -400 – 3600 ms epoch. Prime word onsets at 0 ms (magenta) and target word onsets at 2000 ms (green). Solid red lines indicate decoding accuracy that is significantly above-chance level (50%; cluster t test sums exceeded 95th percentile of null distribution); highlighted by red shading under the curve.*

The mass univariate permutation analysis indicates that the N400 amplitude was significantly reduced for target words that were related and accurately predicted, and for target words that were related but not accurately predicted relative to the unrelated targets. The N400 is most reduced when words are predicted. It is important to note that these findings very closely match those observed in monolinguals in Chapter 3 despite the bilingual participants having a 41% prediction accuracy relative to the 51% in the monolingual participants. This may indicate that bilinguals do not predict the identity of the target words as frequently as monolinguals, but this has not been tested statistically. Strikingly, the unrelated condition appears less negative in the bilinguals than in the monolinguals from Chapter 3. This could suggest that higher entropy in bilinguals leads to reduced prediction error from unrelated target words because bilinguals must allow for possibilities in their other language as observed within Hoversten et al. (2015) where

Table 3. Features - Significantly above chance decoding cluster times

Decoded Word	Feature	Condition	Cluster	Times (ms) Begin	End
Target	Concreteness	Predicted	1	340	980
		Unpredicted	1	320	580
			2	2400	2820
		Unrelated	1	2300	2620
	Length	Predicted	1	1840	1960
		Unpredicted	1	2280	2400
Prime	Concreteness	Predicted	1	200	1500
			1	460	840
			2	940	1060
		Unpredicted	1	240	560
		Unrelated	1	380	620
	Length	Predicted	1	60	160
			1	80	740
			2	260	460
		Unpredicted	1	120	380
		Unrelated	1	260	440

Table 3. Summary of all significantly above-chance decoding cluster start and stop times for word feature decoding. There were no significantly above-chance decoding clusters for target word length in the unrelated condition. Target word onset is 2000 ms.

bilinguals after the target word was processed that it also had a Spanish meaning. Taken all together, these findings are strong evidence that the bilingual participants are predictively pre-activating features of target words when they are successfully predicted or unpredicted but related to the prime word.

3.2 Results of the feature decoding analysis

Figures 3 and 4 show that SVMs reliably decoded EEG data for concreteness and word length features of target words and prime words, respectively. Whether decoding prime or target word features, the trials were binned according to prediction accuracy of the target word and relatedness of the target to the prime which yielded three conditions: *predicted, unpredicted,* and *unrelated*. The exact epochs during which SVMs reliably decoded concreteness and word length features are reported in Table 3. The reported epochs had significantly above-chance (50%) decoding accuracy using the cluster-permutation analyses (exceeded 95th percentile of the null distribution).

As illustrated in Figure 3, significant clusters of above-chance decoding accuracy of concreteness for the target words were found in pre-target epochs when the target word was accurately predicted or unpredicted but related to the prime word. These clusters were concentrated 1680 – 1020 ms prior to target word onset (320 – 980 ms in Figure 3). During the post-target period, concreteness was significantly decodable for unpredicted but related targets and for unrelated target words during a period spanning 300 – 820 ms after target onset (2300 – 2820 ms in Figure 3). Predicted target word length was decodable 160 – 40 ms prior to target onset (1840 – 1960 ms in Figure 3). When the target word was unpredicted but related to the prime word, word length was only decodable during a period spanning 280 – 400 ms after target word onset (2280 – 2400 ms in Figure 3). Surprisingly, unrelated target word length could not be reliably decoded at all. We will entertain potential reasons for this in the discussion.

These decoding findings suggest that accurate prediction of target words pre-activated both semantic and visual features. Related words that were not accurately predicted showed evidence of semantic feature pre-activation, but no evidence for pre-activation of visual features. As such, when participants see the prime word circus, and successfully predict clown for the target word, then they pre-activated “clown” all the way down to the visual feature level. In contrast, if they instead predicted a different related word like acrobat, then some of the semantic features of the target were pre-activated, but after encountering “clown”, there would still be some semantic processing and all visual feature processing to be done. As the two words are visually different, there should be no visual features pre-activated. In the unrelated condition, the predicted words did not overlap with the target words either semantically or visually and all processing must be done after encountering the target word.

Figure 4 shows reliable decoding accuracy for the prime words’ concreteness and word length only prior to target onset. Exact epochs are reported in Table 3. Concreteness of the prime words was reliably decodable for all three conditions during a period spanning 240 – 1060 ms

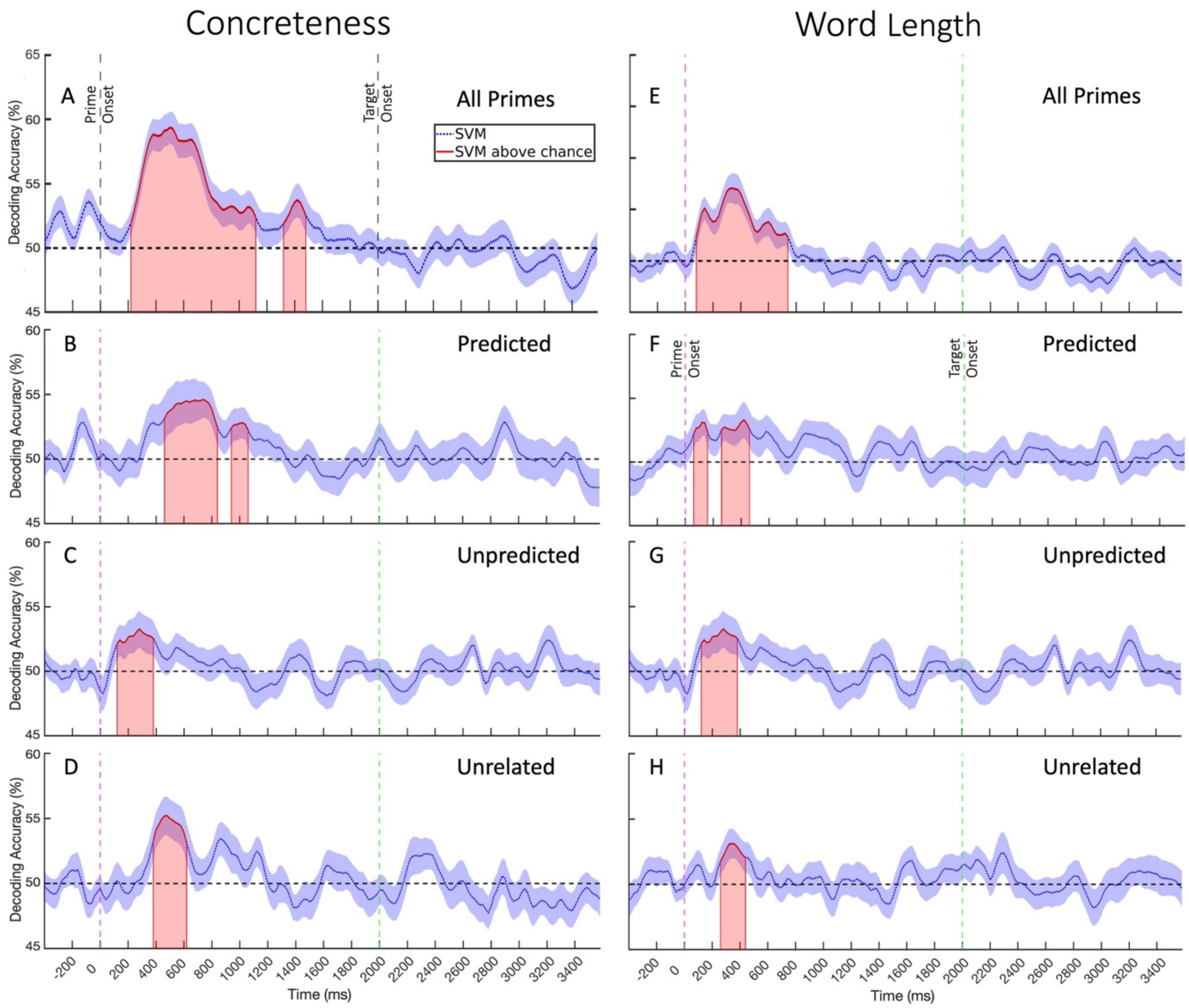

Figure 4. Shows decoding accuracy of prime word concreteness (A – C) and word length (D – F) classifications over the entire -400 – 3600 ms epoch. Prime word onsets at 0 ms (magenta) and target word onsets at 2000 ms (green). Solid red lines indicate decoding accuracy that is significantly above-chance level (50%; cluster t test sums exceeded 95th percentile of null distribution); highlighted by red shading under the curve.

after prime onset. Prime word length was reliably decoded for all three conditions during a period spanning 60 – 460 ms after prime onset.

Although we did not perform direct statistical analyses comparing the reliable decoding periods between prime and target words, the overall decoding patterns suggest that target word decoding was truly classifying the features of the target word and not the prime in the case of predicted target words. For target words, reliable decoding of concreteness spanned a period between 340 and 980 ms after prime onset (1660 – 1020 ms pre-target onset) while prime word concreteness decoding initially spanned 460 – 840 ms dropped to unreliable levels and began again from 940 – 1060 ms. Reliable target word length decoding spanned 1840 – 1960 ms after prime word onset while prime word length decoding spanned a total period of 60 – 460 ms relative to prime onset. These pattern differences suggest that target word feature decoding was not merely reflecting correlations between prime and target word features. For unpredicted but related words, the distinction is less clear as there was much overlap between target word and prime word concreteness decoding (320 – 580 ms and 240 – 560 ms, respectively) with only 20 ms of significant decoding accuracy for target words that was not accounted for by a similar cluster in prime words (560 – 580 ms). The makes it uncertain whether the classifier was truly decoding the target word features or merely reflective of correlation between prime word and target word concreteness when related words were unpredicted.

3.3 Results of feature ERP analysis

Figures 5 and 6 illustrate the ERP effects of concreteness and word length features for the target and prime words, respectively. Exact epochs and member electrodes are reported in Table 4. The ERP analyses were computed using ERP permutation cluster analyses (Maris &

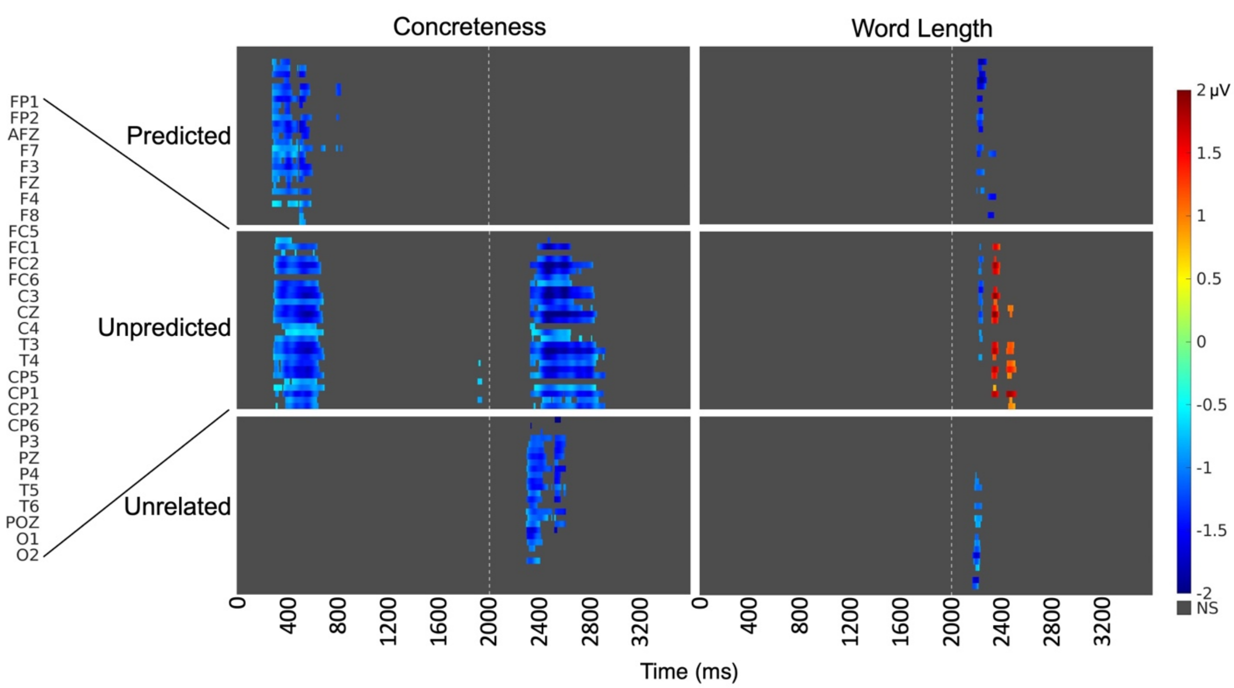

Figure 5. Shows difference wave ERP results (μV) for target concreteness (abstract minus concrete) and word length (long minus short) for each condition. Darkened areas indicate no significant clusters of differences detected within that spatio-temporal range (see supplemental figure S# for unmasked plots). The dashed white line indicates target word onset (2000 ms).

Oostenveld, 2007). As illustrated in Figure 5, both predicted target words and related target words were not predicted showed significant differences of concreteness prior to target word onset. These concreteness effects are primarily concentrated 1720 – 1164 ms prior to target onset (280 – 836 ms in Figure 5). Effects of concreteness were only found post-target concentrated 292 – 604 ms after target onset (2292 – 2604 ms in Figure 5). These findings align well with the decoding findings for each condition, providing further evidence that concreteness is pre-activated when a target word is predicted or unpredicted but related to the prime word. For unpredicted related target words and unrelated target words, significant differences were found post-target. However, as expected, there was no evidence of pre-activation for unrelated target words prior to target onset. Again, these findings are in alignment with the decoding results for these conditions and provide further evidence that while semantic features of related but unpredicted words are partially pre-activated, there is still processing to do after encountering the actual word and that unrelated words are not semantically pre-activated at all. For the target words, we observe no significant differences prior to the target word onset for word-length. All differences are concentrated within a period spanning 196 – 508 ms after target onset (2196 – 2508 in Figure 5) which coincides with the epoch during which the N250 occurs. It is possible that the effects which the classifier identified during the 1840 – 1960 ms significant decoding cluster were too small to be identified by our ERP analysis. Interestingly, there are significant differences observed post-target for the unrelated target words. When combined with the null results in the decoding analysis for this condition, this suggests that there may have been a decrease in signal-to-noise ratio for this condition that prevented the classifier from finding differences between long and short words that were still observable when averaging the data for ERPs.

Figure 6 depicts significant pre-target differences between concrete and abstract prime words for all three conditions that are concentrated between 288 and 984 ms after prime word onset. For unpredicted but related words, we observe post-target word onset concreteness effects that are concentrated 408 – 652 ms after target onset (2408 – 2652 ms in Figure 6).

Significant differences for prime word length were observed almost exclusively prior to target onset with clusters concentrated between 136 and 664 ms after prime word onset. The lone exception was for primes with unrelated target words where significant differences were

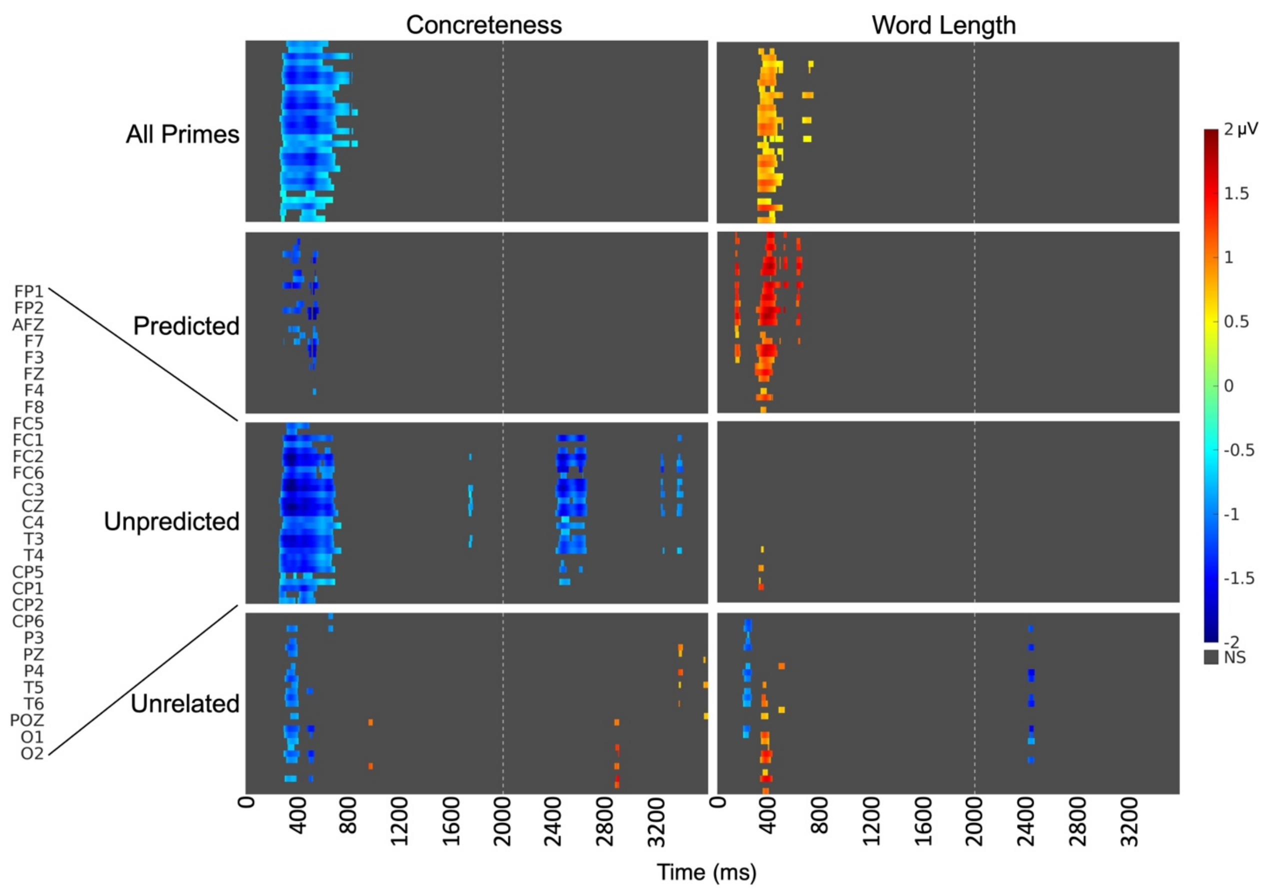

Figure 6. Shows difference wave ERP results (µV) from prime concreteness (abstract minus concrete) and word length (long minus short) for each condition. Red areas indicate significant clusters of differences in that spatio-temporal range (see supplemental figure S# for unmasked plots). The dashed white line indicates target word onset (2000 ms).

observed in a period spanning 416 – 464 ms after target word onset (2416 – 2464 ms in Figure 6). Table 5 summarizes a post-hoc cluster analysis on the paired prime and target word ERP amplitudes of difference waves for features (concreteness and word length). Critically, we observe a cluster from 240 – 420 ms which shows that the target word concreteness effect is larger during this time frame than the prime word concreteness effect. This suggests differential processing of prime words and predicted target word concreteness during this epoch.

Table 4. Features - Significant differences from ERP cluster analyses

Word	Feature	Condition	Cluster	Feature Difference (μV) Min	Feature Difference (μV) Max	Times (ms) Begin	Times (ms) End	Electrodes
Target	Concreteness	Predicted	1	-1.6	-0.5	280	484	F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T6 AFZ FZ CZ PZ
			2	-1.7	-0.6	496	836	F3 F4 F7 F8 FC1 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T6 O1 O2 CZ POZ
		Unpredicted	1	-1.9	-0.5	288	692	FP2 F3 F4 F7 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
			2	-0.9	-0.6	1908	1940	P3 T5 O1
			3	-2.3	-0.5	2324	2920	FP2 F3 F4 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T6 O1 O2 AFZ FZ CZ PZ POZ
		Unrelated	1	-1.8	-0.7	2292	2496	FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 AFZ FZ CZ
			2	-2.1	-0.7	2516	2604	FP1 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 T3 T4 CP1 CP5 FZ CZ
	Length	Predicted	1	-2.1	-0.9	2196	2276	F3 F4 F7 FC2 FC5 FC6 C3 T4 CP5 CP6 P4 AFZ FZ CZ
			2	-1.6	-1.1	2284	2348	CP5 T5 O1
			1	-1.5	-0.8	2212	2248	F3 F4 FC1 FC2 FC5 FC6 C3 C4 T4 CP5 CP6 AFZ FZ CZ
			2	0.7	2.1	2316	2380	F3 F4 FC1 FC2 FC6 C3 C4 CP1 CP2 CP6 P4 T6 AFZ FZ CZ PZ POZ
			3	0.8	1.9	2432	2508	C3 CP1 CP2 P3 P4 O1 O2 CZ PZ POZ
			1	-1.8	-0.6	2168	2240	FC1 FC2 FC6 C4 T4 CP2 CP5 CP6 P3 P4 T5 O1 PZ POZ
Prime	Concreteness	Predicted	1	-1.7	-0.3	264	804	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
			1	-1.5	-0.9	288	352	F3 F7 FC5 C3 T3 CP5
			2	-1.5	-0.8	364	460	FP2 F3 F4 F7 F8 FC5 FC6 C3 T3 T4 CP5 AFZ
			3	-1.8	-0.9	480	508	F3 FC1 FC5 C3 CP1 CP2 CP5 P3 CZ
			4	-2.0	-0.8	524	568	F3 F4 F7 F8 FC1 FC5 FC6 C3 T3 T4 CP1 CP2 CP5 CP6 P3 T6 CZ
			2	-0.9	-0.5	824	868	F4 FC2 FC6 C4 T4 AFZ FZ
		Unpredicted	1	-2.2	-0.5	260	740	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
			2	-0.9	-0.5	1732	1764	FC2 FC6 C3 C4 CP1 CP2 FZ CZ
			3	-1.9	-0.5	2408	2652	F3 F4 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P4 T6 AFZ FZ CZ PZ
			4	-1.4	-0.8	3228	3256	F4 F8 FC1 FC2 FC6 C4 CP6 FZ CZ

		5	-1.2	-0.7	3352	3400	F4 F8 FC1 FC2 FC6 C4 CP6 AFZ FZ CZ
	Unrelated	1	-1.3	-0.7	296	416	F3 F4 FC1 FC2 FC5 FC6 C3 C4 T4 CP1 CP2 CP6 P3 P4 AFZ FZ CZ PZ POZ
		2	-1.4	-1.0	476	532	C3 CP1 CP2 P4 PZ POZ
		3	-1.0	-0.8	644	676	FP1 AFZ
		4	1.0	1.2	956	984	CP5 T5
		5	0.9	1.5	2868	2900	CP5 P3 T5 O1 PZ POZ
		6	0.7	1.2	3364	3396	F4 FC1 FC6 C4 FZ
		7	0.6	0.9	3560	3596	F8 FC6 T4
Length	Predicted	1	0.6	1.6	136	176	FP1 FP2 F4 F7 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP6 FZ CZ
		2	0.7	1.9	292	488	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T4 CP1 CP2 CP5 CP6 P3 P4 T6 O2 AFZ FZ CZ PZ POZ
		1	0.3	1.3	316	512	FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T6 O1 O2 AFZ FZ CZ PZ POZ
		3	1.2	1.7	516	548	FP1 F3 FC5 C3 AFZ FZ
		4	1.1	1.6	612	664	FP2 F3 F4 FC1 FC2 FC5 FC6 C4 CP5 AFZ FZ CZ
		2	0.4	0.9	664	748	F3 F7 FC5 C3 T3
	Unpredicted	1	0.6	1.3	320	356	CP6 P4 T6 POZ
	Unrelated	1	-1.2	-0.6	196	264	FP2 F3 F7 FC1 FC2 FC5 C3 C4 CP1 CP2 AFZ FZ CZ
		2	0.6	1.6	332	424	FC6 C4 T4 CP2 CP6 P3 P4 T6 O2 CZ PZ POZ
		3	0.6	1.0	476	520	FC5 T3
		4	-1.6	-0.8	2416	2464	FC1 FC2 C4 CP1 CP2 CP6 P4 AFZ FZ CZ

Table 4. Summary of all significantly different ERP cluster start times, stop times, and member electrodes between target word and prime word features. For concreteness, a positive feature difference (μV) indicates that abstract words had greater amplitude than concrete words. For word length, a positive feature difference indicates long words had greater amplitude than short words. Target word onset is 2000 ms.

4. Discussion

The present study aimed to investigate the time course of top-down pre-activations during predictive processing in the L2 of Spanish English bilinguals' visual word recognition using two approaches: machine-learning EEG decoding with mass univariate ERP cluster analyses. Within this paradigm, we assumed the selection of contextually relevant language, and if this is the case, then representations of semantic features should be identifiable within the EEG signal earlier than representations of visual features and we should see greater prediction error after unrelated target words than we observe within related but unpredicted or predicted target words. If bilinguals select the context appropriate language (no multiple activation of both languages) then they should be able to anticipate relevant features just like monolinguals. Correct prediction would lead to reduction of prediction error (smaller N400), incorrect prediction would lead to a larger prediction error, but relative to monolinguals the N400 would be reduced, because bilinguals often must switch between languages and, therefore, keep open the possibility that the unexpected input is the other language that needs activating. As was the case in Chapter 3, our prediction task enabled us to distinguish between trials during which the participants successfully predicted the exact target word from trials from an unsuccessful prediction when the target word was related and from trials with unrelated target words that could not be predicted. This allowed us to uncover whether semantic and visual features were anticipated and whether successfully predicted words showed decreases in prediction error. Both the results from the decoding analyses and of the mass univariate ERP analyses showed that semantic features were anticipated prior to sub-lexical features when target words were successfully predicted. Evidence of anticipation of semantic features was also observed for related targets that were not predicted, likely because participants anticipated a word that was related to the presented target word, but not evidence of anticipation of sub-lexical visual features was found in this case. When target words were unrelated, there was no evidence of anticipation of either of these features prior to target word onset. The pre-target decoding and ERP results of the main effect analyses are also consistent with prior findings about semantic prediction. The results from this experiment provide compelling evidence that bilinguals engage in predictive processing in alignment with predictive coding accounts of language comprehension (e.g., Kuperberg & Jaeger, 2016).

Except for one surprising result while decoding word length for unrelated target words, the findings are exactly as would be predicted for these tasks according to top-down predictive pre-activation. The null post-target decoding results for unpredicted unrelated target word length are particularly perplexing as this was both unexpected and a robust effect seen in the monolinguals in Chapter 3. However, we did have evidence of processing prediction error in the ERP analysis for unrelated target words that is in the typical time frame of the N250. This ERP analysis cluster occurs at approximately the same time as a slight increase in decoding accuracy that was not sufficient to be considered significant. Therefore, it is possible that due to a decrease in signal-to-noise ratio the classifier could not reliably pick up on word length differences during this time. One possible source for noise may be the reactivation of the prime word as we observed prime word activity within the time frame during which we would have expected to decode target word length.

In this priming paradigm with a prediction task, the effects of prediction during visual word recognition in L2 do not appear to be delayed relative to the effects found for monolinguals in Chapter 3. We did not directly statistically compare the epochs from bilinguals with native speakers. However, when examining the onsets of main prediction effects and feature effects, they do not seem to differ much from those observed in Chapter 3. Indeed, the one major difference in timing that was observed in bilinguals was earlier activation of word length in predicted words. This may be due to bilinguals' reliance on orthographic features to identify language membership (Casaponsa et al., 2015). It is important to remember that these were highly balanced and proficient bilinguals. Therefore, it is possible that delayed prediction timings are a function of L2 proficiency which we did not investigate here.

Overall, our combination of prediction paradigm, EEG decoding, and mass univariate ERP analyses uncovered anticipation of both concreteness information and word length information within bilinguals reading in L2. Additionally, that successfully anticipated information prevents the need for further processing of that information or calculation of prediction error.

Table 5. Features - Significant differences between target and prime words from ERP cluster analyses

Feature	Condition	Cluster	Target Difference (μV)		Prime Difference (μV)		Target - Prime (μV)		Times (ms)		Electrodes
			Min	Max	Min	Max	Min	Max	Begin	End	
Concreteness	Predicted	1	-0.4	0.4	-0.4	0.3	0.4	0.8	**240**	**420**	F8 FC6 C4 T4 CP2 CP6 P3 P4 T6 O1 O2 POZ
		2	0.8	1.0	0.6	0.8	0.4	0.9	1112	1128	F8 FC6
		3	0.3	1.0	0.1	0.8	0.5	1.1	1132	1152	F8 FC6 T4 CP6
		4	0.3	1.1	0.1	0.7	0.5	1.0	1276	1300	FC6 C4 T4 CP2 CP6 T6
		5	0.3	0.5	0.0	0.2	0.6	0.6	1376	1388	CP6 T6
		6	-0.5	0.2	-0.6	0.3	0.5	0.9	1460	1476	CP6 T5 T6 O1 O2
		7	-0.1	0.7	-0.9	0.0	0.6	1.1	1580	1612	CP6 T6 O1 O2 POZ
		8	-0.2	0.9	-0.7	0.4	0.4	1.1	1616	1692	F4 F8 FC5 FC6 C4 T3 T4 CP1 CP2 CP6 P3 P4 T5 T6 O1 O2 CZ PZ POZ
		9	0.1	1.0	-0.4	0.5	0.5	1.3	1712	1820	F4 F8 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 CZ PZ POZ
		10	0.2	1.2	-0.3	0.5	0.6	1.5	1820	1928	FP1 F3 F4 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
		11	0.1	0.9	-0.3	0.6	0.6	1.1	1972	2052	FC2 FC5 FC6 C4 T3 T4 CP2 CP6 P4 T6 CZ
		12	0.0	0.9	-0.6	0.2	0.5	0.8	2080	2104	FC6 T4 CP2 CP6 P4 T6 O2
		13	-0.1	1.0	-0.7	0.3	0.6	0.9	2112	2128	F8 FC5 FC6 T3 T4 CP2 CP5 CP6 P3 P4 T6 O2 POZ
		14	0.0	1.2	-0.7	0.4	0.4	1.3	2156	2288	FP2 F4 F8 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T6 CZ POZ
		15	0.0	1.2	-0.8	0.3	0.5	1.2	2356	2536	F8 FC6 C4 T4 CP1 CP2 CP6 P3 P4 T5 T6 O1 O2 CZ PZ POZ
		16	-0.2	0.9	-0.8	0.2	0.5	0.9	2600	2664	F8 FC6 T4 CP1 CP2 CP6 P3 P4 T6 O1 O2 CZ POZ
		17	-0.4	0.7	-0.6	0.3	0.6	0.7	2676	2700	CP2 P4 O1 POZ
		18	0.2	0.7	-0.6	0.1	0.5	1.5	2764	2836	FP2 F8 FC6 T4 CP6

Unpredicted	1	0.0	0.2	-0.2	0.0	-0.7	-0.5	124	164	F8 FC6 T4
	2	-0.4	0.2	-0.5	0.4	-0.9	-0.3	**240**	**476**	F4 F8 FC6 C3 C4 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 PZ POZ
	3	-0.1	0.3	0.0	0.4	0.4	0.8	**540**	**612**	CP2 P3 P4 T5 O1 O2 PZ POZ
	4	0.6	0.8	0.7	0.9	0.4	0.6	756	772	O1 O2
	5	1.2	1.6	0.6	1.0	0.4	0.5	944	956	O1 O2 POZ
	6	0.4	1.5	0.8	1.3	-0.8	-0.5	976	1024	F8 FC6 C4 T4
	7	-0.4	0.7	-0.2	0.4	0.3	1.2	**2308**	**2460**	F3 FC1 FC2 C3 CP1 CP2 CP6 P3 P4 T6 O1 O2 FZ CZ PZ POZ
	8	-0.1	0.5	0.0	0.4	0.4	1.1	**2488**	**2604**	F4 FC2 C3 CP1 CP2 CP6 P3 P4 O1 O2 FZ CZ PZ POZ
	9	0.3	2.2	-0.1	1.5	0.4	1.2	**2636**	**2996**	F3 F4 FC1 FC2 C3 C4 CP1 CP2 CP6 P3 P4 T6 O1 O2 AFZ FZ CZ PZ POZ
	10	0.9	1.8	0.2	1.1	0.7	0.8	3016	3076	CP2 P4 O2
	11	0.2	0.8	0.5	1.1	-0.7	-0.5	3420	3440	FP2 F8 T4
Unrelated	1	-0.4	0.3	-0.3	0.2	0.6	1.2	8	200	C3 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 CZ PZ POZ
	2	-0.2	0.1	-0.2	0.2	0.7	1.0	208	236	FP2 F4 FC1 FC2 CP1 CP2 CZ
	3	-0.3	0.4	-0.5	0.2	-1.6	-0.7	308	428	FP1 F3 F4 F7 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 AFZ FZ CZ PZ POZ
	4	0.0	0.3	-0.6	-0.3	-1.5	-0.9	492	508	C3 CP1 CP5 PZ
	5	-0.5	-0.1	0.1	0.2	-1.4	-1.1	544	564	F4 FC2 AFZ FZ
	6	-0.7	0.5	-0.6	0.7	-1.5	-0.8	568	700	FP1 FP2 F3 F4 F8 FC2 FC6 C3 C4 T4 CP1 CP2 CP6 P4 T6 AFZ FZ CZ PZ
	7	-0.4	0.1	0.2	1.2	-1.3	-1.0	712	736	FP1 FP2 F8 T4 AFZ FZ
	8	-0.5	0.2	0.5	1.1	-1.7	-0.8	808	936	FP2 F4 FC2 FC6 C4 T4 CP6 AFZ FZ
	9	-0.1	0.3	-0.3	0.7	0.8	1.2	956	1032	T3 CP5 P3 T5
	10	-0.5	-0.1	-0.9	-0.6	0.9	1.3	**2224**	**2248**	FC5 T3 CP5 T5
	11	-0.9	1.7	-1.2	0.4	0.8	2.5	**2256**	**2788**	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 T5 O1 O2 AFZ FZ CZ PZ POZ
	12	-0.5	1.2	-1.3	-0.5	0.8	1.8	3572	3584	FP1 F3 F4 FC1 FC2 FC5 FC6 T4

		13	-0.3	0.0	-1.0	-0.5	0.7	1.3	3592	3592	CP5 T5
Length	Predicted	1	-0.5	0.5	-0.2	0.3	0.6	2.0	**140**	**252**	FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T4 CP1 CP2 CP5 CP6 P4 AFZ FZ CZ PZ
		2	-0.4	0.4	-0.7	0.2	0.7	2.4	**276**	**368**	FP1 FP2 F3 F4 F7 FC2 FC5 FC6 C3 C4 T4 CP1 CP2 CP5 CP6 P3 P4 T6 O1 O2 AFZ FZ CZ PZ POZ
		3	-0.3	0.4	-0.5	0.6	0.8	2.4	**372**	**468**	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T4 CP1 CP2 CP5 CP6 P3 P4 T6 O2 AFZ FZ CZ POZ
		4	0.2	0.4	0.4	0.7	1.1	2.0	**472**	**488**	F3 F7 FC5 C3 CP5 P3
		5	-0.2	0.7	-1.5	0.5	1.0	2.4	**496**	**540**	FP1 FP2 F3 F4 F7 FC1 FC5 FC6 C3 CP1 CP5 P3 AFZ FZ CZ
		6	-0.3	1.5	-1.8	0.1	0.9	2.4	**612**	**808**	FP2 F3 F4 FC1 FC5 FC6 C3 CP5 CP6 P4 AFZ FZ CZ
		7	-0.1	1.0	-1.7	-0.7	0.9	2.4	916	1020	FP2 F3 F4 FC2 FC5 FC6 C3 CP1 CP2 CP5 CP6 P4 AFZ FZ CZ
		8	-0.1	0.5	-0.5	-0.1	0.9	1.6	1532	1564	FC5 C3 CP5
		9	0.3	0.5	-0.2	0.2	1.1	1.9	**2204**	**2244**	F3 F4 FC6 AFZ FZ
		10	0.8	1.0	0.8	1.0	1.2	1.3	2680	2692	C3 CP1
	Unpredicted	1	-0.2	0.4	-0.2	0.6	-0.9	-0.4	0	28	FP1 FP2 F4 FC2 FC6 C4 T4 CP2 CP6 T6 AFZ FZ CZ
		2	0.0	0.3	-0.2	0.0	-1.1	-0.7	200	212	FP1 FC1 AFZ FZ
		3	-0.3	0.2	-0.2	0.3	-1.4	-0.6	216	272	FP1 FP2 F3 F4 F7 F8 FC1 FC2 FC6 C4 AFZ FZ
		4	-0.3	0.3	-0.1	0.4	0.7	1.2	**320**	**408**	O2 PZ POZ
		5	-0.2	-0.1	-1.2	-0.8	-1.1	-0.9	756	768	PZ POZ
		6	-0.3	0.0	-0.6	-0.4	0.7	0.9	812	836	FC5 T3
		7	0.1	0.2	-0.4	-0.2	0.6	1.1	1432	1500	FC5 T3
		8	-0.1	0.2	-0.5	-0.2	0.6	1.0	1640	1660	FC5 T3 CP5 T5
		9	0.4	0.7	-0.2	0.0	0.6	1.4	1892	1904	FC5 C3 T3 CP1 CP5 P3 T5
		10	-0.3	0.6	-0.5	0.3	0.6	1.6	1912	1948	FC1 FC5 FC6 C3 C4 T3 CP1 CP2 CP5 CP6 P3 P4 T5 CZ POZ
		11	0.0	0.7	-0.6	0.2	0.7	1.4	2012	2084	FC1 FC5 C3 C4 T3 CP1 CP2 CP5 CP6 P3 T5 CZ PZ POZ

	12	-0.1	0.4	-0.4	0.1	0.7	1.2	2112	2128	T3 CP1 CP5 P3 T5 POZ
	13	0.1	0.6	-0.9	0.1	-1.5	-0.7	2288	2308	CP6 P3 P4 PZ POZ
	14	-0.3	0.8	-1.1	0.5	-1.9	-0.6	2312	2556	F4 FC2 FC6 C3 C4 T4 CP1 CP2 CP5 CP6 P3 P4 T5 T6 O1 O2 AFZ FZ CZ PZ POZ
	15	-0.5	0.1	-0.1	0.6	-1.4	-0.7	3404	3508	P3 T6 O1 O2 PZ POZ
	16	-0.5	-0.1	-0.3	0.5	-1.6	-0.8	3572	3596	FC2 C3 C4 CP1 CP2 P3 P4 T6 O2 CZ PZ
Unrelated	1	-0.4	0.0	-0.1	0.2	0.7	1.7	**320**	**332**	F4 FC2 FC6 C4 T4 CP2 CP6 P4 T6 CZ PZ POZ
	2	-0.4	0.4	-0.2	0.6	0.7	2.0	**340**	**440**	F3 F4 F7 F8 FC1 FC2 FC5 FC6 C3 C4 T3 T4 CP2 CP5 CP6 P3 P4 T6 AFZ FZ CZ PZ POZ
	3	-0.8	0.1	-0.4	0.1	-2.2	-0.8	908	1048	FP1 FP2 F4 FC1 FC2 FC6 C4 AFZ FZ
	4	-0.7	-0.2	-0.3	0.2	-1.5	-0.8	1700	1712	F4 FC2 C4 CP2 CP6 P4 T6 FZ
	5	-0.9	0.1	-0.1	0.8	-2.1	-0.8	1724	1768	FP1 FP2 F3 F4 FC1 FC2 C3 C4 CP1 CP2 CP6 P4 FZ CZ
	6	-0.4	0.6	-0.4	0.5	0.6	2.4	**2072**	**2236**	FC6 C3 T4 CP5 CP6 P3 P4 T5 O1 O2 PZ POZ
	7	-1.1	0.7	-0.4	0.9	-2.4	-0.7	2352	2520	FP1 F3 F4 F7 FC1 FC5 FC6 C3 C4 T3 T4 CP1 CP2 CP5 CP6 P3 P4 O2 AFZ FZ CZ POZ

Table 5. Summary of all significantly different ERP cluster start times, stop times, and member electrodes for comparison between target and prime word feature effects. Clusters which overlap with significant clusters from the original target or prime word analyses are in ***bold.*** *The max and min difference values (Target – Prime) indicates direction of the difference; positive values indicate that target word had a larger difference between the features and negative values indicates that the prime word had larger differences. Target word onset is 2000 ms.*

.

5. Conclusion

The present experiment highlights the time course of top-down activation by predictive processing during L2 word recognition within Spanish-English bilinguals. We utilized a combination of EEG decoding and mass univariate ERP analysis to show that semantic features – such as concreteness – and visual features – such as word length – are activated before a reader encounters a predicted word. Therefore, these findings provide a critical extension of our knowledge of the predictive processing time course within bilinguals.

Appendix

This appendix contains supplemental tables and figures that could not be included within the main paper. Figure S1 shows ERP waveforms for all electrode sites. Figures S2 – S4 depict the unmasked ERP difference waves for all the conditions of the study: Figure S2 shows the main effects waveforms, Figure S3 shows target word waveforms, and Figure S4 shows the prime word waveforms. Tables S1 and S2 show the artifact rejection summaries for word length and concreteness, respectively.

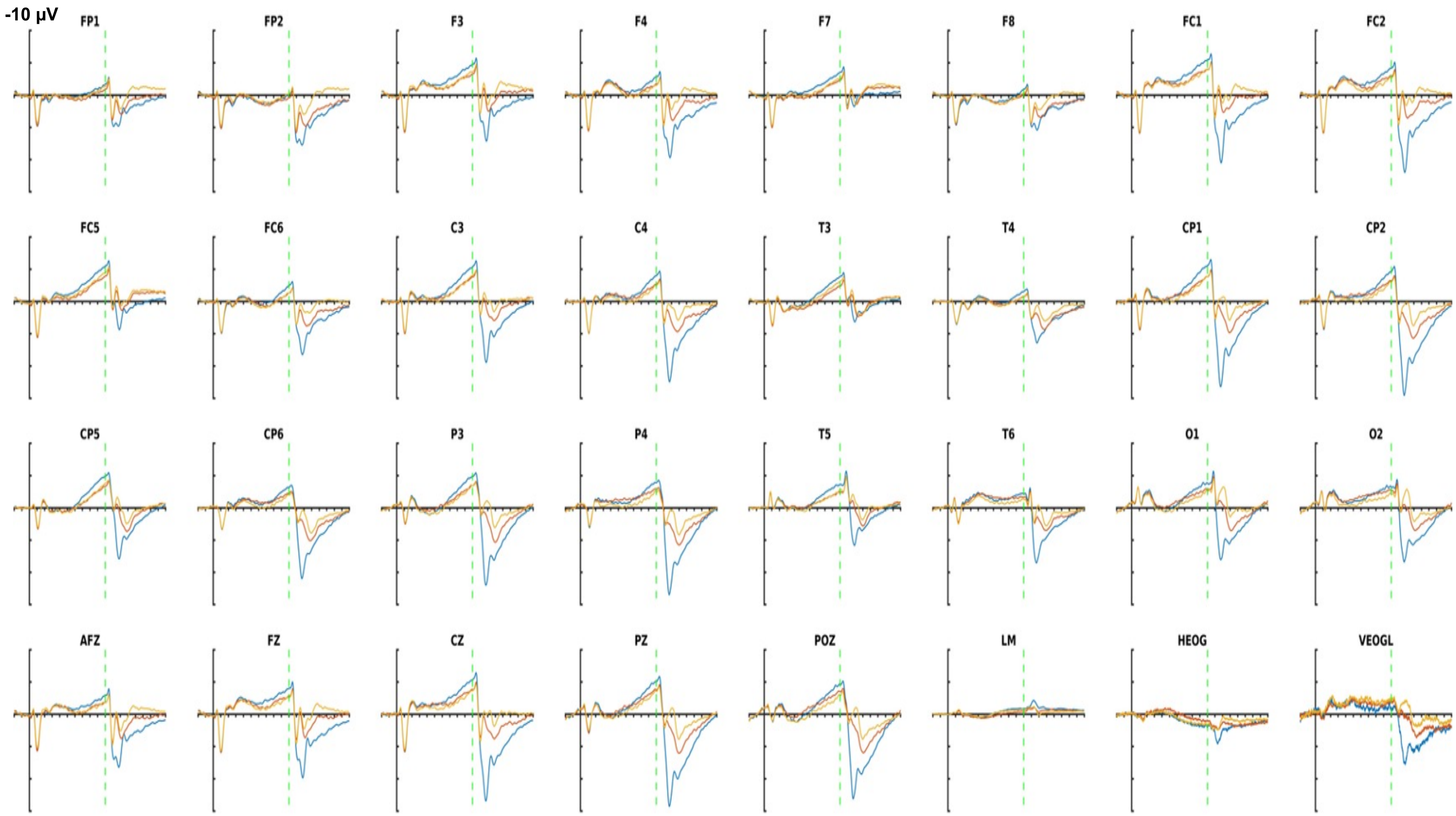

Figure S1. Shows all electrode ERP plots for predicted (blue) – predicted target word, unpredicted (red) – unpredicted related target word, and unrelated (yellow) – unpredicted unrelated target word – conditions. Target word onset is 2000 ms indicated by the dashed green line.

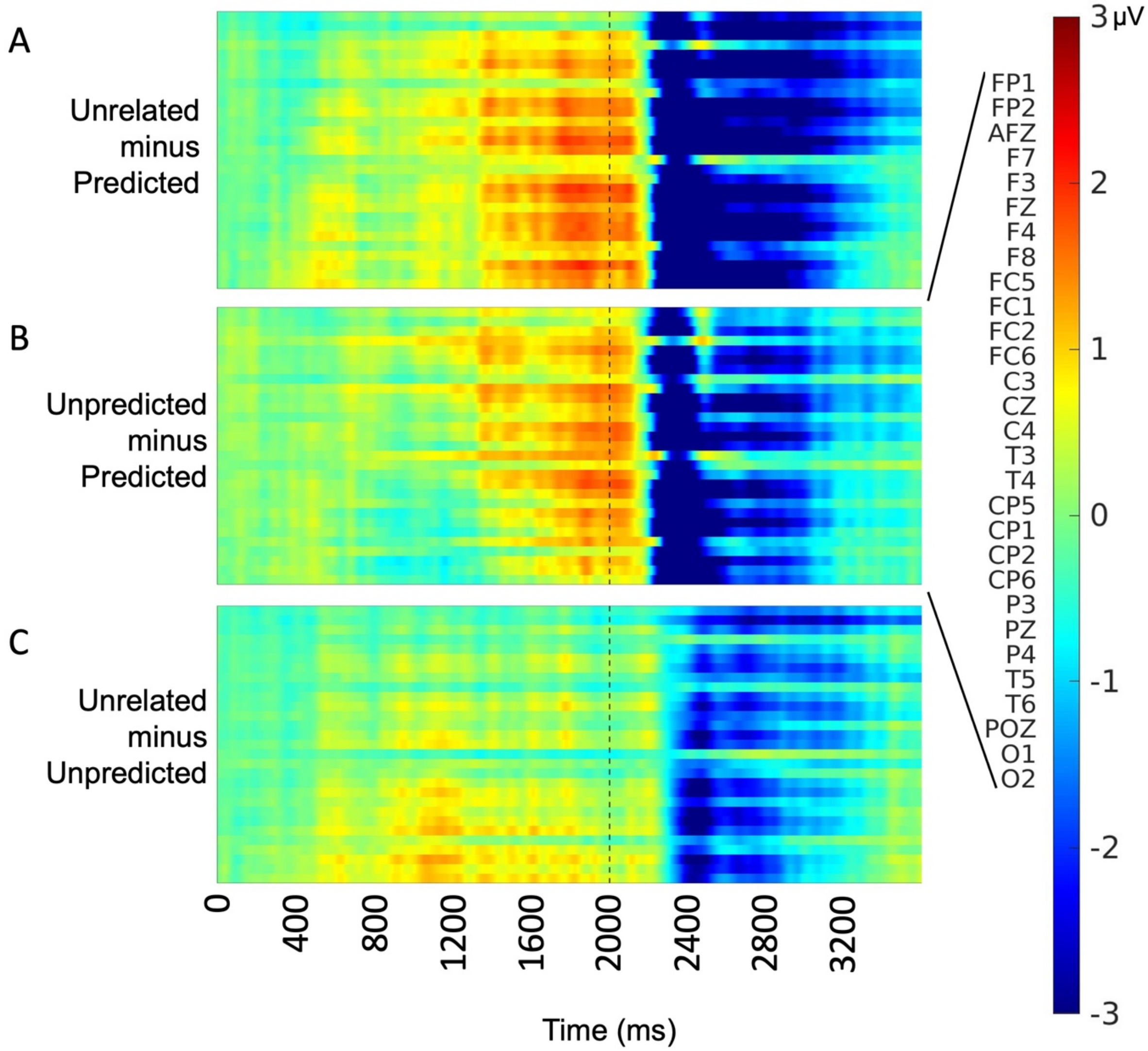

Figure S2. Shows the unmasked difference waves for main effects of target words. A) is unrelated minus predicted target words, B) is unpredicted minus predicted target words, and C) is unrelated minus unpredicted target words. Target word onset is 2000 ms.

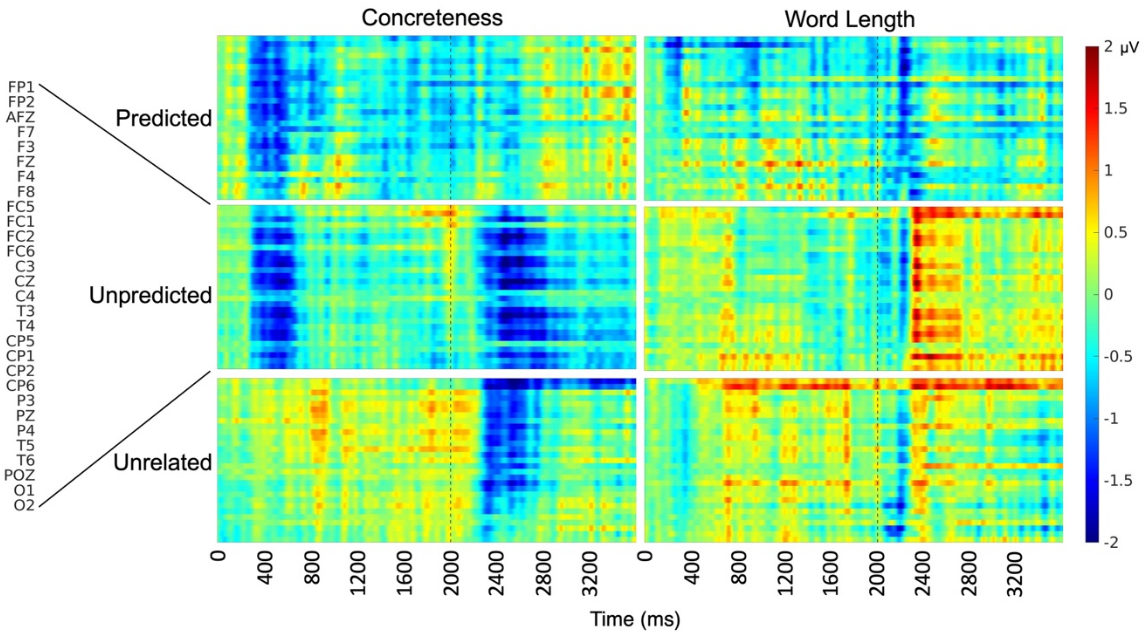

Figure S3. Shows the unmasked difference waves for feature effects of predicted target words, unpredicted – unpredicted related target word, and unrelated – unpredicted unrelated target word –. Target word onset is 2000 ms.

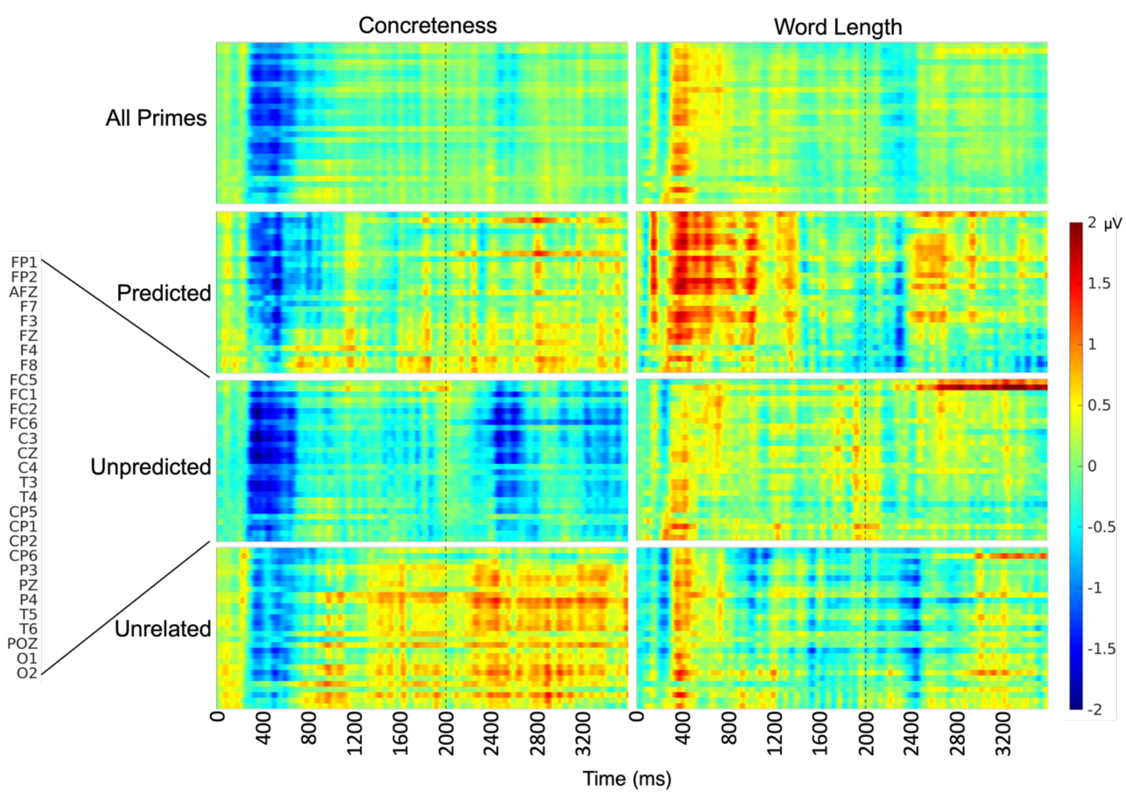

Figure S4. Shows the unmasked difference waves for feature effects of prime words. Conditions are relative to the target word: predicted – predicted target word, unpredicted – unpredicted related target word, and unrelated – unpredicted unrelated target word. Target word onset is 2000 ms.

Subject	Table S1. Length Artifact Rejection											
	Target											
	Predicted				Unpredicted				Unrelated			
	Acc. Long	Rej. Long	Acc. Short	Rej. Short	Acc. Long	Rej. Long	Acc. Short	Rej. Short	Acc. Long	Rej. Long	Acc. Short	Rej. Short
1	42	0	59	2	53	1	75	0	48	0	63	2
2	49	0	46	4	49	0	73	8	40	1	60	3
3	24	0	30	0	67	0	106	0	53	0	66	0
4	43	0	57	0	53	0	80	0	48	0	65	0
5	55	0	67	0	45	1	63	1	43	0	71	0
6	27	0	51	0	64	0	85	0	53	0	66	0
7	37	0	45	0	59	0	92	0	48	0	64	0
8	34	10	40	14	40	12	66	17	39	9	48	17
9	32	0	45	0	60	0	88	0	44	0	61	0
10	32	0	36	0	64	0	101	0	48	0	65	0
11	32	0	27	0	69	0	104	0	43	0	70	0
12	29	0	36	0	56	0	86	0	46	0	59	1
13	45	0	55	0	56	0	76	0	43	0	70	1
14	44	0	52	0	57	0	79	0	43	0	70	1
15	44	2	72	1	45	0	63	0	53	0	66	0
16	45	0	65	0	46	0	71	0	53	0	66	0
17	20	0	31	0	75	0	101	2	46	1	64	1
18	34	1	46	1	60	1	90	0	46	2	65	0
19	35	0	37	0	59	1	100	0	45	3	62	2
20	36	0	52	0	60	0	85	0	47	1	65	0
21	43	0	54	0	58	0	77	0	43	0	71	0
22	42	0	56	0	57	0	75	0	43	0	71	0
23	37	0	45	0	64	0	86	0	43	0	71	0
24	44	0	63	0	57	0	68	0	43	0	71	0
25	39	0	42	0	62	0	88	1	43	0	70	1
26	32	0	48	0	58	0	83	2	52	1	62	0
27	51	3	74	7	33	3	51	3	50	3	61	5
28	54	0	72	0	36	1	62	2	53	0	66	0
29	78	0	116	0	13	0	20	0	53	0	65	0
30	25	1	48	0	65	0	87	1	51	0	65	0
Subject	Prime											
	Predicted				Unpredicted				Unrelated			
	Acc. Long	Rej. Long	Acc. Short	Rej. Short	Acc. Long	Rej. Long	Acc. Short	Rej. Short	Acc. Long	Rej. Long	Acc. Short	Rej. Short

1	67	1	39	2	97	2	43	0	86	1	41	1
2	66	1	36	0	98	2	48	3	72	6	35	1
3	36	0	28	0	136	0	52	0	83	0	46	0
4	64	0	45	0	104	0	39	0	87	0	42	0
5	85	0	50	0	82	2	38	0	85	0	38	0
6	60	0	28	0	112	0	52	0	83	0	46	0
7	63	0	29	0	105	0	55	0	86	0	42	0
8	62	16	25	6	71	19	42	11	69	18	31	11
9	55	0	34	0	106	0	48	0	83	0	38	0
10	52	0	32	0	116	0	52	0	87	0	42	0
11	45	0	25	0	125	0	63	0	85	0	36	0
12	51	0	28	0	103	0	45	0	72	1	42	0
13	73	0	39	0	97	0	49	0	84	1	38	0
14	70	1	40	0	99	0	48	0	83	2	38	0
15	87	3	49	2	82	0	28	0	83	0	46	0
16	92	0	43	0	80	0	37	0	83	0	46	0
17	38	0	24	0	127	0	57	0	85	1	42	0
18	62	1	33	1	104	1	49	1	84	3	42	0
19	66	0	28	0	100	1	56	0	82	4	41	1
20	67	0	35	0	101	0	49	0	87	0	42	0
21	63	0	42	0	107	0	46	0	85	0	38	0
22	67	0	38	0	101	0	50	0	85	0	38	0
23	50	0	33	0	120	0	55	0	85	0	38	0
24	77	0	32	1	93	0	55	0	85	0	38	0
25	60	0	35	0	110	0	52	1	85	0	37	1
26	62	0	30	0	106	2	49	0	80	0	44	1
27	98	10	48	3	60	4	25	4	75	8	40	5
28	97	0	47	0	73	2	33	0	83	0	46	0
29	147	0	70	0	25	0	10	0	81	0	45	0
30	59	0	29	1	112	1	50	0	80	1	45	0

Table S1. Artifact rejection summary of word length for each condition.

Subject	Table S2. Concreteness Artifact Rejection											
	Target											
	Predicted				Unpredicted				Unrelated			
	Acc. High	Rej. High	Acc. Low	Rej. Low	Acc. High	Rej. High	Acc. Low	Rej. Low	Acc. High	Rej. High	Acc. Low	Rej. Low
1	69	1	74	3	90	0	80	2	78	2	78	2
2	60	2	64	2	95	1	82	9	75	2	68	5
3	33	0	44	0	127	0	116	0	80	0	80	0
4	73	0	68	0	87	0	92	0	80	0	80	0
5	91	0	75	0	67	2	84	0	80	0	80	0
6	52	0	56	0	108	0	104	0	80	0	80	0
7	64	0	50	0	96	0	110	0	80	0	79	0
8	52	16	52	13	71	21	77	18	57	23	70	10
9	50	0	66	0	105	0	89	0	76	0	75	0
10	54	0	48	0	106	0	112	0	80	0	80	0
11	42	0	44	0	118	0	116	0	80	0	78	0
12	43	0	52	0	99	0	96	0	73	1	69	0
13	72	0	66	0	88	0	94	0	80	0	79	1
14	66	0	66	1	94	0	93	0	79	1	79	1
15	76	1	91	4	83	0	64	0	80	0	80	0
16	72	0	88	0	88	0	72	0	80	0	80	0
17	39	0	30	0	116	0	125	3	77	2	80	0
18	68	1	47	1	90	1	111	1	79	1	78	2
19	56	0	58	0	103	1	101	0	75	4	78	2
20	57	0	68	0	103	0	92	0	79	1	80	0
21	60	0	72	0	100	0	88	0	80	0	80	0
22	61	0	78	0	98	0	81	0	80	0	80	0
23	49	0	62	0	111	0	98	0	80	0	80	0
24	69	1	76	0	90	0	84	0	80	0	80	0
25	55	1	61	0	103	1	99	0	79	1	80	0
26	60	0	57	0	97	1	100	1	76	1	78	0
27	85	9	98	4	59	5	54	4	72	8	73	6
28	85	0	91	0	73	2	68	1	80	0	80	0
29	130	0	143	0	30	0	17	0	78	0	78	0
30	54	1	53	0	104	1	107	0	78	1	78	0
Subject	Prime											
	Predicted				Unpredicted				Unrelated			
	Acc. High	Rej. High	Acc. Low	Rej. Low	Acc. High	Rej. High	Acc. Low	Rej. Low	Acc. High	Rej. High	Acc. Low	Rej. Low

1	59	1	84	3	65	0	105	2	72	2	84	2
2	48	3	76	1	79	1	98	9	60	2	83	5
3	29	0	48	0	112	0	131	0	58	0	102	0
4	60	0	81	0	65	0	114	0	74	0	86	0
5	69	0	97	0	60	2	91	0	67	0	93	0
6	48	0	60	0	93	0	119	0	58	0	102	0
7	57	0	57	0	68	0	138	0	74	0	85	0
8	42	17	62	12	50	16	98	23	60	14	67	19
9	48	0	68	0	75	0	119	0	67	0	84	0
10	47	0	55	0	78	0	140	0	74	0	86	0
11	36	0	50	0	96	0	138	0	65	0	93	0
12	43	0	52	0	86	0	109	0	50	1	92	0
13	58	0	80	0	74	0	108	0	66	1	93	0
14	54	0	78	1	78	0	109	0	66	1	92	1
15	71	2	96	3	68	0	79	0	58	0	102	0
16	64	0	96	0	77	0	83	0	58	0	102	0
17	33	0	36	0	87	0	154	3	73	1	84	1
18	51	1	64	1	72	1	129	1	73	1	84	2
19	44	0	70	0	81	0	123	1	72	2	81	4
20	47	0	78	0	78	0	117	0	74	0	85	1
21	47	0	85	0	85	0	103	0	67	0	93	0
22	58	0	81	0	73	0	106	0	67	0	93	0
23	47	0	64	0	85	0	124	0	67	0	93	0
24	54	1	91	0	77	0	97	0	67	0	93	0
25	44	0	72	1	88	0	114	1	66	1	93	0
26	54	0	63	0	83	1	114	1	56	1	98	0
27	79	8	104	5	47	5	66	4	53	5	92	9
28	79	0	97	0	61	1	80	2	58	0	102	0
29	120	0	153	0	21	0	26	0	58	0	98	0
30	50	0	57	1	91	0	120	1	58	0	98	1

Table S2. Artifact rejection summary for concreteness for each condition.

References

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition*, *136*, 135–149. https://doi.org/10.1016/j.cognition.2014.10.017

Casaponsa, A., Carreiras, M., & Duñabeitia, J. A. (2015). How do bilinguals identify the language of the words they read? *Brain Research*, *1624*, 153–166. https://doi.org/10.1016/j.brainres.2015.07.035

Chun, E., & Kaan, E. (2019). L2 Prediction during complex sentence processing. *Journal of Cultural Cognitive Science*, *3*(2), 203–216. https://doi.org/10.1007/s41809-019-00038-0

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2017). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition*, *20*(5), 917–930. https://doi.org/10.1017/S1366728916000547

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2019). Prediction and integration of semantics during L2 and L1 listening. *Language, Cognition and Neuroscience*, *34*(7), 881–900. https://doi.org/10.1080/23273798.2019.1591469

Dijkstra, T., & Heuven, W. J. B. van. (2002). Modeling bilingual word recognition: Past, present and future. *Bilingualism: Language and Cognition*, *5*(3), 219–224. https://doi.org/10.1017/S1366728902283017

Europarat (Ed.). (2020). *Common European framework of reference for languages: Learning, teaching, assessment ; companion volume*. Council of Europe Publishing.

Federmeier, K. D. (2022). Connecting and considering: Electrophysiology provides insights into comprehension. *Psychophysiology*, *59*(1), e13940. https://doi.org/10.1111/psyp.13940

Frenck-Mestre, C., German, E. S., & Foucart, A. (2014). Qualitative Differences In Native And Nonnative Semantic Processing as Revealed by ERPs. In R. R. Heredia & J. Altarriba (Eds.), *Foundations of Bilingual Memory* (pp. 237–255). Springer. https://doi.org/10.1007/978-1-4614-9218-4_12

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056. https://doi.org/10.1073/pnas.1216438110

Grisoni, L., Tomasello, R., & Pulvermüller, F. (2021). Correlated Brain Indexes of Semantic Prediction and Prediction Error: Brain Localization and Category Specificity. Cerebral Cortex, 31(3), 1553–1568. https://doi.org/10.1093/cercor/bhaa308

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research*, *28*(2), 191–215. https://doi.org/10.1177/0267658312437990

Grüter, T., Rohde, H., & Schafer, A. (2014). The role of discourse-level expectations in non-native speakers' referential choices. *Proceedings of the Annual Boston University Conference on Language Development*. https://par.nsf.gov/biblio/10028988-role-discourse-level-expectations-non-native-speakers-referential-choices

Gullifer, J. W., & Titone, D. (2020). Characterizing the social diversity of bilingualism using language entropy. *Bilingualism: Language and Cognition*, *23*(2), 283–294. https://doi.org/10.1017/S1366728919000026

Hale, J. T. (2011). What a Rational Parser Would Do. *Cognitive Science*, *35*(3), 399–443. https://doi.org/10.1111/j.1551-6709.2010.01145.x

Holcomb, P. J., & Grainger, J. (2007). Exploring the temporal dynamics of visual word recognition in the masked repetition priming paradigm using event-related potentials. *Brain Research*, *1180*, 39–58. https://doi.org/10.1016/j.brainres.2007.06.110

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research*, *29*(1), 33–56. https://doi.org/10.1177/0267658312461803

Hoversten, L. J., Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Language Membership Identification Precedes Semantic Access: Suppression during Bilingual Word Recognition. *Journal of Cognitive Neuroscience*, *27*(11), 2108–2116. https://doi.org/10.1162/jocn_a_00844

Hoversten, L. J., Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Early processing of orthographic language membership information in bilingual visual word recognition: Evidence from ERPs. *Neuropsychologia*, *103*, 183–190. https://doi.org/10.1016/j.neuropsychologia.2017.07.026

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*(1), 57–83. https://doi.org/10.1016/j.cognition.2012.10.013

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism*, *4*(2), 257–282. https://doi.org/10.1075/lab.4.2.05kaa

Kaan, E. (2023). The neurocognition of prediction in second language processing and learning. In *The Routledge Handbook of Second Language Acquisition and Neurolinguistics*. Taylor-Francis. https://www.taylorfrancis.com/chapters/edit/10.4324/9781003190912-31/neurocognition-prediction-second-language-processing-learning-edith-kaan

Kuperberg, G. R. (2021). Tea With Milk? A Hierarchical Generative Framework of Sequential Event Comprehension. *Topics in Cognitive Science*, *13*(1), 256–298. https://doi.org/10.1111/tops.12518

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language*, *63*(4), 447–464. https://doi.org/10.1016/j.jml.2010.07.003

Martin, C. D., Thierry, G., Kuipers, J.-R., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language*, *69*(4), 574–588. https://doi.org/10.1016/j.jml.2013.08.001

Newman, A. J., Tremblay, A., Nichols, E. S., Neville, H. J., & Ullman, M. T. (2012). The Influence of Language Proficiency on Lexical Semantic Processing in Native and Late Learners of English. *Journal of Cognitive Neuroscience*, *24*(5), 1205–1223. https://doi.org/10.1162/jocn_a_00143

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), Article 1. https://doi.org/10.1038/4580

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Chapter 5

General Conclusion

The studies described in this dissertation utilized a combination of MVPA classification and mass univariate cluster analysis to examine the time course of predictive processing during visual word recognition. Specifically, the experiments aimed to 1) establish the best-performing decoding tool for decoding results from visual word priming paradigms from a selection of readily accessible options, 2) specify the time course of pre-activation of lexico-semantic and visual word features in native American-English speakers, and 3) compare that time course with the time course of Spanish-English bilinguals reading in L2 (English). The time course of feature pre-activation has critical implications for determining the architecture behind predictive processing in both native and non-native readers. If readers are pre-activating word information in a top-down fashion and at all levels of representation, then that would provide evidence of an automatic predictive coding model in which higher levels in the cortical hierarchy are continuously making top-down predictions, then calculating prediction error which is passed up to update the readers' higher-level representations.

The study in Chapter 2 established SVM as a superior option from three highly accessible EEG decoding machine learning algorithms. We then used SVMs – in combination with mass univariate ERP analyses – to provide evidence that when during prediction tasks, readers can pre-activate lexico-semantic and sublexical visual features and whether they do so in a top-down fashion – higher level predictions, such as anticipated semantic features, are made first and those predictions inform the predictions about lower-level features, like word length – as would be expected within a predictive coding model of language processing (Kuperberg & Jaeger, 2016). Additionally, predictive coding models require prediction error to update higher level representations (Rao & Ballard, 1999). One important aspect of the N400 ERP component is that it has been shown to be sensitive to semantic expectation violations (Kutas & Hillyard, 1980) and can be considered a reflection of prediction error within predictive processing in language (Eddine et al., 2023). The reduced N400 effects when participants successfully predict a word and increased N400 when predictions are incorrect are both reflective of this prediction error calculation within the present study. Next, we looked at predictive processing within bilinguals using the same paradigm and analyses. The bilingual group in Chapter 4, not only showed evidence of lexico-semantic prediction and visual feature pre-activation, but the visual feature decoding was clearly before target onset. We found similar evidence of prediction error within our bilingual study as well. These findings suggest that bilinguals predict in L2 in a similar fashion to monolinguals in their L1. Together these findings provide compelling evidence in support of predictive coding models of anticipation within language processing whether bilingual or not and regardless of whether using L1 or L2.

Prior studies have suggested the possibility that while bilinguals may predict in L2 in a similar fashion to native speakers, there may be latency delays in processing these predictions (Frenck-Mestre et al., 2014; Kaan, 2023; Newman et al., 2012). Although there are no direct statistical comparisons of timings between the native English speakers in Chapter 3 and the bilinguals in Chapter 4, the overall patterns of results did not suggest any delays in prediction timing for the Spanish-English bilinguals relative to the native English speakers. Indeed, within bilinguals pre-activation of visual features was observed earlier than in monolinguals which may be due to bilinguals' reliance on word length for language identification (Casaponsa et al., 2015). An important caveat to this finding is that due to some overlap between prime and target word pre-target concreteness effects in both types of analyses it is difficult to separate out precise onsets for target word pre-activation. Thus, it is possible that there are differences in the timing of these effects between the two groups that could not be isolated by this experimental setup.

Gaining a more precise account of the prediction time course could be partially addressed through modification of the current paradigm. One limitation of the paradigm is that in unpredicted trials whether the target word is related or unrelated, there is no way of knowing if the participant successfully settled on a prediction. For example, after seeing the prime word “circus” a participant may predict “acrobat” or the participant may be unable to come up with a prediction before the target word appears. In either case, this would fall under the appropriate unpredicted condition. Due to this, there will be additional noise within these conditions which can impact the ability to detect differences within these conditions. One solution to this problem would be to modify the paradigm to have participants explicitly identify their predicted word, either through vocalization or by reporting their predicted word post-trial. Such a modification would allow more clarity on the prediction time course when encountering an unsuccessfully predicted but related word. Additionally, we achieved reliable decoding of these lexical and sub-lexical features despite not explicitly manipulating the concreteness or word length of target words. While this is promising for more naturalistic settings to study these feature activations, it may be necessary to explicitly manipulate semantic features to gain a more precise separation of the effects of concreteness within prime and target words.

Overall, the evidence from these studies suggests that both native speakers and non-native speakers engage in predictive processing in a similar top-down fashion while calculating prediction error when their predictions are incorrect. This aligns well with predictive coding accounts of language processing (Eddine et al., 2023; Kuperberg & Jaeger, 2016). However, the studies highlight the need for further investigation into the precise prediction time course differences between native and non-native speakers. Moreover, these studies also highlight the possibility that prediction may not always occur at all levels of processing, and this may be particularly true of native speakers. Finally, these studies were all conducted at the word level within a priming paradigm. Thus, there is a need to perform similar analyses at other language processing levels – such as within sentences – and in more natural settings which do not explicitly instruct the participant to predict.

With the rise of predictive coding as a potential unifying theory in cognitive processing, including language processing, the study of the time course of predictive processing during language comprehension is becoming increasingly important. The present work contributes to our understanding of this time course in both first and second languages.

References

Eddine, S. N., Brothers, T. A., Wang, L., Spratling, M., & Kuperberg, G. (2023). A Predictive Coding Model of the N400. *bioRxiv*, 2023–04.

Frenck-Mestre, C., German, E. S., & Foucart, A. (2014). Qualitative Differences In Native And Nonnative Semantic Processing as Revealed by ERPs. In R. R. Heredia & J. Altarriba (Eds.), *Foundations of Bilingual Memory* (pp. 237–255). Springer. https://doi.org/10.1007/978-1-4614-9218-4_12

Kaan, E. (2023). The neurocognition of prediction in second language processing and learning. In *The Routledge Handbook of Second Language Acquisition and Neurolinguistics*. Taylor-Francis. https://www.taylorfrancis.com/chapters/edit/10.4324/9781003190912-31/neurocognition-prediction-second-language-processing-learning-edith-kaan

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology*, *11*(2), 99–116. https://doi.org/10.1016/0301-0511(80)90046-0

Newman, A. J., Tremblay, A., Nichols, E. S., Neville, H. J., & Ullman, M. T. (2012). The Influence of Language Proficiency on Lexical Semantic Processing in Native and Late Learners of English. *Journal of Cognitive Neuroscience*, *24*(5), 1205–1223. https://doi.org/10.1162/jocn_a_00143

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), Article 1. https://doi.org/10.1038/4580