

Context is Key: The Importance of Deep Reading Comprehension

A dissertation submitted by

Edward J. Alexander

in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in

Psychology

Tufts University

August 2024

© 2024, Edward J. Alexander

Advisor: Gina Kuperberg, M.D., Ph.D.

Abstract

Psycholinguistic research aims to understand how people make sense of language in their everyday lives. However, most of this research studies language under experimental conditions in which people are instructed to specifically monitor (and indicate) when there is a breakdown in their understanding. Moreover, there is an assumption that people equally attend to the individual, carefully crafted sentences in our experiments as they would to more naturalistic discourses. There is increasing evidence, however, that the depth with which participants attend to single sentences is shallower than the deep style of comprehension that people use when reading longer narratives. This work has shown that the depth of comprehension has a vital impact on the cognitive operations involved in language comprehension, and on the resulting signals observable in electroencephalography (EEG), such as the N400 and P600 components. In this thesis, I present three studies: in the first, participants viewed sentences in a context where the detection of errors was important to their goal of understanding, and were instructed to either correct errors or repeat them verbatim to determine if the P600 is related to error correction processes. In the second, in order to determine if the P600 is directly tied to deep comprehension, I contrasted isolated non-naturalistic sentences with naturalistic discourses. Finally, I use these same naturalistic stimuli and collect predictability measures on every word to determine if lexical predictability interacts with frequency and semantic relatedness in naturalistic contexts.

Table of Contents

Chapter 1: General Introduction	1
Chapter 2: The P600 does not reflect error correction	4
Chapter Introduction	4
Introduction	4
Methods	9
Development of materials	9
Participants	10
Procedure and tasks	11
EEG data acquisition	12
Behavioral data analysis	13
Speech onset latencies	13
Correction Distance	14
ERP Data: Statistical analysis	14
Relationship between Correction Distance and ERP amplitude	16
Results	16
Behavioral results	16
Acceptability judgments: Accuracy and reaction times	16
Speech onset latencies	16
Correction Distance	17
ERP Results	18

600–1000ms.....	18
1000-1400ms	19
300-600ms	20
Single trial analysis: Relationship between late positivity effects and Correction Distance..	20
Discussion.....	22
The P600 is unlikely to purely reflect an explicit task-related categorical judgment	23
The P600 may be linked to prolonged decision making during reprocessing of the anomalous input	24
A later frontally distributed positivity may reflect the successful re-establishment of plausibility	25
Chapter 3: Contributions of Naturalistic Stimuli on the P600.....	28
General Introduction.....	28
Study 1 (Experiment 1a and 1b).....	34
Introduction	34
Methods	35
Participants	35
Design and Materials	36
Tasks and Procedure.....	36
EEG Data Acquisition and Preprocessing	37
ERP Statistical Analysis Strategy.....	38
Results	39
Behavioral Results.....	39

ERP Results	39
P600 (600-1000ms)	39
N400 (300-500ms).....	40
Discussion.....	40
Study 2 (Experiment 2).....	42
Introduction	42
Methods	43
Participants	43
Design and Materials	43
Procedure	44
Results	46
Behavioral Results.....	46
ERP Results	46
P600 (600-1000ms)	46
N400 (300-500ms).....	47
Discussion.....	47
General Discussion	49
The P600 is evoked when error detection disrupts deep language comprehension	50
Chapter 4: Contributions of Naturalistic Stimuli on the N400.....	61
Introduction	61
Present Study	65

Methods	67
Participants	67
Construction of Stimuli	67
Operationalization of Lexical Characteristics	68
Operationalization of Lexical Predictability (Cloze Probability).....	68
ERP Procedure.....	70
EEG recording	71
Data analysis.....	72
Results	73
1. Do frequency and lexical predictability interact in naturalistic contexts?.....	73
2. What is the role of semantic relatedness in naturalistic contexts?	75
Methods	75
Results	76
General Discussion	79
1. Interaction of frequency and predictability	79
2. Effects of semantic relatedness and predictability	80
2.1. Interactive effects of Semantic relatedness and predictability	80
2.2. Interactive effects of Semantic relatedness and predictability	82
3. Hierarchical generative models and predictive coding can explain interactions between local and contextual variables.....	83
4. Future directions	84
Chapter 5: General Discussion	86

Summary of findings	86
Broader implications.....	88

List of Figures

- Figure 1:** *Top:* Grand-average event-related potentials (ERPs) evoked by plausible (black) and anomalous words (red) in the Repetition task (left) and the Correction task (right), plotted at three midline electrode sites (FPz, CPz, Oz). Positive voltage is plotted downward. *Bottom:* Voltage maps showing the scalp topographies of the ERP effects produced by the anomalous *minus* plausible words within the 300-500ms, 600-1000ms and 1000-1400ms time windows. 18
- Figure 2** *Top:* An example of an anomalous sentence and its correction, with a Correction Distance of 6 characters, including both deletions (red) and insertions (green). *Bottom Left:* A scatter plot showing the relationship between the amplitude of the very late frontal positivity (averaged across FPz/2, Fz/4, AF4; 1000-1400ms) and the Correction Distance between anomalous sentences and participants' verbal responses. *Bottom Right:* A topographic voltage map, showing the frontal scalp distribution of the Correction Distance effect. 21
- Figure 3.** Self-paced reading times as a function of sentence position across all experiments. Each data point depicts the overall average time in milliseconds, with error bars representing ± 1 standard error of the mean calculated within participants. Sentence-final words could appear either two or three words past the critical word. 55
- Figure 4.** Grand-average event-related potentials (ERPs) in Experiments 1a and 1b, plotted at Pz. The bar graphs show the average voltages within our spatiotemporal region of interest (600-1000ms, CP1/2, Pz/3/4, Oz/1/2) for the semantic and syntactic P600 effects with error bars showing ± 1 standard error of the mean calculated within-participants. Topographic voltage maps show the distribution of the P600 effects over the scalp surface within the same region of interest. 59
- Figure 5.** Grand-average event-related potentials (ERPs) in Experiments 1a and 2, plotted at Pz. The bar graphs show the average voltages within our spatiotemporal region of interest (600-1000ms, CP1/2,

Pz/3/4, Oz/1/2) for the semantic and syntactic P600 effects with error bars showing ± 1 standard error of the mean calculated within-participants. Topographic voltage maps show the distribution of the P600 effects over the scalp surface within the same region of interest. 60

Figure 6. Averaged cloze probability to content words as a function of word-in-sentence position relative to the final word of sentence 1. Regression lines are calculated within each sentence individually, and gray regions represent standard error. Opacity of data points represents the number of data points for each word position. 70

Figure 7. Averaged N400 amplitudes to content words in our spatial region of interest (P3/z/4, C3/z/4, CP1/2) plotted with all trials (left), only trials in the top third of values for cloze probability (center), and only trials in the bottom third of values for cloze probability (right). The gray line represents the grand average ERP within this spatial region of interest. To compute the effect of frequency, ERP data was binned temporally into 25ms time windows, and model 2 was fit individually at each time point. At each time point, the dotted line represents the estimated ERP amplitude to words that are 1 standard deviation higher (dark blue) or 1 standard deviation lower (light blue) frequency than average, based on the beta weights extracted from the models. 74

Figure 8. Left: Model-estimated partial effects (beta weights) for the effect of cloze probability and semantic relatedness with preceding content words over time. The semantically associated words that immediately preceded the content word (n-1) reduced the amplitude of the N400 response, but there were no significant effects of semantic relatedness for earlier sentence positions. Right: Averaged N400 amplitudes to content words in our spatial region of interest (P3/z/4, C3/z/4, CP1/2) plotted with trials with above median values for cloze probability (top right), and only trials with below median values for cloze probability (bottom right). The gray line represents the grand average ERP within this spatial region of interest. To compute the effect of semantic relatedness, ERP data was binned temporally into 25ms time windows, and model 1 was fit individually at each time point. At each time point, the dotted line

represents the estimated ERP amplitude to words that are 1 standard deviation more semantically related to the preceding content word (dark blue) or 1 standard deviation less semantically related (light blue) than average, based on the beta weights extracted from the models. 78

List of Tables

Table 1. Example stimuli with critical word characteristics (means, with standard deviations in parentheses)	10
Table 2. Behavioral results with means and standard deviations	17
Table 3. Significant results from the 2×2 cluster-mass permutation analyses	19
Table 4. Mean accuracies for responses to questions in Experiments 1a, 1b, and 2. Standard deviations are reported in parentheses. Accuracy represents percentage of correct responses to intermittent comprehension questions (Experiment 1a), explicit acceptability judgments (Experiment 1b), or a combination of multiple-choice and true/false comprehension questions (Experiment 2).	57
Table 5. Stimuli used in each experiment. Stimuli were identical between Experiments 1a and 1b. For Experiment 2, stimuli from Experiments 1a and 1b were embedded in larger contexts as described in Methods.	58
Table 6. Effects of Frequency and Cloze on the N400	74
Table 7. Effect of frequency, follow-up analyses	75
Table 8. Effects of Semantic Relatedness with prior words and Cloze on the N400	77

Chapter 1: General Introduction

Comprehending written language is among the most impressive human capacities. We are able to quickly infer meaning as we read, incrementally building up the underlying message of a sentence as we proceed through it word-by-word (Van Petten & Kutas, 1990, 1991; see discussion in Kuperberg, Brothers, & Wlotko, 2020; Brothers, Wlotko, Warnke, & Kuperberg, 2020). Perhaps most impressively of all, we can accomplish this despite obstacles such as ambiguous phrasing, unfamiliar handwriting, misspellings, and outright conceptual errors in our input, from which we can quickly recover.

Electroencephalography (EEG), a method for measuring electrical activity in the brain in real time, has enabled researchers to directly study the cognitive processes involved in error recovery and incremental language comprehension. In particular, averaging multiple EEGs that are time-locked to particular events of interest (resulting in an averaged waveform known as an Event-Related Potential or ERP) allows us to identify consistent patterns in brain activity following stimulus presentation, known as components. Studying how errors and unexpected input in language affect these components gives us a window into how the brain recognizes and recovers from this input.

In this thesis, we examine two ERP components that have been influential in the language comprehension literature: the P600, a positive-going component often observed 600-1000ms after an anomalous or dispreferred word which has been related to error monitoring and reprocessing, and the N400, a negative-going component often observed 300-500ms after unpredicted words which is sensitive to the difficulty of lexico-semantic access of the input in context.

The functional interpretation of the P600 is particularly controversial. It is unclear whether it indexes detection of conflict, syntactic reanalysis, or actual correction of the input – processes which are often conflated in descriptions of the P600. Some have even questioned its relevance to language cognition, arguing that it instead reflects processes relevant to artificial task demands such as explicit acceptability judgments.

In Chapter 2, we examined the relationship between error correction and the P600 by asking participants to speak aloud either corrected or unaltered versions of anomalous sentences. The experiment was calibrated to make error correction difficult when participants were asked to repeat anomalies. If P600 effects to anomalous sentences are unaffected by the task demands, then the P600 does not index the process of error correction. Importantly, however, this chapter does not directly address the question of the P600's relevance specifically to natural reading, as both the tasks and the stimuli are unlike any typical natural comprehension.

In Chapter 3, having established the role of the P600 within a tightly controlled experimental manipulation, we return to the question of its relevance to real-world language cognition. In order to test this, we report three experiments with variations on the same stimuli and different tasks. In the first and second experiments, we presented single-sentence versions of the stimuli, while participants either answered infrequent comprehension questions (Experiment 1a) or made explicit acceptability judgments (Experiment 1b). We then contrast this directly to an expanded version of the stimuli and task, where we encourage deep comprehension by embedding stimuli in narrative contexts and adding a variety of challenging comprehension questions, which ask participants to make inferences about the content, remember details, or provide continuations of stories (Experiment 2). Critically, in this experiment, detecting the presence of anomalies is never directly useful to the demands of the tasks, except in service of general comprehension. If the P600 indexes solely the cognitive operations underlying explicit judgments, then we should observe enhancement of the P600 only when explicit judgments are required in the Experiment 1b, and not in the more naturalistic Experiment 2.

In Chapter 4, we use these same naturalistic contexts and tasks to answer questions about the real-world language cognition relevance of a different ERP component, the N400, which is known to relate to the predictability of stimuli from context. Instead of examining critical words that are expected or anomalous, we tag and analyze thousands of distinct content words in our stimuli and collect predictability measures for every word. We then fit a series of mixed effects regression models to address whether effects of lexical predictability in naturalistic contexts can override effects of lower-level factors

on the N400 such as word frequency and semantic relatedness. Uniquely, this experiment collects ERPs to every content word over a collection of unrelated paragraphs.

Finally, in Chapter 5, we summarize the findings of the previous chapters, and argue that ERP research must incorporate both controlled experimental and task-free naturalistic designs, and that successful reading comprehension requires establishment of a situation model. By understanding how these components respond in more demanding, naturalistic contexts, we open the door to better understanding how the brain deals with difficult or unexpected input in everyday language. This could inform our understanding of how developing children and adults learn language and develop better interventions and teaching strategies.

Chapter 2: The P600 does not reflect error correction

Chapter Introduction

I am particularly indebted to Dr. Trevor Brothers for collaboration on this project. We collaborated on the original concept, paradigm, and stimuli, and he assisted me throughout the data collection process. My advisor, Dr. Gina Kuperberg, MD. Ph.D., also assisted extensively with the writing, helping me connect multiple literatures, and providing me with helpful references -- too many times to count! Finally, I'd like to thank Cassidy Prejean, Elizabeth Kenneally, and Maya Lazarus, who assisted with data collection and analysis, and Sophie Greene, who assisted with data analysis.

The chapter presented here is based on an early version of a manuscript. As it was difficult for me to physically attend campus due to the COVID-19 pandemic, and because I was starting full time employment, Gina Kuperberg and Trevor Brothers graciously took over this project, ran an additional experiment to supplement the results presented here, and developed a new theory of the P600. The full body of work is currently under review in the Journal of Experimental Psychology.

This work was funded by the National Institute of Child Health and Human Development (R01 HD082527 to G.R.K.).

Introduction

Language comprehension requires us to infer a communicator's intended message as linguistic inputs unfold quickly in real time. However, this process is sometimes complicated by the presence of errors – misinterpretations, misperceptions or anomalies in the input that can disrupt the process of comprehension. Upon detecting an error, instead of abandoning comprehension altogether, the brain often engages in additional neural analysis, but the precise nature of this processing remains unclear. In this study, we addressed this question using a novel within-subjects task manipulation. This approach allowed us to determine whether the neural activity produced linguistic error reflects reprocessing of the prior

input, or whether it additionally reflects attempts to correct the error and establish a new plausible interpretation¹.

The P600 is a positive-going event-related potential (ERP) with a posterior scalp distribution, which is often observed 600-1000ms after comprehenders encounter an anomalous or dispreferred continuation. P600 effects are often elicited by syntactic violations (e.g. Osterhout, 1993; Osterhout and Holcomb, 1992; Hagoort, Brown, & Groothusen, 1993), orthographic violations (e.g. Munte, Heinze, Matzke, Wieringa & Johannes, 1998; Vissers, Chwilla & Kolk, 2006; Ito & Nieuwland, 2016), and, of most relevance to this paper, semantic violations (e.g. Kuperberg, Sitnikova, Caplan & Holcomb, 2003; Kolk, Chwilla, van Herten, & Oor, 2003; Hoeks, Stowe and Doedens, 2004; Kim & Osterhout, 2005; see Kuperberg, 2007, for a review). Critically, however, a P600 is not *always* produced by semantic anomalies (e.g. Ainsworth-Darnell, Shulman & Boland, 1998; Gunter, Friederici, & Hahne, 1999; Friederici, Pfeifer & Hahne, 1993; Osterhout & Nicol, 1999; Brothers, Wlotko, Warnke, & Kuperberg, 2020, Experiment 1). Rather, for a semantic anomaly to produce a P600 effect, it must be perceived as disrupting the process of comprehending the high-level message currently being communicated. Consequently, whether or not a semantic anomaly triggers a P600 can be influenced by multiple internal and external factors that impact comprehension (see Kuperberg, 2007 for a review). These include task demands (Kolk et al., 2003; Geyer et al., 2006; Xiang & Kuperberg, 2005; Royle, Drury, & Steinhauer, 2013; Alexander, Brothers & Kuperberg, in preparation), the presence of a prior context that establishes a situation model (Nieuwland & Van Berkum, 2005; Brothers, Wlotko, Warnke, & Kuperberg, 2020; Alexander, Brothers & Kuperberg, in preparation), and whether the anomaly is “semantically attracted” to preceding words in the context (Kuperberg, Sitnikova, Caplan & Holcomb, 2003; Kolk, Chwilla, van Herten & Oor, 2003; Kim & Osterhout, 2005, Experiment 1). For example, a large P600 effect is

¹ Throughout this chapter, I will use the term “reprocessing” to refer to any re-evaluation of the prior input or prior interpretation, at any level of representation (without necessarily involving any attempt to correct the error), and we will use the term “error correction” in a descriptive fashion to refer to any attempt to correct the error and establish a coherent interpretation after detecting an anomaly, either through active editing of the input (correction), or through shifts in prior beliefs about its surface structure.

produced by words like “eat” in sentences like “*Every morning at breakfast the eggs would *eat*” because the combination of the words, “*breakfast*”, “*eggs*” and “*eat*” provides highly reliable evidence for a specific familiar event (see Kuperberg, 2016). The semantic anomaly on “*eat*” *conflicts* with the communication of this message, disrupting the process of incremental comprehension and committing the comprehender to additional downstream processing, reflected in the posterior P600 response.

Although there has been debate about the precise nature of the conflict that triggers the P600 (Kuperberg, 2007; van de Meerendonk, Kolk, Chwilla, & Vissers, 2009; Kim & Osterhout, 2005), most models have assumed that the purpose of this additional processing is to *correct* the eliciting error in order to re-establish coherence. This assumption was inherited from earlier models of syntactic processing, in which this additional processing was thought to reflect a process of “revision” of an initial parse, either during a second stage of processing (in serial models, Frazier & Rayner, 1982, van Gompel, Pickering, & Traxler, 2001) or through competition between multiple structures, activated in parallel (Just & Carpenter, 1992; Spivey & Tanenhaus, 1998; MacDonald, Pearlmutter & Seidenberg, 1994; Jurafsky, 1996, or “correction” – the process of editing ill-formed anomalous constructions (Kaan & Swaab, 2003, but see Fodor & Inoue, 1994). More recent “noisy channel” frameworks assume that comprehenders retain some uncertainty about the surface structure of prior linguistic inputs, and so the divisions between “revision” (a shift in belief about prior input) and “correction” (edits to the surface structure) are less clear-cut (Levy, 2008; 2009; 2011; Gibson, Bergen & Piantadosi, 2013; Ryskin, Stearns, Bergen, Eddy, Fedorenko, & Gibson, 2021). In these frameworks too, however, a basic assumption is that error-driven processing, reflected in the P600, functions to correct an error and re-establish coherence.

This assumption, however, may be wrong. As argued by Herman Kolk and colleagues, it is possible that the detection of an error instead leads comprehenders to reprocess the critical word and aspects of the prior context, simply to gather more information and determine whether the input was correctly perceived the first time around (van Herten, Kolk & Chwilla, 2005; see van de Meerendonk et al., 2009, for a review). Importantly, on this account, reprocessing is not necessarily a precursor to error correction, and these two processes can be dissociated.

The goal of the present study was to distinguish between these accounts by determining whether the posterior P600 is linked to reprocessing, error correction, or some combination of the two. Participants read plausible and semantically attracted anomalous sentences (e.g. “*The dark caves were exploring/explored*”) as ERPs were recorded at the scalp surface. Critically, we used a within-subjects task manipulation to influence the likelihood of error correction across blocks. In both tasks, participants first judged whether or not each sentence was acceptable. This ensured that they consciously detected any errors (see Sanford et al. 2011). In the “Repetition” task, participants then simply repeated the sentence that they had just read. This allowed for reprocessing but discouraged any attempts to correct the anomalous sentences. In the “Correction” task, participants spoke each sentence aloud after first correcting any semantic anomalies. This resulted in a 2×2 within-subjects design, crossing Task (Repetition or Correction) and Sentence Type (Plausible or Anomalous).

We considered three possible outcomes. The first was that the P600 *only covaries* with reprocessing, but is not directly linked to attempts to re-establish coherence through error correction (cf. van de Meerendonk et al., 2009). For example, after reading, “*The dark caves were exploring*”, participants might reprocess the input to verify that it was indeed anomalous, but this reprocessing would not entail actively editing the input or shifting to an alternative interpretation. This account would predict a larger P600 to anomalous (versus plausible) sentences in both the Repetition and Correction tasks, with no difference in the magnitude of this effect across tasks.

A second possibility is that reprocessing necessarily involves error correction in attempts to re-establish a coherent interpretation, which is reflected in the P600 response. These attempts at error correction may involve actively correcting the surface form of the input (see Kim & Osterhout, 2005; Kim & Sikos, 2011), or, within a noisy channel framework, they may entail shifting to an alternative, lower-probability belief about its surface structure or interpretation (Levy, 2008; Gibson, Bergen & Piantadosi, 2013; Ryskin et al., 2021). These accounts would predict a robust posterior P600 effect to the anomalous (*versus* plausible) sentences in the Correction task, but no such effect in the Repetition task, where attempts to correct the error should be blocked.

Finally, we considered the possibility that the posterior P600 effect is linked to reprocessing, but that error correction and attempts to re-establish coherence would be reflected in a separate ERP component with a different latency or time-course. One potential candidate is a *frontally distributed* late positivity. Late frontal positives between 600-1000ms are typically produced by unexpected but plausible continuations that trigger a large shift in the situation model, either because they are inconsistent with a previously established schema (Kuperberg, Brothers, & Wlotko, 2020; Brothers, Wlotko, Warnke, & Kuperberg, 2020) or because they are highly informative about a new, unpredicted schema (Brothers, Greene, & Kuperberg, 2020). Late positivities with a frontal or widespread distribution are also sometimes produced by dispreferred continuations in syntactically ambiguous (e.g. Osterhout & Holcomb, 1992; Van Berkum, Brown & Hagoort, 1999; Kaan & Swaab, 2003) and syntactically complex sentences (e.g. Friederici, Hahne & Saddy, 2002) – situations in which it is likely that comprehenders are able to re-establish coherence through syntactic revision (but see Kaan, Harris, Gibson & Holcomb, 2000; Kaan & Swaab, 2003). Thus, in the current experiment, it is possible that semantic anomalies will generate a posteriorly distributed P600 effect in both the Repetition and the Correction tasks (reflecting reprocessing), but that an additional frontal positivity would be elicited by anomalies that are explicitly corrected. This late frontal positivity might be observed at the same time as the posterior P600 effect, which would suggest that reprocessing and error correction proceeded in parallel. Alternatively, it may onset at a still later stage of processing, suggesting that attempts to re-establish coherence in the Correction task begin only after an initial stage of reprocessing.

In addition to this between-task comparison, we also asked how the magnitude of linguistic corrections influenced online neural activity, trial-by-trial. For this analysis, we transcribed participants' responses and calculated the *Correction Distance* between each presented sentence and the participant's verbal correction. We then fitted linear mixed effects models to determine whether, in the anomalous sentences, this continuous measure of Correction Distance explained additional variance in the amplitude of late ERP responses, over and above any categorical effect of Task.

Methods

Development of materials

Each participant read 192 sentences in total: 96 anomalous and 96 plausible. All the sentences followed the form “[article/pronoun] [adjective] [noun] [was/were/had been] [verb]”, and, in all cases, the critical verb was semantically attracted to the preceding noun (see Table 1 for examples). Of the 96 anomalous sentences seen by each participant, 32 had inanimate subjects, and ended with an -ing verb that could be corrected syntactically by changing it to -ed (e.g. “*The overgrown lawn was mowing*” → “*mowed*”). Another 32 sentences had animate subjects, and ended with an -ed verb that could be syntactically corrected by changing it to “-ing” (e.g. “*The folk musicians were strummed*” → “*strumming*”). The remaining 32 sentences had inanimate subjects that ended either with an -ing verb (17 sentences) or with an -ed verb (or an irregular past-tense completion; 13 sentences), which required more substantial edits to restore coherence, for example by adding a new argument (e.g. “*The judge’s gavel was *pardoned*” → “*The judge pardoned the defendant*”).

The 64 syntactically correctable sentences and their plausible counterparts were constructed in quadruplets. We began with a set of 128 verbs that could plausibly take an animate or inanimate subject (e.g. “mow”). We wrote four sentences for each verb — two with an animate subject and two with an inanimate subject — such that the verb’s inflection rendered the sentence either plausible or anomalous (e.g. “*Their responsible son was mowing/mowed*”; “*The overgrown lawn was mowing/mowed*”). These items were rotated across four lists.

The 32 remaining anomalous sentences and their plausible counterparts were constructed in pairs. We began with 64 inanimate subjects and combined each of these with a semantically associated verb that rendered the sentence either plausible or anomalous (e.g. “*The judge’s gavel was banged/pardoned*”). Within this set, we matched the plausible and anomalous verbs in length (number of letters), frequency (SUBTLEX-US, log per million; Brysbaert & New, 2009), concreteness (Brysbaert, Warriner, & Kuperman, 2014), and orthographic neighborhood density (Yarkoni, Balota, & Yap, 2008), see Table 1.

Across conditions, we also matched the semantic attraction between the verb and the preceding context, operationalized using Latent Semantic Analysis (LSA: Landauer, Foltz, & Laham, 1998; Landauer & Dumais, 1997). For these analyses, semantic similarity values (SSV) were extracted using the pairwise comparison tool (document-to-term) at <http://lsa.colorado.edu/> (all $ts < 1.33$, all $ps > .18$). These items were also rotated across the four lists. Thus, each scenario and critical verb combination appeared once across the four lists, with no repetition of sentence contexts within an experimental session.

Table 1. Example stimuli with critical word characteristics (means, with standard deviations in parentheses)

Condition	Sentence Examples	Length	Freq.	OLD20	Conc.	SSV
Plausible	<i>The overgrown lawn was <u>mowed</u>.</i>	7.73	1.97	2.12	3.49	0.13
	<i>The judge's gavel was <u>banged</u>.</i>	(1.70)	(0.77)	(0.65)	(0.70)	(0.13)
Anomalous	<i>The overgrown lawn was <u>mowing</u>.</i>	7.68	2.03	2.14	3.50	0.14
	<i>The judge's gavel was <u>pardoned</u>.</i>	(1.75)	(0.74)	(0.67)	(0.71)	(0.14)

Participants

We report data from 21 participants with a mean age of 18.9 years. Data from six additional participants were excluded: five due to excessive EEG artifact (greater than 30% of epochs rejected) and one due to an accuracy rate below 75% in the acceptability judgment task (see below).

All participants were recruited from Tufts University and the surrounding community. An initial sample of 27 participants was included. All were screened based on the following exclusion criteria: Exposure to any language other than English before the age of five years, a history of psychiatric or neurological diagnoses or injury, the use of psychoactive medication within the preceding six months. In

addition, all participants were right-handed, as assessed using the modified Edinburgh Handedness Inventory (Oldfield, 1971).

In addition to being asked their age, participants were also asked about their gender identity using the options of “Male”, “Female”, and an “Other” free-response field (12 identified as male, 15 female, and 0 non-binary). Participants were also given an option to provide information about their race/ethnicity, where they were asked whether they were Hispanic or Latino (Yes/No response), and what their race was (checking all that applied of the following options: Asian, Black or African-American, Native American or Alaska Native, Native Hawaiian or other Pacific Islander, White or Caucasian, and a free response field of Other (Hispanic/Latino: 1; Asian: 1, Black or African-American: 1, White or Caucasian: 23, Asian & White: 2).

All participants provided written informed consent and were paid for their time. All protocols were approved by the Social, Behavioral, and Educational Research Institutional Review Board at Tufts University.

Procedure and tasks

Participants sat in a dimly lit, quiet room, approximately 1.5 meters from the computer screen. Each participant carried out both the Repetition and the Correction task, with the order of these tasks counterbalanced across participants. Each task constituted two blocks of trials. Before the experiment began, participants were given instructions and example items for the first task that they would carry out, but they were not informed that their task would switch in the middle of the experiment. They then started the first task with 13 practice trials, followed by the two blocks of experimental items. After completing the first task, they completed a questionnaire about their performance (see below). They then received new instructions and example items for the second task. The new task also began with 13 practice trials, followed by two blocks of experimental items, and a post-task questionnaire. In total, the entire experiment took approximately two hours.

In both tasks, stimuli were presented using PsychoPy 1.83 software (Peirce, 2007), and were displayed on a LCD monitor using white Arial font on a black background. Stimulus lists and task order were both counterbalanced across participants. Blocks were presented in random order, and the order of trials within each block was also randomized. Before each sentence trial, a fixation cross appeared on the screen for 700ms. Sentences were presented one word at a time. Each word appeared in the middle of the screen for 450ms, with a 50ms blank screen interstimulus interval. The final word of each sentence appeared with a period. Following the sentence-final word and interstimulus interval, a fixation cross appeared for 1000ms.

In both tasks, as soon as each sentence ended, participants pressed a button to indicate whether it was plausible or anomalous. Then, 1500ms after the onset of the sentence-final word, a speech bubble appeared, cueing participants to make a verbal response. This speech bubble stayed on the screen until participants pressed the spacebar, which brought up a 700ms fixation cross before the next trial began. In the Repetition task, participants were told that, as soon as the speech cue appeared, they should repeat the prior sentence verbatim, out loud. In the Correction task, participants were told to speak plausible sentences verbatim, but to “correct” any anomalous sentences and to speak the corrected version aloud. In both tasks, to avoid speech-related EEG artifact, participants were instructed to not begin speaking until the speech bubble appeared on the screen.

The entire session was audio recorded for later transcription and analysis of behavioral data, as described below.

EEG data acquisition

EEG data were recorded using a BioSemi ActiveTwo system (biosemi.com) with 32 Ag/AgCl active electrodes in a modified 10/20 system montage. Signals were digitized at 512 Hz and a passband of DC - 104 Hz. Preprocessing was carried out using EEGLab (Delorme & Makeig, 2004) and ERPLab (Lopez-Calderon & Luck, 2014) in the MATLAB environment. After importing the data for processing,

the EEG was referenced offline to the average of the right and left mastoids. A 0.05 - 30 Hz bandpass 24 dB per octave filter was applied to the continuous data.

The EEG was then segmented into epochs spanning from 300ms before until 1400ms after the presentation of critical words. Independent Components Analysis (using the extended infomax algorithm implemented in EEGLab) was then applied to the epoched data to correct for blinks and other eye movement artifacts. Then, trials with any residual artifacts were rejected using a semi-automated procedure, with participant-specific artifact detection thresholds. After artifact rejection, there remained on average 41.9 trials per condition ($SD = 4.04$). The rejection rate was slightly greater in the plausible sentences (14.0%) than in the anomalous sentences (11.5%), as reflected by a main effect of Sentence Type ($F = 5.84, p = 0.025$), but this did not differ between the two tasks (no main effect of Task and no interaction between Task and Sentence Type, $F_s < 1.9, p_s > 0.19$). All trials were included in the analyses, regardless of accuracy in making acceptability judgments.

Behavioral data analysis

Before all behavioral analyses, we rejected trials in which participants failed to make a button response (mean: 1.8% of trials; SD: 2.0). We also rejected trials with speech onset latencies of greater than four seconds (mean: 2.5% of trials; SD: 6.0). On the remaining trials, we examined differences across the four conditions in three different behavioral measures: (a) acceptability judgments (accuracy and reaction times), (b) speech onset latencies, and (c) the Correction Distance (in characters) between each participant's verbal response and the originally presented sentence. For each of these measures, we carried out a 2×2 repeated measures ANOVA crossing Sentence Type and Task.

Speech onset latencies

Speech onset latencies were calculated by experimenters who were blind to condition. On each trial, raw speech waveforms were inspected in Audacity (www.audacityteam.org) and the difference was calculated in milliseconds between the onset of the sentence-final word and speech onset. Trials on which

participants restarted a sentence, amended a previously uttered word, or paused for more than 1000ms between words, were excluded from further analysis.

Correction Distance

To calculate Correction Distance, participants' verbal responses on each trial were transcribed by experimenters who were blind to condition. For each transcription, we used a modified Levenshtein distance method (based on Levenshtein, 1966) to compute the minimum number of character additions and subtractions needed to change the presented sentence into the participant's transcribed response. We calculated this distance using characters in order to provide a reliable metric of change, following previous work on noisy channel models of comprehension (Levy, 2008). Because changing word order can lead to an artificially large Levenshtein distance with little corresponding change in meaning (e.g. "*The boy mowed the lawn*" → "*The lawn was mowed by the boy*"), we allowed for, (a) unlimited transposition of characters, and (b) the addition or subtraction of spaces or other non-alphabetical characters without changing the total Correction Distance. For example, changing "*The overgrown lawn was mowing*" to "*The overgrown lawn was mowed*" resulted in a Correction Distance of 5, as 'ing' was removed and 'ed' was added. Changing the same sentence to "*The gardener was mowing the overgrown lawn*" resulted in a Correction Distance of 11, as the letters "*the gardener*" were added, and all other changes could be accounted for by transposition of characters or the addition of spaces.

ERP Data: Statistical analysis

To carry out statistical analyses of our ERP data, we adopted a mass univariate approach, carrying out tests at multiple time points and electrode sites (excluding ocular electrodes), and using a permutation-based cluster mass test to account for multiple comparisons (Groppe, Urbach & Kutas, 2011; Luck, 2014). As discussed in the Introduction, it was unclear whether the late positivity effect in the Correction task would localize to posterior sites (reflecting a P600) or to more frontal electrode sites. By carrying out tests at all sites, we were able to remain agnostic about the precise spatial location of effects.

Recent simulations in our lab show that, when such tests are carried out within time windows that correspond to components of interest, this approach does not sacrifice power to detect ERP effects and does not inflate Type 1 error rate (Fields & Kuperberg, 2020). We selected three time windows for analyses. Of most interest was the 600-1000ms window, which has previously been used to capture the late posterior positivity/P600 effect (e.g. see Kuperberg, Brothers, & Wlotko, 2020). In addition, we carried out analyses within a later time window (1000-1400ms) to explore the possibility of still later effects following the traditional P600 time window. Finally, we carried out an analysis between 300-600ms to capture the N400 effect where, based on previous studies using semantically-attracted stimuli (e.g. Kuperberg, Sitnikova, Caplan & Holcomb, 2003; Kim & Osterhout, 2005, Experiment 1), we hypothesized there would be no significant differences across conditions.

Within each of these time windows, at each sampling point and electrode site, we carried out 2x2 repeated measures ANOVAs that crossed Sentence Type and Task (both within-participant factors) using a prestimulus baseline of -300ms. To carry out these analyses, we used the Mass Univariate ERP Toolbox (Groppe, Urbach & Kutas, 2011) and the Factorial Mass Univariate ERP Toolbox (Fields, 2017; Fields & Kuperberg, 2020). Consecutive data points at electrode sites within 8cm of one another, which exceeded a pre-set uncorrected p-value of 0.05 or less, were considered clusters. The individual F-statistics within each cluster were summed to yield a cluster mass statistic. We then repeated this procedure 10,000 times, randomly reassigning the four condition labels at each time point and electrode site. For each randomization, we took the largest cluster mass statistic, and in this way, created a null distribution for this statistic. We then compared our observed cluster-level test statistic with this null distribution. Any clusters falling within the top 5% of the distribution were considered significant.

Any significant interactions between Sentence Type and Task were followed up with planned repeated measures ANOVAs that directly compared ERPs between the anomalous and plausible critical words in the Repetition and Correction tasks separately. In these follow-ups, we once again took a mass univariate approach, using the same time windows, electrode sites, and parameters as in the original analysis. This avoids the circularity or “double dipping” that can arise from carrying out follow-ups in

only spatial regions that show significant effects (see Kriegeskorte, Simmons, Bellgowan & Baker, 2009 for discussion). However, to interpret the source of interactions, we only considered effects at electrode sites that closely overlapped with cluster(s) that showed the original interaction effect.

Relationship between Correction Distance and ERP amplitude

Finally, we examined the relationship between evoked neural responses and a single-trial measure of Correction Distance using linear mixed effects analyses (see Results for precise models). All regression analyses were fit using the lme4 package (Bates et al., 2015) in R (R core team, 2014). Following Barr et al. (2013), these regression analyses included the maximal random effects structure.

Results

Behavioral results

Acceptability judgments: Accuracy and reaction times

Participants' overall accuracy in judging sentence acceptability was 90.5% (SD: 5.7%). There was no difference in accuracy between the anomalous and plausible sentences, and this did not differ between the two tasks (no main effects or interactions involving Sentence Type or Task on accuracy, all F s < 1.8). Participants were slower to judge acceptability in the Correction task than in the Repetition task (Task: $F(1,20) = 27.25, p < .001$), and there was a Task by Sentence Type interaction that was driven by longer times to judge the accuracy of the anomalous than the plausible sentences in the Correction task ($F(1,20) = 4.78, p < .001$), but no effect of Sentence Type in the Repetition task ($F(1,20) = 0.78, p = .44$).

Speech onset latencies

Similar to the acceptability judgments, speech onset latencies were longer in the Correction than the Repetition task ($F(1,20) = 18.35, p < .001$), and there was again an interaction between Task and Sentence Type ($F(1,20) = 29.56, p < .001$). In the Correction task, speech onset latencies were longer to

the anomalous than the plausible sentences ($F(1,20) = 41.23, p < .001$), but there were no such differences in the Repetition task ($F(1,20) = 0.47, p = .50$).

Table 2. Behavioral results with means and standard deviations

	<i>Repetition task</i>		<i>Correction Task</i>	
	Plausible	Anomalous	Plausible	Anomalous
Judgment accuracy	0.91 (0.28)	0.92 (0.27)	0.87 (0.34)	0.92 (0.27)
Judgment reaction time (ms)	1264 (549)	1290 (528)	1503 (681)	1702 (691)
Naming latency (ms)	2366 (329)	2387 (294)	2630 (549)	2969 (623)
Correction Distance (characters)	0.7 (3.0)	0.9 (2.9)	1.7 (4.5)	8.1 (6.2)

All reaction times were measured from the onset of the critical word. Correction Distances in the Repetition task (less than one character on average) reflect infrequent production errors.

Correction Distance

As expected, we also observed a significant interaction between Sentence Type and Task ($F(1,20) = 309.35, p < .001$) on our measure of Correction Distance. In the Repetition task, participants were asked to repeat each sentence verbatim, and so there were no differences in Correction Distance between plausible and anomalous sentences ($F(1,20) = 1.09, p = .31$). In contrast, in the Correction task, participants consistently corrected anomalous sentences, resulting in a clear effect Sentence Type ($F(1,20) = 541.59, p < .001$).

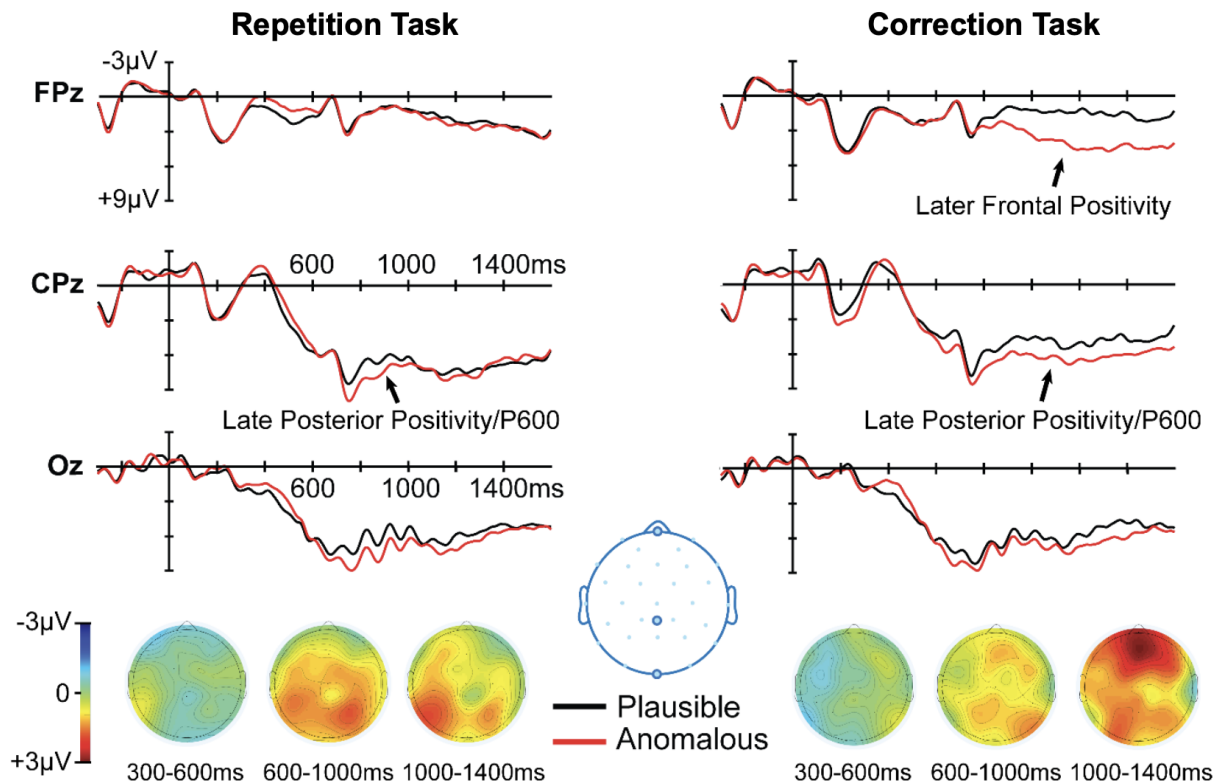


Figure 1: *Top:* Grand-average event-related potentials (ERPs) evoked by plausible (black) and anomalous words (red) in the Repetition task (left) and the Correction task (right), plotted at three midline electrode sites (FPz, CPz, Oz). Positive voltage is plotted downward. *Bottom:* Voltage maps showing the scalp topographies of the ERP effects produced by the anomalous *minus* plausible words within the 300-500ms, 600-1000ms and 1000-1400ms time windows.

ERP Results

600–1000ms

As shown in Figure 1, we observed a late posterior positivity between 600-1000ms – a P600 effect – which was larger to the anomalous than the plausible words in both tasks. This main effect of Sentence Type resulted in a significant cluster that was maximal at posterior electrodes, see Table 3. In this time window there were no significant clusters showing an interaction between Task and Sentence Type (all cluster *p*-values > .2), which suggests the magnitude of the P600 effect was similar across tasks.

Indeed, post-hoc analyses revealed significant P600 effects in both the Repetition and Correction tasks (see Table 3). Of note, the effect in the Correction task appeared to show a more frontal distribution (spatial mass peak: Fz) than the effect in the Repetition task (spatial mass peak: CP5).

1000-1400ms

The late positivity effect continued into the 1000-1400ms time window in both the Repetition and the Correction tasks (see Figure 1). At *frontal* sites, however, the effect appeared to be much more robust in the Correction than the Repetition task. Statistically, this was reflected by a cluster that showed an interaction between Sentence Type and Task, which was limited to frontal sites, see Table 3. Follow-up analyses in each task separately confirmed that, within this very late time window, positivity effects were evoked in both tasks (main effects of Sentence Type), but that it had a more frontal peak in the Correction than the Repetition task, driving the interaction between Task and Sentence Type within this frontal region.

Table 3. Significant results from the 2×2 cluster-mass permutation analyses

<i>Time Window</i>	<i>Effect</i>	<i>Cluster p-value</i>	<i>Spatial extent</i>	<i>Spatial peak</i>	<i>Temporal extent</i>	<i>Temporal peak</i>
600-1000ms						
Across both tasks	Sentence Type	< 0.001	FP1/z, AF3/4, F3/4/7/8/z, FC1/2/3/4/5/6/z, CP1/2/5/6/z, T5/6, P3/z/4, O1/z/2	P4	627-1000ms	822ms
Repetition task (post-hoc)	Sentence Type	0.004	Fz/3/8, FC1/2/5/6, Cz/3/4, CPz/1/2/5/6, T5/6, Pz/3/4, Oz/1/2	CP5	609-1000ms	758ms

Correction task (post-hoc)	Sentence Type	0.023	FPz/1/2, AF3/4, Fz/3/4/7, FC1/2/5/6, Cz/3/4, CPz/1/5, T6, Pz/3/4, O2	Fz	637-1000ms	1000ms
-------------------------------	------------------	-------	--	----	------------	--------

1000-1400ms

Across both tasks	Sentence Type x Task	0.030	FP1/2/z, AF3/4, F4/8/z, FC1/2/6, CPz	AF4	1000- 1400ms	1025ms
Repetition task (follow-up)	Sentence Type	0.013	AF3, Fz/3, FC5, C3/4, CP1/5/6, T5, Pz/3/4, Oz/1/2	T5	1000- 1400ms	1291ms
Correction task (follow-up)	Sentence Type	0.002	FPz/1/2, AF3/4, Fz/3/4/7/8, FC1/2/5/6, Cz/3/4, CPz/1/5, T6, Pz/3/4, O1/2	Fz	1000- 1400ms	1018ms

Spatial extent: All electrodes with at least one significant time point included in the cluster. *Spatial peak:* The electrode site with the largest spatial cluster mass. *Temporal extent:* All time points with at least one significant electrode in the cluster. *Temporal peak:* The time point with the largest cluster mass across electrodes.

300-600ms

As expected, we observed no significant N400 differences in this time-window: the 2 x 2 (Sentence Type x Task), and cluster mass analyses showed no significant main effects or interactions, all $ps > .1$.

Single trial analysis: Relationship between late positivity effects and Correction Distance

As described above, the mass univariate analyses revealed a significant interaction between Task and Sentence Type at frontal sites between 1000-1400ms, which was driven by a larger late frontal positivity to anomalies in the Correction than the Repetition task. If this very late frontal component is

indeed linked to successful correction, then, within this cluster, the difficulty of correction should predict the amplitude of this component across trials.

To test this hypothesis, and to examine the relationship between extracted single *trial* responses for all semantically anomalous critical words in the Correction task, these single-trial ERPs were then averaged across electrodes and time-points within two regions of interest: a posterior P600 region (600-1000ms; Pz/3/4, Oz/1/2) and a very late frontal positivity region (1000-1400ms; FPz/1/2, AF3/4, Fz/3/4).

We found that, during the Correction task, Correction Distance failed to predict the magnitude of the posterior P600 ($b = 0.22$, $t = 0.52$, $p = .61$). Instead, Correction Distance predicted the amplitude of the later frontal positivity ($b = 1.81$, $t = 3.23$, $p = .005$), with a larger late frontal positivity on anomalous trials that required more extensive edits (see Figure 2).

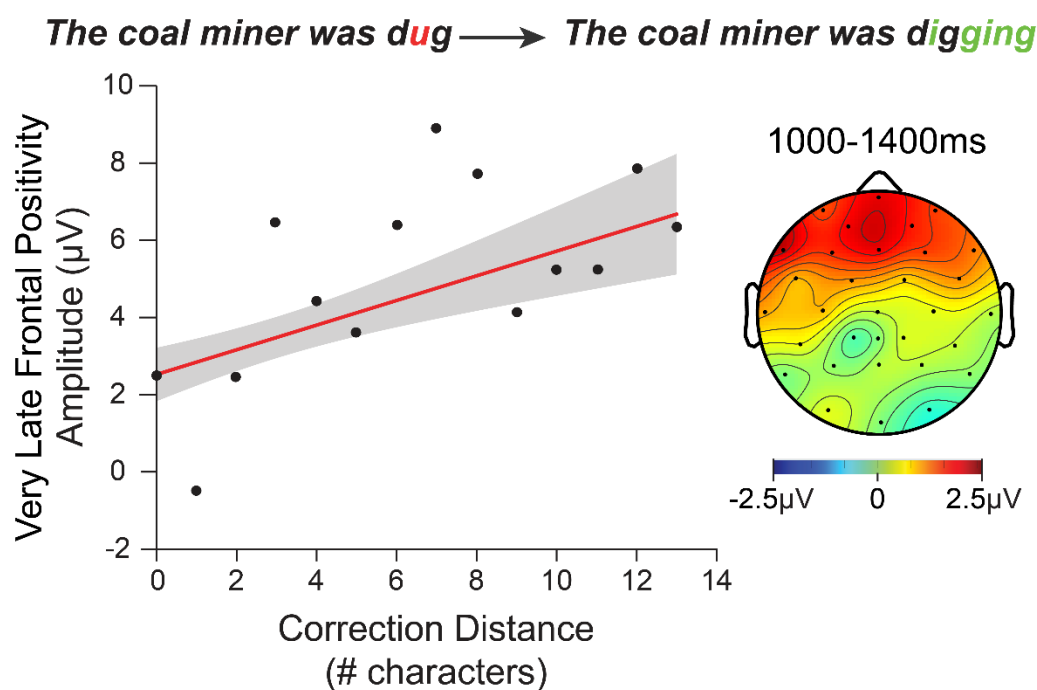


Figure 2 *Top*: An example of an anomalous sentence and its correction, with a Correction Distance of 6 characters, including both deletions (red) and insertions (green). *Bottom Left*: A scatter plot showing the relationship between the amplitude of the very late frontal positivity (averaged across FPz/2, Fz/4, AF4; 1000-1400ms) and the Correction Distance between anomalous sentences and participants' verbal

responses. *Bottom Right*: A topographic voltage map, showing the frontal scalp distribution of the Correction Distance effect.

Discussion

In this study, we attempted to disentangle the neural mechanisms of error-based processing. We recorded ERPs at the scalp surface and presented participants with plausible sentences and sentences containing semantically attracted anomalies. On each trial, participants were instructed to indicate whether or not they detected an error, and then to either: (a) repeat the sentence without any modification (Repetition task), or (b) correct the sentence so that it made sense (Correction task). In both tasks, the anomalous words elicited a larger P600 than the plausible words between 600-1000ms. Importantly, however, the magnitude of this posteriorly distributed effect was not sensitive to error correction: we observed similar P600 effects in both the Repetition and the Correction tasks. Moreover, in the Correction task, the amplitude of this posterior P600 was not sensitive to the degree of correction across trials (i.e. the number of characters changed between the anomalies and participants' corrections). Instead, explicit correction influenced a still later frontally distributed positivity (1000-1400ms). Moreover, the amplitude of this later frontal positivity was predicted by trial-by-trial Correction Distance, with anomalies that require more extensive correction producing larger late frontal positivities.

The P600 does not index error correction

Before concluding that the posterior P600 response is not directly linked to error correction of anomalous inputs, it is important to consider an alternative interpretation: Is it possible that, even though participants were instructed to repeat each sentence verbatim in the Repetition task, they simultaneously engaged in covert corrections? We believe that this is unlikely for two reasons. First, it would be very difficult to internally correct the anomalous sentences while also maintaining their original form in working memory for later repetition. If participants had engaged in this type of dual task, it would have led to substantial interference, resulting in longer speech onset latencies to the anomalies (*versus* plausible sentences) in the Repetition task. However, we found no evidence of such interference.

Our finding that the P600 is not linked to error correction is problematic for “noisy channel” theories of sentence comprehension that have linked the P600 to “error correction” (Gibson, Bergen & Piantadosi, 2013; Ryskin et al., 2021). These theories posit that, in addition to maintaining high probability beliefs about observed linguistic inputs, comprehenders also assign a small amount of probability mass to possible near-neighbors of these inputs (Levy, 2008; Levy, 2011; Levy, Bicknell, Slattery & Rayner, 2009). For example, when reading the sentence “*The overgrown lawn was mowing*”, comprehenders might retain a small amount of belief for alternative surface forms (e.g. “*The overgrown lawn was mowed*”). When the linguistic anomaly is detected, comprehenders engage in correction/belief updating if this reinterpretation is consistent with one of these alternative near-neighbors. The P600 response is assumed to directly reflect the process of shifting beliefs towards alternative near-neighbors to come to an alternative plausible interpretation (see Gibson, Bergen & Piantadosi, 2013; Ryskin et al., 2021). However, if the P600 was linked to error correction, then we would have seen a selective enhancement of the P600 effect in the Correction task.

The P600 is unlikely to purely reflect an explicit task-related categorical judgment

Another possibility to consider is that, in both the Repetition and Correction tasks, the P600 reflected *only* the detection/categorization of anomalies, as required by the task. However, previous studies have shown that simply detecting and making an overt judgment about semantic anomalies is not sufficient to trigger a P600 effect. For example, in a recent study by Brothers, Wlotko, Warnke & Kuperberg, 2020 (Experiment 1), we found that *no* P600 effect was produced by semantic anomalies in short sentence contexts (*James locked the laptop/gardener**), even when participants correctly judged them to be anomalous (for similar findings, see Gunter, Friederici & Hahne, 1999; Osterhout & Nicol, 1999). The same anomalies, however, produced a robust P600 when they were embedded in longer discourse contexts (Brothers, Wlotko, Warnke & Kuperberg, 2020, Experiment 2). This was taken as evidence that, in order to evoke a P600 effect, an anomalous input must disrupt the broader goal of comprehension, thereby committing comprehenders to reprocessing the linguistic input.

The P600 may be linked to prolonged decision making during reprocessing of the anomalous input

We suggest that current set of findings can be understood within a decision making account of the P600 component that builds from monitoring account of the P600 (van de Meerendonk, Kolk, Chwilla, & Vissers, 2009; see also Brothers, Zeitlin, Choi Perrachione, Choi & Kuperberg, 2021). Briefly, we suggest that the P600 indexes a prolonged decision-making process – an extended accumulation of “error evidence” that is collected during *reprocessing* of an anomalous input. This reprocessing serves the crucial purpose of confirming that the error stemmed from an external source (an external error within the linguistic input), rather than from an internal processing error that can arise from the comprehender’s earlier misanalysis (e.g. Altmann, 1998; Osterhout & Holcomb, 1992; 1993; Hagoort, Brown, & Groothusen, 1993), or misperceptions of the input (Levy, Bicknell, Slattery & Rayner, 2009; e.g. one word being misrecognized as another (Kaufman & Obler, 1995; Gregg & Inhoff, 2016). The P600 may track the amount of “error evidence” accumulated during reprocessing. Reprocessing and decision making terminate only when the comprehender is confident that the error stems from an external (rather than internal) source.

This account links the P600 to a large body of evidence that the brain engages sequential evidence accumulation algorithms to support decision making (cf. Ratcliff, 1978; Forstmann, Ratcliff, & Wagenmakers, 2016). In particular, it links to work in non-linguistic domains that has linked other prolonged posterior late positivities with the accumulation of evidence to increase confidence about previous choices (Murphy, Robertson, Harty, & O’Connell, 2015; Boldt & Yeung, 2015; Desender, Murphy, Boldt, Verguts, & Yeung, 2019; Desender, Ridderinkhof, & Murphy, 2021). These non-linguistic posterior positivities are closely related to the P3b ERP component, and so these links are of particular relevance to previous discussions relating the P600 to this family of components (see Coulson, King & Kutas, 1998; Osterhout, Kim & Kuperberg, 2012; Sassenhagen & Fiebach, 2019; Sassenhagen, Schlesewsky, and Bornkessel-Schlesewsky, 2014).

Reprocessing and prolonged decision making are only initiated if an anomaly disrupts the build-up of a high-level interpretation

A key claim of this theory is that reprocessing and prolonged decision making is only initiated if the linguistic anomaly is perceived as *disrupting* the ongoing construction of a high-level situation model (see Kuperberg, Brothers, & Wlotko, 2020; Brothers, Wlotko, Warnke, & Kuperberg, 2020). If no situation model has been established before an anomaly is encountered, then no P600 will not be produced (e.g. Brothers, Wlotko, Warnke, & Kuperberg, 2020). Many factors can influence whether or not comprehenders will establish a situation model during comprehension, including their internal goals, external task requirements, and the presence of additional linguistic or non-linguistic context (see Kuperberg, 2007 for a review of factors influencing the semantic P600). In the present study, where the sentences were very short, we suggest that the key triggering factor was the strong semantic attraction between the individual words. These semantically-attracted words provided highly reliable cues for the retrieval of a canonical schema stored in long-term memory (e.g. “dark-caves-exploring” or “judge-gavel-pardon”, see Kuperberg, 2016 for discussion), allowing comprehenders to establish a situation model.²

A later frontally distributed positivity may reflect the successful re-establishment of plausibility

Although the posterior P600 effect to the semantic anomalies did not differ across the two tasks, we found evidence that a still-later frontally distributed positivity was sensitive to correction. Semantic anomalies produced an enhanced frontal positivity only in the Correction Task, and the amplitude of this component increased with Correction Distance on each trial.

² This “cue reliability” account of semantic attraction can be understood within a probabilistic framework (see Kuperberg, 2016). This framework differs from earlier “stream-based” accounts that attributed the effects of semantic attraction to the engagement of a semantic combinatorial stream (Kim & Osterhout, 2005), a plausibility heuristic (van Herten, Kolk, & Chwilla, 2005), or to the tracking of semantic associations and other semantic relationship within memory (see Kuperberg, 2007).

As noted in the Introduction, late frontal positivities are typically seen in plausible sentences when an unexpected word induces leads to a large update of the current situation model (Kuperberg, Brothers, & Wlotko, 2020; Brothers, Wlotko, Warnke, & Kuperberg, 2020).

The frontal positivity in the current study was observed in a later time-window than previously reported (1000-1400ms vs. 600-1000ms), and, unlike in these previous studies, it was triggered by critical words that were anomalous in context rather than simply unexpected. However, we suggest that this frontal effect may have similarly reflected a successful large shift in the situation model. In the present case, this shift would have *resulted* from correcting the anomalous input. Consistent with this account, the greater the shift between the anomalous sentence and the generated correction (as indexed by our Correction Distance measure), the greater the amplitude of this late frontal positivity.³

Implications for natural language comprehension

One set of open questions concerns the role of reprocessing and error correction during task-free reading comprehension. Here, we have argued that the reprocessing triggered by a linguistic error does not necessarily require a commitment to correct this error. This raises the question of how often comprehenders *spontaneously* fix errors when they are not explicitly instructed to do so? We suggest that this will depend on the *outcomes* of reprocessing and the nature and source of the error that triggered it. In cases when the error stems from a misanalysis or misperception of the prior input, a second chance to process the input should yield the correct interpretation. In cases when the comprehender encounters an overt production error, reprocessing should reduce uncertainty about where in the sentence, and at what level of linguistic representation it originated. The comprehender could then use this information to decide whether to commit to a specific correction. In other cases, however, the source of an error may not

³ An alternative interpretation is that this very late frontal response instead reflects a distinct process of generating and selecting an appropriate correction, or perhaps the downstream process of updating a previously activated speech plan. Future ERP studies investigating covert correction mechanisms during online comprehension will be important for disentangling these alternative accounts.

be so obvious, and so, after engaging in multiple rounds of reprocessing, readers may opt for a “good enough” interpretation of the sentence (Christianson et al., 2001; Ferreira, 2003; Ferreira et al., 2002). This would involve re-engaging a default forward-looking processing mode without ever resolving the initial error.

If there are indeed two distinct ERP responses associated with reprocessing (late posterior positivity/P600) and the successful re-establishment of coherence (late frontal positivity), our current findings provide a novel method for tracking successful error correction during task-free reading comprehension. For example, if comprehenders only engage in spontaneous corrections in response to easy-to-correction anomalies, then these types of anomalies should produce both a posterior and a frontal positivity, while anomalous inputs with larger Correction Distances may only produce a posterior positivity/P600 effect. When testing these predictions, it will be important to examine the distribution of late positivities in both early (600-1000ms) and late time-windows (1000-1400ms), and to determine whether these neural signatures vary across individuals with different reading abilities.

Chapter 3: Contributions of Naturalistic Stimuli on the P600

In the previous chapter, participants read short, four-to-five-word sentences (“The lawn was mowing”) while performing tasks that required different levels of involvement with the stimuli. We showed that asking participants to correct errors in sentences, compared to repeating the errors verbatim, resulted in increased late positive EEG activity related to reprocessing (and associated decision-making). We presented a case that these late positivities, including the P600, reflect activity important to general language processing.

Here, we focus more specifically on the question of whether artificially imposed task demands are necessary preconditions for observing P600s to semantic and syntactic errors by presenting single sentence stimuli alone with passive comprehension questions, with overt acceptability judgments, and embedded in multi-sentence scenarios and a task that encouraged deep comprehension comparable to natural reading.

This chapter builds upon work by Suiping Wang, who previously completed data collection for experiments 1a and 1b, which used single-sentence stimuli. This chapter extends this work by adding a deep comprehension environment (longer multi-sentence scenarios and a set of tasks designed to require deep comprehension).

Gina Kuperberg and Trevor Brothers assisted extensively with study design and writing. I additionally thank Cassidy Prejean, Elizabeth Kenneally, and Maya Lazarus, who assisted with data collection and analysis.

General Introduction

The P600 is a late positive event related potential (ERP) observed at the scalp surface between 500-1000ms and maximal over posterior electrodes sites. It is frequently triggered by linguistic anomalies during language comprehension, and its amplitude is enhanced when comprehenders are explicitly instructed to detect these anomalies. What does this tell us about language comprehension? Does the

sensitivity of the P600 to task demands present an artificial confound imposed by the experimenter? Or does it yield insights into the neurocognitive mechanisms engaged during error processing during deep language comprehension? In the present study, we address this question by systematically examining the modulation of the P600 produced by the same set of semantic and syntactic anomalies under three different experimental and task environments

The P600 was initially characterized as a neural index of syntactic processing (Hagoort, Brown, & Groothusen, 1993; Osterhout & Holcomb, 1992; 1993). It is now clear, however, that a robust P600 can be produced by multiple different types of linguistic anomalies during language comprehension. For example, a larger P600 is evoked by misspelled words (orthographic anomalies) than by correctly-spelled words (e.g. Munte, Heinze, Matzke, Wieringa & Johannes, 1998; Vissers, Chwilla & Kolk, 2006; Ito & Nieuwland, 2016). In addition a “semantic P600 effect” is produced by different types of semantic anomalies, ranging from local selection restriction violations within sentences (e.g. Paczynski & Kuperberg, 2011, 2012; Kuperberg, Brothers, & Wlotko, 2020), to words that violate presuppositions in discourse contexts (e.g. Xiang & Kuperberg, 2015; Jouravlev et al., 2016; Shetreet, Alexander, Romoli, Chierchia & Kuperberg, 2019). In contrast, a P600 effect is less likely to be produced by words that are unlikely but not anomalous in relation to their prior context. These semantically unexpected words produce only an earlier N400 effect between 300-500ms, which is thought to reflect the retrieval or access to the semantic features of individual words (Kutas & Federmeier, 2011; Kuperberg 2016).⁴

In many of the studies that have reported a P600 to linguistic anomalies, participants were asked to make explicit judgments about the acceptability or coherence of the linguistic materials. It has been pointed out that, because this task is directly relevant to the manipulation of interest (the presence or absence of an anomaly), it should enhance the amplitude of any P600 produced by the linguistic violation.

⁴ An N400 effect often precedes the P600 effect produced by semantic anomalies. However, a P600 effect can also be seen without an earlier N400 effect (see Kuperberg, 2007). If the P600 begins early within the 300-500ms time window, thereby obscuring the appearance of the opposite-polarity N400 at the scalp surface due to spatiotemporal component overlap (see Kuperberg, Kreher, Sitnikova, Caplan & Holcomb, 2007; Brouwer & Crocker, 2017) and/or if the incoming word is semantically associated with its prior context (see Kuperberg, 2016, and see Shetreet, Alexander, Romoli, Chierchia & Kuperberg, 2019, Supplementary Materials for detailed discussion).

This intuition has been confirmed by a few studies that have directly examined the effect of task on the magnitude of the P600 effect. When participants make active acceptability or coherence judgments, the P600 produced by both syntactic (Osterhout, Allen, McLaughlin, & Inoue, 2002; Royle, Drury, & Steinhauer, 2013) and semantic violations (Kolk et al., 2003; Xiang & Kuperberg, 2005 ; Geyer et al., 2006; Royle, Drury, & Steinhauer, 2013) is larger than under more passive task demands (e.g. answering intermittent comprehension questions).

These observations have led to much debate about the role of task on the P600. Two different perspectives have emerged. The first is that the explicit requirement to track the presence/absence of linguistic anomalies superimposes an unnecessary and artificial decision-making requirement on language comprehension. This decision making is not believed to be directly linked to the preparation of the motor response itself because participants are usually asked to delay this response until well after the appearance of the anomaly. However, as pointed out by Schacht et al. (2014), even at earlier points in the sentence, “violations would provide task-relevant information allowing for a ‘no’ decision.”

The implication of these arguments is that the sensitivity of the P600 to task requirements presents an inherent “confound”. As such, it has been argued that finding a P600 when participants carry out a judgment task tells us little about the architecture of the language processing system, and is unlikely to generalize to “naturalistic” task-free comprehension (e.g. Vega-Mendoza, Pickering, & Nieuwland, 2021). The solution that is usually offered is that, to better simulate natural comprehension, participants should be asked to simply read sentences, while answering intermittent questions to check their understanding.

The second perspective on the relationship between the P600 and task is that instructions to explicitly judge anomalies encourage participants to engage decision making processes that play a crucial role in supporting deep language comprehension. Linguistic errors are common in everyday communication (Nakatani & Hirschberg, 1994; Nooteboom, 1980). Overt production errors appear frequently in natural speech, as well as in texts and emails. In addition, “internal” comprehension errors

can arise from temporary misparses of ambiguous or complex sentences (e.g. Altmann, 1998), or from misperceptions such as “slips of the eye” (Gregg & Inhoff, 2016; Kaufman & Obler, 1995).

As discussed in Chapter 2, detecting such errors, and distinguishing whether they stem from an external (versus internal) source, requires us to reprocess our input and engage in an extended decision-making process, which plays an important role in paving the way for downstream processes, such as adaptation, and learning. Proponents of this second view have pointed out that requiring participants to make explicit error judgments during experimental studies ensures that we actually capture the neurocognitive mechanisms of interest. They point out that, in a lab environment, asking people to engage in more passive tasks with only occasional comprehension checks may be problematic because it may discourage deep comprehension. Instead, participants may focus on the surface structure of the text (cf. Kintsch & van Dijk, 1978; Graesser, Singer & Trabasso, 1994) and ignore the presence of anomalies altogether. Indeed, there is evidence that if comprehenders fail to consciously detect a semantic anomaly, they do not produce a P600 effect at all (Sanford et al., 2011).

One difference between these two perspectives rests on a subtle but important distinction between “task relevance” and “goal relevance”. During deep comprehension, the overarching *goal* is to infer the message or situation model being communicated. The imposition of an additional task, such as asking participants to press a button depending on whether or not an anomaly has been detected, is often closely linked to this goal. However, goal-relevance and task-relevance can be dissociated: It is possible to successfully make an anomaly judgment without being engaged in deep comprehension, and it is possible to engage in deep comprehension without pressing a button to indicate one’s decision. The question of interest, therefore, is whether the P600 is primarily relevant for the goal of deep comprehension, or whether it is *only* modulated by explicit task requirements.

There is some existing evidence that speaks to this question. Several studies have reported that a P600 can be evoked by semantic and syntactic anomalies, even when participants are not explicitly required to make judgments, so long as they are engaged in deep comprehension. This is most likely to occur when the anomalies appear in multi-clause sentence or discourse contexts (e.g. Nieuwland & Van

Berkum, 2005; van de Meerendonk, Kolk, Vissers & Chwilla, 2010; DeLong, Quante & Kutas, 2014; Szewczyk & Schriefers, 2011; Münte, Heinze, Matzke, Wieringa & Johannes, 1998; Quante, Bolte & Zwitserlood, 2018 ; Pijnaker, et al., 2010; Hanulíková et al., 2012, native accent conditions; Dien, Michelson & Franklin, 2010; Jouravlev et al., 2016).

Conversely, there are times when no P600 is evoked by linguistic anomalies, even when participants are asked to make explicit anomaly judgments. For example, syntactic P600s are typically absent or greatly reduced to syntactic anomalies in jabberwocky sentences, even though participants correctly judge them to be anomalies (Münte, Matzke & Johannes, 1997; Ericsson, Olofsson, Nordin, Rudolfsson & Sandstrom, 2008; Canseco-Gonzalez, 2000; Canseco-Gonzalez et al., 1997; Yamada & Neville, 2006; though see Hahne & Jescheniak, 2001). In these sentences, all content words are replaced by pseudowords and so the comprehender knows that no message is being conveyed. In addition, no P600 effect is evoked by correctly judged semantic anomalies that appear in minimal contexts (e.g. Brothers, Wlotko, Warnke, & Kuperberg, 2020, Experiment 1; Gunter, Friederici & Hahn, 1999; Osterhout & Nicol, 1999), where it is again unclear to the comprehender whether a message is being communicated.⁵

The experiments discussed above provide some evidence against a task-centric view of the P600 response. They suggest that anomaly judgment tasks are neither necessary nor sufficient to produce a P600 effect to linguistic anomalies. Nonetheless, concerns about the role of task in interpreting the P600 persist in the literature. Only a few studies have systematically investigated its sensitivity to varying task demands and, as noted above, these studies have mainly contrasted two extremes in which participants are either asked to make judgments about the presence/absence of an anomaly in single sentences, or asked to simply read these sentences and respond to occasional comprehension questions (e.g. Xiang &

⁵ Interestingly, however, even in minimal contexts, semantic anomalies can produce a P600 if they are semantically “attracted” to a preceding word (Kim & Osterhout, 2005; Chow & Phillips, 2013; Kim, Oines, & Miyake, 2018). In these cases, it may be clearer to the comprehender that the communicator had an event in mind that was being communicated (e.g. *The hearty meals were devouring...*’ from Kim & Osterhout, 2005, Experiment 1; Kuperberg, Sitnikova, Caplan & Holcomb, 2003; Kolk, Chwilla, van Herten & Oor, 2003).

Kuperberg, 2015; Kolk et al., 2003; Osterhout, Allen, McLaughlin, & Inoue, 2002; Royle, Drury, & Steinhauer, 2013; Geyer et al., 2006).

In the present study, we therefore carried out a series of three experiments in which participants read sets of sentences with semantic and syntactic anomalies that previous studies have shown to produce robust semantic and syntactic P600 effects (Kuperberg et al., 2003; 2006; 2007). In all three experiments, ERPs were recorded during self-paced reading, which allowed participants to directly control the rate of information processing. The first two experiments sought to replicate previous work that has examined the P600 responses evoked by semantic and syntactic anomalies appearing in single sentences under the two task extremes. In Experiment 1a, participants passively read sentences and answered intermittent comprehension questions. In Experiment 1b, a different group of participants performed a classic anomaly judgment task in which they judged whether or not each sentence made sense. Based on previous findings, we expected that the P600 evoked by both semantic and syntactic anomalies would be larger in Experiment 1b than Experiment 1a.

We then carried out a third experiment that was designed to more closely emulate the process of natural but deep reading comprehension. In this experiment, the same set of semantic and syntactic anomalies were embedded in coherent, multi-sentence contexts. No explicit decisions about the presence or absence of linguistic anomalies were required, but, to encourage deep reading comprehension, participants responded to an engaging comprehension question that appeared after every trial. By providing richer contexts and encouraging participants to attend carefully to the entire discourse scenario, we hoped to make stronger inferences about the relevance of the P600 to anomaly processing in more naturalistic language contexts. If, under these experimental conditions, the P600 effect is larger than in the passive comprehension task in Experiment 1a, this would support the claim that the P600 effect reflects brain activity that is relevant to the goal of deep language comprehension.

Study 1 (Experiment 1a and 1b)

Introduction

The aim of this study was to confirm the findings of previous studies examining the effect of task on the P600. We directly contrasted the P600 effects to semantic and syntactic anomalies in single sentences while participants were engaged in passive comprehension (Experiment 1a) or in active anomaly judgments (Experiment 1b). In both experiments, participants read four types of sentences, and we measured ERPs on critical verbs (see Kuperberg et al., 2006). These verbs were 1) *congruous*, 2) *real-world incongruous* (semantically unrelated to the prior context and inconsistent with real-world knowledge, but not anomalous), 3) *semantically anomalous* (violating the animacy-based selectional restriction constraints of the subject noun, but nonetheless semantically related to prior words in the context), or 4) *syntactically anomalous*, with an incorrect morphosyntactic marker (see Table 5).

In Experiment 1a, participants answered intermittent comprehension questions. These comprehension questions were relatively undemanding, and they never referred directly to the critical verbs, even when following anomalous sentences. In Experiment 1b, participants were asked to make explicit acceptability judgments at the end of each sentence.

We examined the P600 effect to both the semantic and syntactic anomalies in both tasks. Based on our previous studies using these materials (Kuperberg et al., 2003; 2006; 2007), we expected that, when participants carried out the explicit judgment task (Experiment 1b), we would see robust P600 effects to both the semantic anomalies and the syntactic anomalies, relative to the congruous critical words. Also, on the basis of previous studies that have contrasted these tasks with different stimulus materials (Kolk, Chwilla, van Herten, & Oor, 2003; Xiang & Kuperberg, 2015), we expected that these P600 effects would be larger in Experiment 1b than in Experiment 1a. We were particularly interested in whether, in the passive comprehension task (Experiment 1a), we would see a simple *attenuation* of the P600 response, or whether this ERP effect would be completely absent without the explicit requirement to make a binary categorization judgment about the stimulus materials (cf. Schacht et al., 2014).

In addition to examining the P600 produced by the semantic and syntactic anomalies, we also examined the effect of task on the magnitude of the earlier N400 effect produced by the real-world incongruities. The N400 is thought to reflect the ease of retrieving semantically unexpected information (Kutas & Hillyard, 1984; Kutas & Federmeier, 2011). This would help determine whether judgment tasks have a selective influence on the processing of linguistic errors (indexed by the P600), or if task demands have a more general effect on the retrieval of lexico-semantic information (indexed by the N400). Note that, based on our previous studies using these materials, we did not expect to see an earlier N400 effect on the semantically anomalous verbs, both because they were semantically related to the prior context, resulting in increased semantic facilitation, and because the P600 is likely to have begun in the N400 time window, resulting in component overlap (see Kuperberg et al., 2007 for discussion).

Methods

Participants

For Experiment 1a, we report data from 24 native English speakers (7 females, average age: 20.2 years). Data from five additional participants were excluded due to excessive ERP artifact (see below). For Experiment 1b, we report data from 16 different participants (8 female, average age: 19.4 years). Twenty-one originally participated, but five were subsequently excluded (two due to poor accuracy in the acceptability judgment task, and three because of excessive ERP artifact, see below).

All participants were native speakers of English with no significant exposure to any language other than English before the age of five. They were recruited from Tufts University and the surrounding community. All were right-handed, as assessed using the modified Edinburgh Handedness Inventory (Oldfield, 1971; White & Ashton, 1976), and all had normal or corrected-to-normal vision. None were taking any medication, and participants were screened to exclude the presence of psychiatric disorders and neurological disorders/injury. They were paid for their time and written consent was obtained before participation, following the guidelines of the Tufts Human Subjects Research Committee.

Design and Materials

The materials have been previously described in detail by Kuperberg et al. (2006). In total, there were 180 sentences, with 45 items presented in each of the four conditions. Briefly, all sentences contained a short, introductory context, an animate subject and an Agent-Patient verb. The congruous sentences continued with a coherent critical word (e.g. *Every morning at breakfast the boys would only eat...*). The real-world incongruous sentences were constructed by replacing the critical verb of each congruous sentence with a verb from a different sentence (e.g. *Every morning at breakfast the boys would only plant...*). The syntactically anomalous sentences were created by either violating subject-verb agreement or by using a finite verb in place of an infinitival verb (...*the boys would only eats toast and jam.*). The semantically anomalous sentences were constructed by replacing the animate noun phrase with an inanimate one, which was related to the prior context (...*the eggs would only eat toast and jam.*). Sentences continued with an additional two to three words following the critical verb. On average, sentences were 11.2 words long, with 7.9 words prior to the critical verb and 2.3 words following the critical verb.

Sentences were counterbalanced across four lists, ensuring that each participant read each sentence context and critical word verb only once, but that, across participants, the same critical verb appeared in all four conditions. In order to increase the proportion of congruous sentences within each list, we added 23 or 24 filler congruous sentences to each list. Thus, approximately two-thirds of sentences in each list contained some type of anomaly or incongruity, and one-third were congruous. Within each list, the order of conditions was randomized.

Tasks and Procedure

Participants were randomly assigned to one of the four counterbalanced lists. They sat in a comfortable chair in a dimly lit, quiet room, approximately 100cm from the computer screen. Before the main experiment, they carried out a practice session (eight trials). At the beginning of each trial, the word “READY” appeared in the center of the screen. Participants then pressed a button on a gamepad to begin, and they continued to press the same button in order to self-pace through the trial (response hand was counterbalanced across participants). Each trial began with a centered fixation cross, followed

by each word of the sentence. Participants were discouraged from blinking while they were reading the sentences, but encouraged to blink between trials while “READY” was on the screen.

In Experiment 1a, participants were told that they should respond to a question that would appear following some of the sentences by pressing either a “yes” or a “no” button on a response box. In all trials, the sentence-final word was followed by a blank screen, and, following 41% of sentences — 15 of the 45 experimental sentences in each condition and all of the congruous filler sentences — this was followed by the comprehension question. There were equal numbers of questions in which the correct answer was “yes” as “no”.

In Experiment 1b, participants were told that, following each sentence, they would see a “?”, at which point they should indicate whether or not the sentence was acceptable or unacceptable by pressing either the “yes” or the “no” button on a response box. They were told that they should judge the sentence to be unacceptable if it contained grammatical errors or if the content of the sentence seemed odd in any way. The “?” appeared 700ms after the disappearance of the final word of each sentence.

EEG Data Acquisition and Preprocessing

In both experiments, EEG was recorded from the scalp using 29 tin electrodes, secured on the scalp surface by an elastic cap (Electro-Cap International, Inc., Eaton, OH) in a modified 10/20 system montage. Electrodes were also attached below the left eye and at the outer canthus of the right eye in order to monitor vertical and horizontal eye movements, as well as on the left and right mastoids. Impedances were kept below 10 k Ω for the eyes and below 5 k Ω at other sites. The EEG signal was amplified by an Isolated Bioelectric Amplifier System Model HandW-32/BA (SA Instrumentation Co., San Diego, CA) with a bandpass of 0.01 to 40 Hz. It was continuously sampled at 200 Hz by an analog-to-digital converter and referenced online to the left mastoid. After recording, EEG data were time-locked to the critical word using a -100ms pre-stimulus baseline and averaged across trials in each of the four conditions. All trials contaminated with either eye-movement artifact or amplifier blocking were identified, and participants with artifact on more than 40% of trials were excluded. This resulted in the exclusion of five participants in Experiment 1a, and three participants in Experiment 1b. Of the remaining

participants in Experiment 1a, there were, on average, 39 trials per condition (range: 26-45) after artifact rejection, with no significant differences in artifact rejection rates across conditions ($F(3,92) = 0.20, p = .90$). In Experiment 1b, there remained, on average, 41 trials per condition (range: 36-44), with no significant differences in rejection rates across conditions ($F(3,60) = 1.24, p = .34$).

ERP Statistical Analysis Strategy

Based on previous studies (Shetreet et al., 2019; Kuperberg, Brothers, & Wlotko, 2020; Brothers et al., 2020), we operationalized the P600 as the average voltage across all time points between 600-1000ms, averaged across eight posterior electrode sites (CP1/2, Pz/3/4, Oz/1/2). Within each experiment, we asked (a) whether the semantic P600 and syntactic P600 effects were significant by contrasting each of these anomalies with the congruous words, and (b) whether the amplitude of the P600 differed between the syntactic and animacy anomalies. We then examined the influence of Experimental Task on the semantic and syntactic P600 effects by calculating the magnitude of each effect in each participant (anomalous *minus* congruous) and carrying out a 2 x 2 ANOVA that crossed Experimental Task (between-participant: Exp 1b vs. Exp 1a) and P600 Anomaly Effect (within-participant: syntactic P600 effect vs. semantic P600 effect).

We operationalized the N400 as the average voltage between 300-500ms across six central electrode sites (Cz/3/4, CP1/2). We first verified that a significant N400 effect was produced by the real-world incongruities (vs. congruous critical words) in each experiment, and we then carried out a between-subjects t-test to determine whether the magnitude of this effect differed between the two Experimental tasks.

All analyses were carried out using all artifact-free trials, regardless of accuracy. A Greenhouse–Geisser correction was applied to all analyses with more than one degree of freedom in the numerator (Greenhouse & Geisser, 1959). In these cases, the original degrees of freedom with the corrected p-value are reported. We also constructed linearly interpolated voltage maps to show the scalp distribution of ERP effects within the time windows of interest, using the EEGLab plugin in MATLAB.

Results

Behavioral Results

In Experiment 1a, participants were generally accurate in answering the yes/no comprehension questions (93% correct, with no significant differences across conditions, $F(3, 69) = 1.37, p = .26$). In Experiment 1b, participants' judgment accuracy was also high (96% on average). An ANOVA revealed significant differences across the four conditions ($F(3, 69) = 8.92, p < .001$) due to greater accuracy in classifying the semantically anomalous (99%) and the syntactically anomalous (99%) sentences as unacceptable/odd than in classifying the real-world incongruous sentences as unacceptable/odd (94%) or the congruous (93%) sentences as acceptable.

ERP Results

P600 (600-1000ms)

In Experiment 1a, both the semantic and the syntactic anomalies produced significant P600 effects (semantic anomalies vs. congruous: $t(23) = -3.57, p < .001$; syntactic anomalies vs. congruous: $t(23) = -4.63, p < .001$), and the amplitude of the P600 did not differ between the semantic and syntactic anomalies ($t(23) = 1.05, p = .31$).

Similarly, in Experiment 1b when participants carried out explicit judgments, large P600 effects were produced by both the semantic anomalies ($t(15) = -9.70, p < .001$) and the syntactic anomalies ($t(15) = -13.42, p < .001$). In this Experiment, however, the syntactic P600 was larger than the semantic P600 ($t(15) = 8.60, p < .001$).

As shown in Figure 1, the magnitude of both the semantic and syntactic P600 effects appeared to be larger in Experiment 1b, when participants made explicit acceptability judgments, than in Experiment 1a when participants answered intermittent comprehension questions, and this effect of task appeared to be greater for the syntactic than the semantic P600 effect. This was confirmed by a 2 x 2 ANOVA that crossed Experimental Task (Exp 1a vs. 1b) with P600 Anomaly Effect (semantic P600 effect vs. syntactic P600 effect), which showed both a significant main effect of Experimental Task ($F(1,76) = 139, p$

$< .001$), and a significant interaction between Experimental Task and P600 Anomaly Effect ($F(1,76) = 18.5, p < .001$).

N400 (300-500ms)

As expected, the real-world incongruous critical words elicited larger N400 responses than congruous critical words in both Experiment 1a ($t(23) = 3.65, p = .001$) and Experiment 1b ($t(15) = 5.14, p < .001$). In contrast to the P600 effects, however, the magnitude of this effect did not differ between the two tasks ($t(38) = 1.29, p = .20$).

Discussion

In these two experiments, participants read sentences at their own pace while answering intermittent comprehension questions (Exp. 1a) or judging the acceptability of each sentence (Exp. 1b). Qualitatively, the pattern of P600 responses were quite similar across the two experiments. Even when participants were reading solely for comprehension, semantic and syntactic anomalies elicited significant P600 effects relative to congruous sentences. This is consistent with previous findings that have reported P600 effects in the absence of explicit task requirements (e.g. Nieuwland & Van Berkum, 2005; van de Meerendonk, Kolk, Vissers & Chwilla, 2010; DeLong, Quante & Kutas, 2014; Szewczyk & Schriefers, 2011; Münte, Heinze, Matzke, Wieringa & Johannes, 1998; Quante, Bolte & Zwitterlood, 2018 ; Pijnaker, et al., 2010; Hanulíková et al., 2012, native accent conditions; Dien, Michelson & Franklin, 2010; Jouravlev et al., 2016), and suggests that an overt judgment is not a necessary precondition for eliciting a P600 effect. Quantitatively, however, task exerted a robust effect on the P600 effect, which was significantly larger in the explicit judgment task (Exp. 1b) than in the more passive comprehension task (Exp. 1a). This difference held true for both the semantic and the syntactic P600 effects. Finally, we also observed a larger N400 to real-world incongruities compared to the congruous critical words. However, in contrast to the robust effects of task on the P600, task did not significantly modulate the magnitude of the N400 effect.

As we outlined in the Introduction, there are two possible explanations for this pattern of results. First, the explicit judgment task may have introduced an artificial confound. On this account, the enhanced semantic and syntactic P600 effects observed in Experiment 1b were driven by explicit task requirements that are irrelevant to the goal of language comprehension, while the small P600 responses observed in Experiment 1a indexed the processing of anomalies during “natural” language comprehension.

An alternative interpretation of these findings is that the requirement to overtly judge coherence in Experiment 1b encouraged neurocognitive mechanisms that play an important role in deep, “natural” language comprehension — namely, the processing of comprehension errors (van de Meerendonk, et al., 2008; van Herten, Chwilla & Kolk, 2006; see also Brothers, Zeitlin, Choi Perrachione, Choi & Kuperberg, 2021). On this account, requiring an explicit coherence judgment increased participants’ engagement with the stimulus materials, and increased the likelihood that they detected and monitored for linguistic anomalies. That is, they were more likely to have adopted a mode of deep comprehension in which they were motivated to build a high-level representation of the events in each sentence (a “situation model” Van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). After encountering an anomaly that interrupted the build-up of this model, participants were more likely to perceive it as a comprehension error, triggering a P600 response. Note that, in Experiment 1a (passive comprehension), the presence of a robust N400 effect suggests that participants were processing the lexico-semantic properties of the critical words. This rules out an interpretation in which participants failed to attend to sentence materials altogether. Rather, we suggest that, in this experiment, participants encoded a more superficial semantic representation of the text (cf. Kintsch & van Dijk, 1978; Graesser, Singer & Trabasso, 1994) to answer the relatively infrequent comprehension questions. As a consequence, they may have ignored the linguistic anomalies or failed to detect them altogether (see Sanford et al., 2011), resulting in reduced P600 amplitudes.

If it is indeed the case that the amplitude of the P600 tracks participants' depth of comprehension, then other task manipulations that encourage deep comprehension should also increase the amplitude of the P600 response. Study 2 was designed to test this hypothesis.

Study 2 (Experiment 2)

Introduction

Explicit acceptability judgments are often used in studies investigating the P600, and, as demonstrated by the two experiments described above, the demands of this task enhance the magnitude of the P600 produced by both semantic and syntactic anomalies. Based on these differences, some have argued that the presence of a P600 effect with such task demands tells us little about “natural” language comprehension. Another view, however, is that, rather than introducing a fundamental confound, judgment tasks influence the P600 by encouraging processes that are engaged in deep comprehension.

In Experiment 2, we aimed to distinguish between these accounts by increasing participants' engagement without the use of an explicit judgment task, and to emulate a deep natural reading comprehension experience. We accomplished this in two ways. First, we embedded the same critical sentences used in Study 1 within rich and interconnected multi-sentence contexts that would motivate participants to construct a complete situation model. In addition, we redesigned our comprehension questions so that they required a deeper understanding of the text. Instead of simple yes/no questions that appeared only after some sentences, questions appeared after every stimulus, and took on a variety of forms. These included multiple-choice comprehension questions that required inference or attention to detail, as well as “open response” questions in which participants had to provide a one-sentence continuation to each scenario. We hypothesized that, with these new task manipulations, comprehenders would engage in the goal of deep language comprehension and perceive the linguistic anomalies as comprehension errors that are relevant to this goal.

To test our hypothesis, we directly compared the magnitude of the semantic and syntactic P600 effects in this study (Experiment 2) with the effects produced during the more passive comprehension of single sentences (Experiment 1a). If an enhancement of the P600 is driven solely by the externally imposed requirement to make an acceptability judgment, then we should see no differences between the effects produced in Experiments 2 and Experiment 1a (neither of which included a judgment task). If our second explanation is correct, and the P600 is relevant to the goal of deep language understanding, then we should see more robust P600 effects in Experiment 2 than in Experiment 1a.

Methods

Participants

We report ERP data from 24 native English speakers (15 female, average age, 21.6 years). Data from one additional participant was excluded due to excessive EEG artifact (see below). Inclusion criteria were the same as in Study 1.

Design and Materials

We measured ERPs to the same critical words in the same critical sentences as those used in Experiments 1a and 1b. However, we made several changes to the materials.

First, we created short, engaging story contexts that preceded each critical sentence. Each mini-story ranged from 3 to 5 sentences in length (mean length = 3.7, SD = 0.6) and the sentences containing the anomalies were always preceded by at least one introductory sentence (average number of introductory sentences: 2.3). The variable position of the critical sentences within the stories ensured that participants were not able to predict which sentence would contain an anomaly or incongruity. In some cases, nouns or proper names within the critical sentence were replaced with pronouns to avoid word repetition. However, the words immediately surrounding the critical word were never changed, and the position of the critical words within the critical sentences was the same as in Experiments 1a and 1b.

Second, to encourage deep reading comprehension, after each mini-story, participants saw one of three types of prompts: (a) an inference-requiring yes/no question (50% requiring a “yes” answer), e.g. “Is the new return policy good for customers?”, (b) a three-answer multiple choice question (choice order randomized), e.g. “When is Melanie’s big soccer game?” (“This weekend”; “Tomorrow”; “June”), or (c) a “Continue the story” prompt, which cued them to provide a plausible continuation of the story, stating this continuation aloud. To be able to answer these questions and produce coherent continuations, participants needed to read the entire scenario, including the final sentence, even if this did not contain an anomalous word. This therefore discouraged participants from disengaging from comprehension after encountering an anomaly.

Procedure

As in Experiments 1a and 1b, participants were randomly assigned to one of four counterbalanced lists, and the mini-stories appeared word-by-word using self-paced visual presentation. The final word of each sentence appeared with a period, and a fixation cross separated each sentence. At the end of each story, another fixation cross appeared, followed by one of the three comprehension prompts described above. Although all words, fixations, and questions were self-paced, words could not be advanced before 250ms had elapsed, fixation crosses could not be advanced before 700ms, and questions could not be advanced before 1250ms. Between each word there was also an interstimulus interval (ISI) of 10ms.

At the end of each mini-story, participants provided a manual or verbal response, depending on the type of comprehension prompt. For the yes/no questions and the three-answer multiple choice questions, numbered options appeared on the screen, and participants selected their preferred choice using a keyboard. For the open-response continuations, participants spoke their responses, which were recorded. The experiment was divided into four blocks, which were separated by short breaks.

EEG Data Acquisition and Preprocessing

EEG data was recorded using a BioSemi ActiveTwo system (biosemi.com) with 32 Ag/AgCl active electrodes in a modified 10/20 system montage. Signals were digitized at 512 Hz and a passband of DC - 104 Hz.

Processing was carried out using EEGLab (Delorme & Makeig, 2004) and ERPLab (Lopez-Calderon & Luck, 2014) analysis packages in the MATLAB environment. After importing the data for processing, the EEG was referenced offline to the average of the right and left mastoids, and a 0.05 - 30 Hz bandpass 24 dB per octave filter was applied to the continuous data.

The EEG, time-locked to each word in the discourse, was segmented into -300 to 1175ms epochs. Independent Components Analysis (using the extended infomax algorithm implemented in EEGLab) was then applied to the epoched data to correct for blinks and other eye-movement artifacts. Epochs containing residual artifacts such as blinks, eye-movements, or muscle artifacts were rejected prior to analysis. On average, there were 40 trials per condition following artifact rejection, with no significant differences in artifact rejection rates across conditions ($F(3,92) = 0.197, p = 0.898$).

ERP Statistical Analysis Strategy

As described in Experiment 1, we first verified that both the semantic and the syntactic anomalies produced significant P600 effects, relative to congruous critical words. We then compared the magnitude of the semantic and syntactic P600 effects produced in this experimental environment with those produced in Exp 1a by carrying out a 2 x 2 ANOVA that crossed Experimental Task (between-participant: Exp 2 vs. Exp 1a) and P600 Anomaly Effect (within-participant: semantic P600 effect vs. syntactic P600 effect). In addition, we carried out an independent samples t-test to determine whether the magnitude of the N400 effect evoked by the real-world incongruities (vs. congruous critical words) differed between the current experiment and Exp 1a.

Results

Behavioral Results

Overall, accuracy for the multiple choice and true/false questions was high (91%; see Table 1). To examine performance on the multiple choice and true/false questions, we performed a repeated measures ANOVA, which revealed no differences in accuracy across conditions ($F(3,69) = 0.39, p = 0.76$). Open response continuations were not included in this analysis.

ERP Results

P600 (600-1000ms)

Once again, relative to the congruous critical words, both the semantic anomalies ($t(23) = -4.65, p < .001$) and the syntactic anomalies ($t(23) = -4.54, p < .001$) produced robust P600 effects. As in Experiment 1a, the amplitude of the P600 did not differ between the semantic and syntactic anomalies ($t(23) = 1.30, p = .207$).

As shown in Figure 2, the magnitude of both the semantic and syntactic P600 effects appeared to be larger in this experiment than in Experiment 1a, which used isolated sentences and intermittent comprehension questions. This was confirmed by a 2 x 2 ANOVA that crossed Experimental Task (Exp 1a vs. Exp 2) with P600 Anomaly Effect (semantic P600 effect vs. syntactic P600 effect), which revealed a significant main effect of Experimental Task ($F(1,92) = 8.24, p = .005$), but no interaction between Experimental Task and P600 Anomaly Effect ($F(92) = 0.16, p = .69$).

A visual inspection of the data suggested that the P600 effects produced in this experiment were not only larger in magnitude than those produced in Experiment 1a, but also longer in duration. To assess this statistically, we computed point-by-point subtractions between anomalies and congruous critical words for each participant at each electrode site. We then calculated the midpoint of each of these P600 “difference waves” from 300-1175ms, using the 50% fractional area latency method (Hansen & Hillyard, 1980; Kiesel et al., 2008). Note that this measure of latency is unaffected by the overall amplitude of the effect. We then averaged these latency values across electrodes within each participant and carried out

independent samples t-tests comparing latency values of the semantic and syntactic P600 effects between the two experiments. These analyses confirmed that the P600 effects were significantly prolonged in Experiment 2 compared to Experiment 1a, both for the semantic anomalies (Exp. 1b: 772.16ms $\pm$ 17.22; Exp. 2: 859.96ms $\pm$ 15.01; $t(46) = 3.84, p < .001$) and syntactic anomalies (Exp. 1b: 763.79ms $\pm$ 12.06; Exp. 2: 835.92ms $\pm$ 16.30; $t(46) = 3.56, p < .001$).

N400 (300-500ms)

Once again, in this experiment, the real-world knowledge incongruities elicited larger N400 responses than the congruous critical words, $t(23) = 4.36, p < .001$. Again, the magnitude of this effect was not modulated by Task, $t(46) = 0.99, p = .325$.

Discussion

In Experiment 2, participants read multi-sentence discourse scenarios, word-by-word at their own pace, and responded to prompts after each scenario that encouraged deep comprehension. Similar to the previous experiments using the same anomalies, we observed robust P600 responses to both the semantic and syntactic anomalies, relative to congruous words. Critically, our between-experiment comparisons showed that both these P600 effects were larger than those observed under the more passive comprehension conditions when the same sentences were presented in isolation, and participants answered less demanding intermittent comprehension questions (Exp. 1a). In contrast, the real-world incongruities evoked a larger N400 than the congruous continuations, and this effect did not differ between Experiment 2 and 1a. This rules out the explanation that the differences in the P600 resulted simply from differential processing of the lexico-semantic features of individual words.

Together, these findings provide strong evidence *against* the claim that P600 to linguistic anomalies is simply an artifact of an externally imposed judgment task. Instead, we take the enhancement of this effect in the present study as evidence that this component indexes error processing mechanisms that are relevant for the goal of deep language comprehension. We will provide a detailed discussion of

this interpretation in the General Discussion. At this point, we emphasize two aspects of the current set of findings that speak to the goal-relevance of the P600.

The first observation is that, in this experiment, the magnitude of the P600 response was enhanced to the same degree to the semantic and syntactic anomalies (relative to Experiment 1a). This differs from the results of Experiment 1b (overt judgment) in which the syntactic P600 was enhanced even more than the semantic P600. We think that difference speaks to the distinction, highlighted in the General Introduction, between “task relevance” and “goal relevance”. We suggest that, in Experiment 1b, the P600 to the *syntactic* anomalies was particularly large because these anomalies were perceived by participants as more salient than the semantic anomalies for the immediate *task* requirement to judge the anomalies (as evidenced by a higher accuracy rate in this condition). Consistent with this interpretation are findings from an early study by Hahne & Friederici, 2002, who showed that the magnitude of the P600 to syntactic anomalies was larger when comprehenders were asked to detect syntactic anomalies, compared to when they were asked to disregard syntactic anomalies and only judge the semantic coherence of each sentence. We suggest that in Experiment 2, the semantic and syntactic anomalies were both perceived as equally relevant to the overarching goal of incrementally inferring the situation model being communicated.

The second observation of note is that, in addition to the enhancement of the *amplitude* of the semantic and syntactic P600s, the *durations* of these P600 responses were prolonged in Experiment 2, relative to Experiment 1a. We suggest that this was for two related reasons. First, the presence of additional discourse content *prior* to the anomaly led to more uncertainty about the source of the processing error. Second, the presence of additional discourse materials and more detailed comprehension questions *after* the critical word may have provided more *incentive* for such reprocessing

General Discussion

In these experiments we sought to answer a basic question: Is the posterior P600 response evoked by linguistic anomalies relevant for natural reading comprehension, or is it an artifact of an explicit instruction to detect the presence of anomalies? To this end, we measured P600s to two types of linguistic anomalies — semantic and syntactic — in three different task environments. First, we compared two tasks that have been commonly used to study the P600 — a passive comprehension task in which participants read a series of single unrelated sentences and answered intermittent True/False questions (Experiment 1a), and a more active task in which comprehenders made explicit acceptability judgments following each sentence (Experiment 1b). In both tasks, we observed qualitatively similar P600 effects to the semantic and syntactic anomalies, with similar time-courses and scalp distributions. However, consistent with previous studies, the amplitude of these P600 effects was significantly larger when comprehenders carried out the active judgment task.

We then examined the P600 under experimental conditions that aimed to better simulate deep text comprehension (Experiment 2). We embedded the same materials in rich, multi-sentence discourse contexts, and participants responded to engaging comprehension questions that appeared after each trial. Although no explicit acceptability judgments were required, we again observed an enhancement of the P600 response to both semantic and syntactic violations (compared to Experiment 1a). In addition, in Experiment 2, the P600 was more prolonged than in the previous single-sentence experiments.

Taken together, these findings provide clear evidence that the P600 response to linguistic errors is not simply an artifact of externally imposed task requirements. Instead, we suggest that the P600 effect was enhanced in Experiment 1b and Experiment 2 because, in comparison to Experiment 1a, they both engaged *deep language comprehension*, and that the P600 is closely linked to prolonged *decision making mechanisms engaged during reprocessing* that play an important role in error processing during deep language comprehension.

In the remainder of this Discussion, we will first provide a clearer definition overview of what we mean by deep language comprehension and explain how this links to the present findings. We will then link the present findings to the reprocessing and decision-making framework summarized in the previous chapter. We then turn to the practical and methodological implications of our findings.

The P600 is evoked when error detection disrupts deep language comprehension

To comprehend language deeply, we must do more than access the meaning of individual words and combine them into simple linguistic events that define “who does what to whom”; we must build a high-level situation model that links these events to real-world knowledge stored in long-term memory. If the detection of an error interferes with the goal of building a situation model, i.e. the goal of deep language comprehension, then this will trigger a P600.

We suggest that the reason why the P600 was enhanced in Experiment 1b (the active judgment task) and in Experiment 2 (discourse comprehension), in comparison with Experiment 1a (more passive single-sentence comprehension), is because both these tasks encouraged participants to actively build a situation model, which requires comprehenders to link events together along referential, temporal, and causal dimensions (Van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). In Experiment 2, this was encouraged by the presence of a discourse context, as well as by frequent and deep inference-requiring comprehension questions. We suggest that, in Experiment 1b, the explicit judgment task similarly encouraged participants to establish coherence across clause boundaries and integrate world knowledge from long-term memory. This was particularly likely given the presence of sentences with real-world incongruities. For example, to be able to make a judgment about whether the sentence “Every morning at breakfast the boys would plant...” makes sense, comprehenders would need to draw upon real-world knowledge about how likely it is that boys might plant at breakfast.

In contrast, in Experiment 1a, it is more likely that, instead of building a situation model, comprehenders constructed a *surface-level representation* of each sentence (cf. Kintsch & van Dijk, 1978; Graesser, Singer & Trabasso, 1994). Without establishing this situation model and engaging in the goal-

directed process of deep comprehension, they were therefore less likely to produce a robust P600 effect to the anomalies. Indeed, as noted earlier, no P600 is produced by linguistic anomalies when they appear in contexts that fail to establish a situation model (e.g. non-attracted semantic anomalies in very short sentence contexts, see Brothers, Wlotko, Warnke, & Kuperberg, 2020, Experiment 1; Gunter et al., 1999; Osterhout & Nicol, 1999) or in Jabberwocky sentences (Müntz, Matzke & Johannes, 1997; Ericsson et al., 2008).

Another reason why we think that comprehenders produced larger P600 effects in Experiments 1b and 2 (versus Experiment 1a) is that they were more likely to have consciously *detected* the comprehension disruptions. In situations where no situation model is established, being able to detect an anomaly is not sufficient to produce P600. For example, in some of the studies cited above, no P600 effect was produced, even though participants engaged in acceptability judgment tasks. However, being able to detect any disruption to deep comprehension of building a situation model does seem to be necessary for producing a P600 effect. This was illustrated in an elegant study by Sanford et al. 2011, who showed that, on trials in which comprehenders failed to detect anomalies, no P600 was observed.

We suggest that, in the passive comprehension task (Experiment 1a), participants failed to consciously detect the anomalies altogether on some proportion of trials. This does not imply that they didn't attend to the sentence materials at all in this experiment: the magnitude of the N400 effect produced by the real-world incongruous versus congruous critical words was the same as in the more active tasks, suggest that participants were accessing the lexico-semantic properties of the individual words. However, we suggest that the demands of reading hundreds of isolated sentences, many of them containing errors, followed by relatively superficial questions after only a proportion of trials (none of which referred to the anomalous inputs), led some participants to ignore some of the anomalies, resulting a smaller P600 effect on average.

The P600 and decision making

Together, these findings suggest that overt task requirements are not a sole determiner of the

P600. Rather, as discussed in Chapter 2, we suggest that P600s are instead produced when anomalous inputs are perceived as disrupting a broader existing goal of comprehension, committing comprehenders to reprocessing and a prolonged process of decision making that terminates only when comprehenders are confident that the disruption stemmed from a true incompatibility within the input (an external error), rather than an internal processing error. On this account, in Experiment 2, when anomalies were encountered in discourse contexts, participants spontaneously engaged similar decision making mechanisms to those they engaged when explicitly instructed to make acceptability judgments in Experiment 1b. Finally, as mentioned in the previous chapter, similar prolonged decision making mechanisms are thought to contribute to other prolonged posterior late positivities that are members of the P300 family of ERP components (Murphy, Robertson, Harty, & O'Connell, 2015; Boldt & Yeung, 2015; Desender, Murphy, Boldt, Verguts, & Yeung, 2019; Desender, Ridderinkhof, & Murphy, 2021). This highlights links between the P600 and this family of ERP components (Coulson, King & Kutas, 1998; Osterhout, Kim & Kuperberg, 2012; Sassenhagen & Fiebach, 2019; Sassenhagen, Schlesewsky, and Bornkessel-Schlesewsky, 2014).

Methodological implications: What is the ideal task?

Finally, our findings have methodological implications for the selection of experimental task in studies of the P600. As explained in the General Introduction, there has been much debate about the optimal task to use in ERP studies, and there has been some criticism of tasks that instruct comprehenders to make explicit judgments about the acceptability or coherence of sentences. To our mind, there is no ideal task. Rather, the selection of a given task depends entirely on the question being addressed by a particular study.

For example, several studies of the P600 have been focused on whether certain words and phrases can act as linguistic cues that lead the brain to *restrict* upcoming states to those that are compatible with the communication model: Can strong semantic attraction override syntactic cues to constrain upcoming inputs (e.g. Kolk, Chwilla, van Herten, & Oor, 2003; Kim & Osterhout, 2005, Experiment 1)? Does the brain use a verb's animacy constraints differently from other types of semantic or pragmatic contextual

constraints (e.g. Kuperberg et al., 2003; Kolk et al., 2003; Brothers, Wlotko, Warnke, & Kuperberg; Kuperberg, Brothers, & Wlotko, 2020)? During discourse comprehension, do certain types of discourse markers (e.g. concessive connectives (Xiang & Kuperberg, 2015) or presupposition triggers (Shetreet et al., 2019; Jouravlev et al., 2016) restrict the comprehender's communication model to specific sets of events or states? One way of addressing these types of questions is to design experiments that manipulate the presence/absence of these linguistic cues, and then to introduce subsequent words that are either compatible or incompatible. If an incompatibility produces a P600 effect, then this provides evidence that the prior linguistic cue did, indeed, constrain the comprehender's communication model. In these situations, we believe the use of an explicit coherence judgment task is appropriate. This is because it ensures that comprehenders detect any incompatibility if it is present. Without such instructions to actively monitor the input, it would be difficult to interpret the absence of a P600 effect to a word that is hypothesized to be incompatible: Was this because participants simply failed to detect the incompatibility? Or did the linguistic cue fail to restrict upcoming inputs, as hypothesized? A similar logic holds if a researcher is interested in whether a particular population is sensitive to a particular linguistic structure. For example, participants reading sentences in their second languages appear to engage in more shallow syntactic processing, as indicated by smaller or absent P600s to semantic reversal anomalies while they make overt acceptability judgments (Zheng & Lemhöfer, 2019).

More generally, as we have discussed, judgment tasks can encourage comprehenders to actively read and build coherence and minimize the use of shallow processing strategies. This is especially important when readers are being asked to comprehend multiple single unrelated sentences. We also note that the use of an overt detection task provides a powerful method of improving statistical power by increasing the amplitude of the P600 response. Finally, it can provide an additional behavioral measure for assessing individual performance, which can also be linked to neural measures across trials or participants.

On the other hand, researchers may be primarily interested in whether comprehenders are able to *spontaneously* engage decision making processes to track errors during online comprehension. In these

cases, then the use of a superimposed judgment task is obviously less appropriate. On the other hand, we caution again that the use of passive comprehension tasks, particularly when reading single sentences, may be equally inappropriate because the absence of a P600 effect would be difficult to interpret: A null effect may simply reflect a failure to engage in deep comprehension altogether, as seen in studies using impoverished sentence contexts (Brothers, Wlotko, Warnke & Kuperberg, 2020). Therefore, when addressing this question, we recommend adopting some or all of the methods employed in Experiment 2 in order to encourage deep comprehension. This includes embedding critical sentences in extended discourse contexts, providing meaningful content that extends beyond the appearance of the anomaly, and including demanding comprehension questions on every trial. In some circumstance it may also be beneficial to measure the P600 under conditions where participants have control over the pace of linguistic input (via self-paced reading), or through simultaneous eye-tracking and EEG (Metzner, von der Malsburg, Vasishth & Rosler, 2015).

Finally, as in the current study, the question of interest may be *how* different types of stimuli and task instructions influence the specific mechanisms indexed by the P600. For example, are these mechanisms more likely to be engaged when comprehenders are incentivized to read for later comprehension or for memory? How do different reading strategies influence these processes? Once again, by viewing the sensitivity of the P600 to task as a tool, rather than as an artifact, we believe that this component can provide valuable insights into the neurocognitive mechanisms underlying goal-directed language comprehension.

..

Supplementary Materials: Self-paced Reading Times

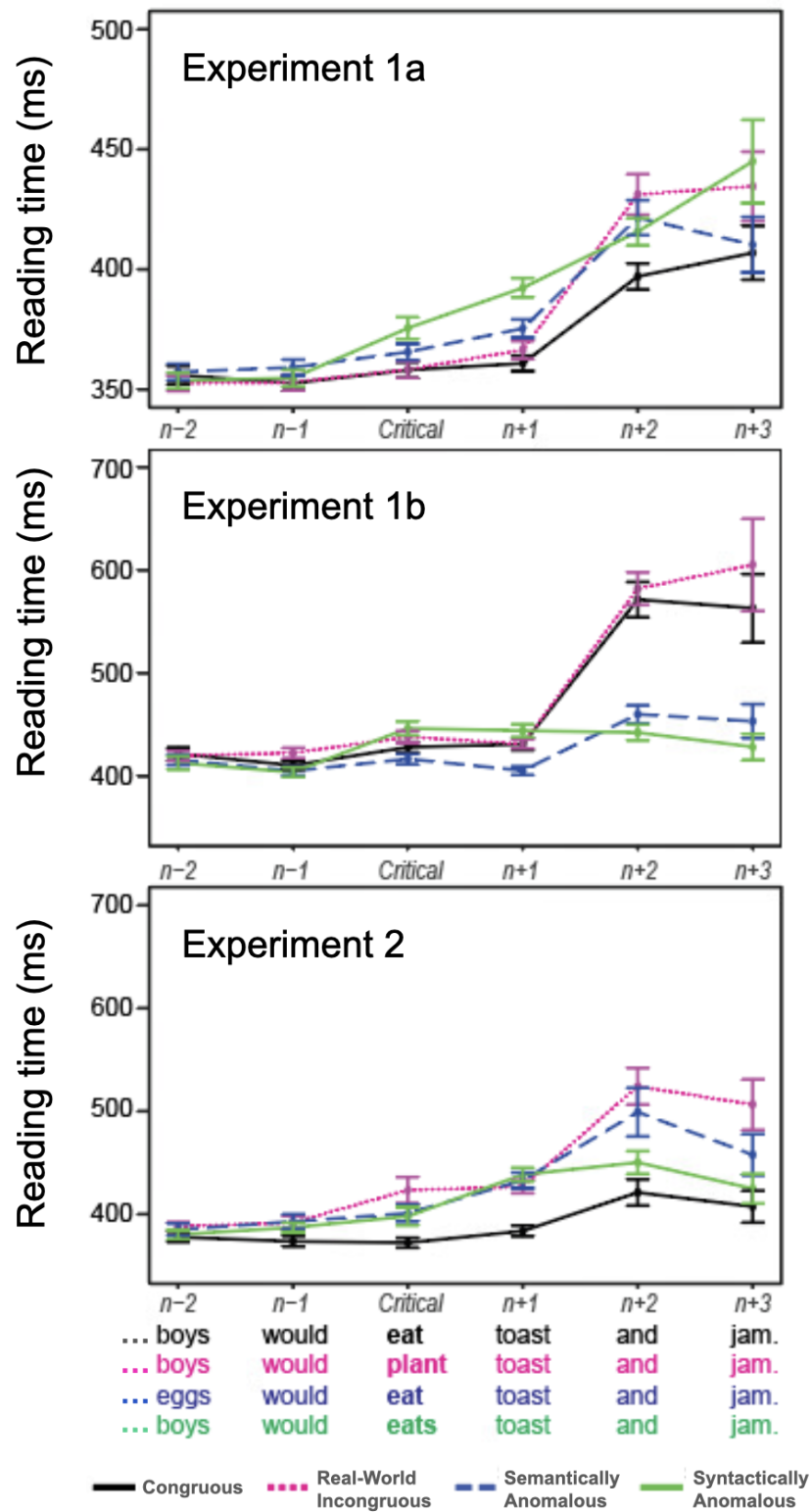


Figure 3. Self-paced reading times as a function of sentence position across all experiments. Each data point depicts the overall average time in milliseconds, with error bars representing ± 1 standard error of the mean calculated within participants. Sentence-final words could appear either two or three words past the critical word.

Experiment	Task	Condition			
		Congruous	Real-world incongruous	Semantically anomalous	Syntactically anomalous
1a	Passive comprehension	0.95 (0.07)	0.93 (0.07)	0.91 (0.07)	0.92 (0.06)
1b	Active judgments	0.93 (0.07)	0.94 (0.04)	0.99 (0.01)	0.99 (0.02)
2	Deep comprehension	0.91 (0.06)	0.91 (0.05)	0.89 (0.09)	0.91 (0.06)

Table 4. Mean accuracies for responses to questions in Experiments 1a, 1b, and 2. Standard deviations are reported in parentheses. Accuracy represents percentage of correct responses to intermittent comprehension questions (Experiment 1a), explicit acceptability judgments (Experiment 1b), or a combination of multiple-choice and true/false comprehension questions (Experiment 2).

Condition	Example Sentence (Experiments 1a and 1b)	Example Scenario (Experiment 2)
Congruous	For breakfast the boys would only <u>eat</u> toast and jam.	Every day, there was a rush to get the kids off to school. We
Real-world	For breakfast the boys would only <u>watch</u>	usually got up with little time
Incongruous	toast and jam.	to spare. For breakfast the
Semantically	For breakfast the eggs would only <u>eat</u> toast	<u>eggs/boys</u> would only
Anomalous	and jam.	<u>eat/watch/eats</u> toast and jam. I
Syntactically	For breakfast the boys would only <u>eats</u> toast	would usually have a bagel
Anomalous	and jam.	and a strong cup of coffee.

Table 5. Stimuli used in each experiment. Stimuli were identical between Experiments 1a and 1b. For Experiment 2, stimuli from Experiments 1a and 1b were embedded in larger contexts as described in Methods.

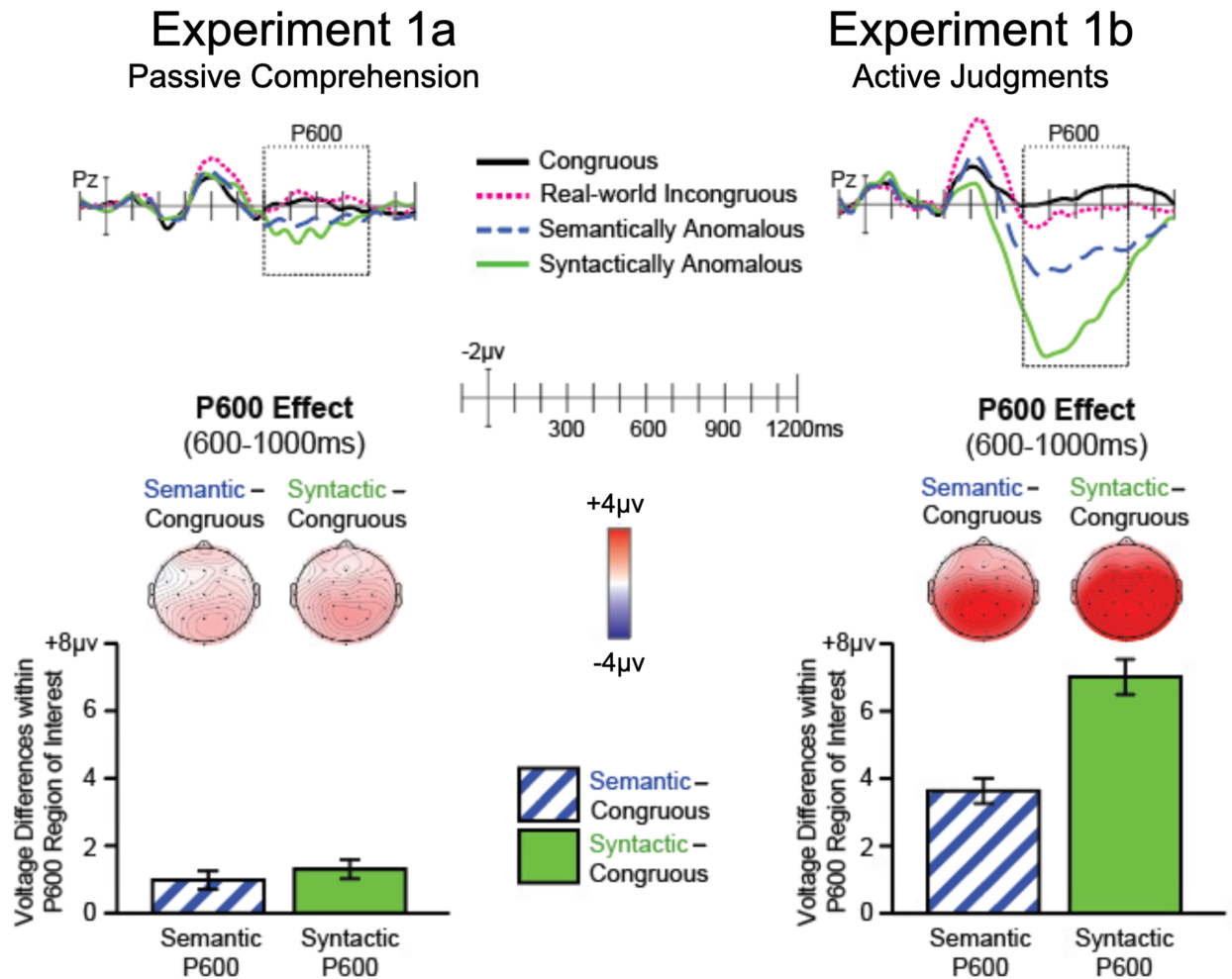


Figure 4. Grand-average event-related potentials (ERPs) in Experiments 1a and 1b, plotted at Pz. The bar graphs show the average voltages within our spatiotemporal region of interest (600-1000ms, CP1/2, Pz/3/4, Oz/1/2) for the semantic and syntactic P600 effects with error bars showing ± 1 standard error of the mean calculated within-participants. Topographic voltage maps show the distribution of the P600 effects over the scalp surface within the same region of interest.

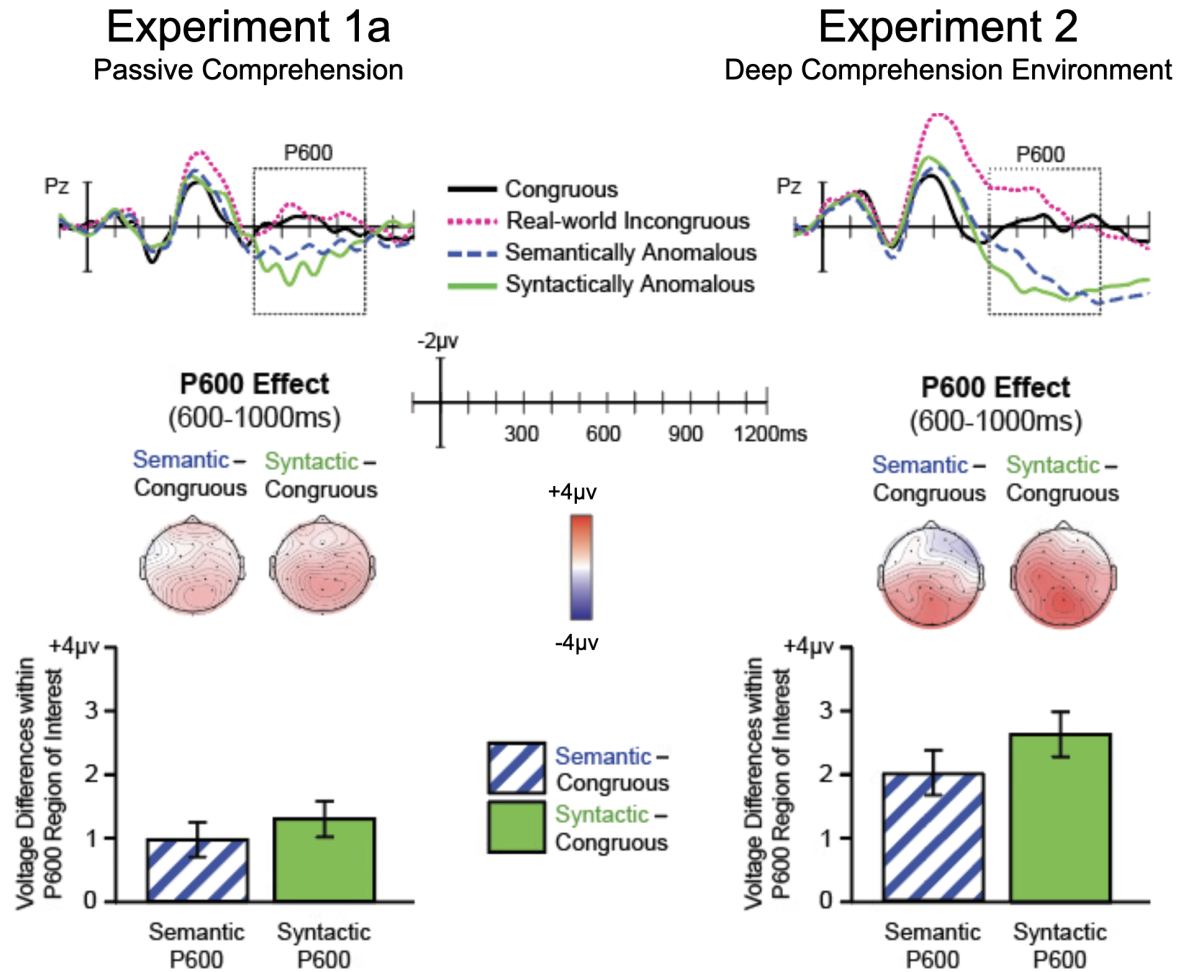


Figure 5. Grand-average event-related potentials (ERPs) in Experiments 1a and 2, plotted at Pz. The bar graphs show the average voltages within our spatiotemporal region of interest (600-1000ms, CP1/2, Pz/3/4, Oz/1/2) for the semantic and syntactic P600 effects with error bars showing ± 1 standard error of the mean calculated within-participants. Topographic voltage maps show the distribution of the P600 effects over the scalp surface within the same region of interest.

Chapter 4: Contributions of Naturalistic Stimuli on the N400

In chapters 2 and 3, we discussed the nature of the P600, a posteriorly distributed ERP component occurring 600-1000ms following the onset of anomalous words. In chapter 3, we emphasized the importance of naturalistic contexts for understanding reading comprehension. In this chapter, we again focus on naturalistic contexts, but switch gears to look at an earlier ERP component known as the N400, a centro-parietally focused ERP component occurring 300-500ms following word onset and thought to reflect initial stages of lexico-semantic access.

In contrast to the previous chapters where we examined averaged ERP responses on "critical words" within our stimuli, here we take a different approach: we analyze N400 responses produced on each word within the multi-sentence discourses used in Chapter 3 (excluding the anomalous inputs). This paradigm afforded us the opportunity to examine how the N400 is modulated in these naturalistic contexts.

Note that the onset of the COVID-19 pandemic interrupted data collection for this study, and we see this work as relatively preliminary in comparison with the previous chapters. Each of the questions we address are associated with large literatures, and so it is likely that the results of each question will be reported in separate future manuscripts.

As in the previous chapter, Trevor Brothers Ph.D. assisted with study design. I thank Cassidy Prejean, Elizabeth Kenneally, and Maya Lazarus, who assisted with data collection and analysis. In addition, Trevor Brothers Ph.D. and Anthony Yacovone Ph.D. provided assistance with writing and analysis. Gina Kuperberg, MD. Ph.D. also assisted with the writing.

Introduction

The N400 component is a negative-going event related potential (ERP) typically observed between 300 and 500ms following word onset and that is maximal over central posterior electrodes. It was first described as being larger to words that are anomalous vs congruent in context (Kutas & Hillyard,

1980). However, soon afterwards it was shown to be influenced by the predictability of words within plausible sentences as well (Kutas & Hillyard, 1984). For example, N400s were attenuated to the highly predictable word *stamp* in “He mailed the letter without a stamp” relative to the unpredicted plausible word *hour* in “The bill was due at the end of the hour”.

This attenuating effect of predictability on the N400 has since been replicated many times in the experimental literature (e.g. DeLong, Urbach & Kutas, 2005; Van Berkum, Hagoort & Brown, 1999; Federmeier & Kutas, 1999; DeLong, Chan & Kutas, 2019). In these experimental studies, predictability is usually operationalized as cloze probability – the percentage of respondents in an offline norming task completing a sentence fragment with a given word. This effect of predictability on the N400 is usually interpreted as reflecting the ease of accessing lexico-semantic features of incoming words. This access is facilitated when lexico-semantic information has been probabilistically pre-activated by prior context in a top-down fashion in advance of encountering the input.

Most of the initial work establishing the effect of predictability on the N400 used tightly controlled sets of constructed stimuli, which were typically isolated sentences, each revolving around a set of ‘critical words’. This allowed experimenters to isolate the effect of predictability by holding lexical and contextual variables constant, including the position of critical words within sentences. However, it has long been known that the N400 is evoked by all words in sentences. For example, in a seminal study by Van Petten and Kutas (1991), the amplitude of the N400 became progressively smaller to later words in congruent sentences (see also Payne, Lee & Federmeier, 2015; Dambacher, Kliegl, Hofmann, & Jacobs, 2006). It is also known that the N400 can be affected by extra-sentential semantic context in paragraphs. However, this latter work has typically involved comparisons between contextually congruent vs incongruent critical words (see e.g. Van Berkum, Hagoort & Grown, 1999).

More recently, following a general move in psycholinguistics towards using more ecologically valid stimuli, several researchers have begun to look at the effects of predictability on the N400 in these more naturalistic contexts (see Brennan, 2016). These studies have typically made use of language corpora or book chapters (e.g. Heilbron, Armeni, Schoffelen, Hagoort & De Lange, 2022; Brennan &

Hale, 2019), and include a much wider variety of sentence structures. Just as critically, this approach allows experimenters to measure ERPs to every word.

Since collecting cloze probability for every word in longer contexts would be costly in time and resources, these newer studies have operationalized predictability in different ways. Earlier studies used n-gram probability, which measured occurrences of specific combinations of words, but correlated poorly with cloze probability (e.g. Goodkind & Bicknell, 2018; Brennan & Hale, 2019; Yan & Jaeger, 2019). However, recent advances in large language models (LLMs) have enabled researchers to generate predictions that highly correlate with cloze probability from human norming studies (Michaelov, Coulson, & Bergen, 2022; Heilbron et al., 2022).

Instead of carefully controlling for a large number of potentially confounding variables by using tightly controlled stimuli, these experiments relied on methods such as mixed effects regression to statistically control for confounding variables on every word.

Collectively, these studies have replicated the finding that the N400 is sensitive to lexical predictability in naturalistic contexts (Heilbron et al., 2022; Brennan & Hale, 2019; Michaelov, Coulson, & Bergen, 2022; Goodkind & Bicknell, 2018; Yan & Jaeger, 2019). However, these studies have not yet examined many other questions that have been asked in the experimental literature. For example, we know from this literature that the N400 is sensitive to lexical factors such as frequency and semantic relatedness, and that these factors can interact with lexical predictability on the N400. We know very little, however, about how these findings generalize to naturalistic contexts, where behavioral studies have shown contradictory results. In this chapter, we examine these findings in a maximally naturalistic context.

1. Do frequency and lexical predictability interact in naturalistic contexts?

The first open question concerns the interaction of lexical variables and lexical predictability. It has long been known from the experimental literature that the N400 is sensitive to certain lexical characteristics for words presented in isolation. For example, words that are more frequent (e.g. cold)

produce a smaller N400 compared to uncommon words (e.g. wart) when these words are presented in isolation (Rugg, 1990; Van Petten & Kutas, 1990). This interaction between lexical predictability and frequency has also been observed over the course of short, isolated sentences, where words appearing later in congruent sentences (which were presumed to be more predictable) have been shown to have smaller effects of frequency (Van Petten & Kutas, 1991; Payne, Lee & Federmeier, 2015). This interaction has also been replicated using cloze probability in isolated sentences (Dambacher, Kliegl, Hofmann, & Jacobs, 2006), and in MEG using model-derived cloze probability in single sentences (Huizeling, Arana, Hagoort, & Schoffelen, 2022).

However, this interaction has not yet been demonstrated during naturalistic reading beyond short, single sentence contexts. Understanding whether this interaction on the N400 would replicate in natural contexts is particularly important given evidence from studies using eye-tracking as a dependent measure. These studies have suggested that, instead of interacting, frequency and predictability instead have completely independent effects on comprehension in naturalistic reading (see Staub, 2015, table 2 for review of studies). Further, two studies using fixation-related potentials (FRPs) failed to find an interaction between frequency and predictability on the N400 (Kretzschman, et al., 2015; Degno et al., 2018), and a recent fMRI study using model-derived cloze probability in larger discourse fragments found independent effects of both frequency and predictability with no evidence for an interaction (Shain et al., 2023).

2. What is the role of semantic relatedness in naturalistic contexts?

The second open question regards the relationship between predictability and simple semantic associations during reading comprehension. It has been known for many years that the N400 is influenced by single word contexts in semantic priming paradigms (e.g. Bentin, McCarthy & Wood, 1985; Rugg, 1985). Under automatic conditions (i.e. very short stimulus onset asynchronies between prime and target), this effect is thought to be driven primarily by bottom-up ‘spreading activation’, which

is different from the type of top-down predictive preactivation that is thought to drive the probabilistic effects of broader contexts.

There is some existing experimental evidence suggesting that semantic relatedness interacts with effects of context during language comprehension. However, these studies have primarily examined congruent vs incongruent critical words, or scrambled vs congruent contexts. For example, the effect of semantic relatedness between word pairs on the N400 appears to be smaller in congruent vs incongruent sentences (Van Petten, Weckerly, McIsaac & Kutas, 1997; Simpson, Peterson, Casteel, & Burgess, 1989; Boudewyn, Long & Swaab, 2015).

In addition, effects of semantic relatedness between word pairs in sentences seem to only be observed when lexical associates are close to the critical word, and not when they are further away within the sentence (e.g. opened–closed in "She fixed the sticky drawer so that it *opened* and *closed* easily." vs gun–shoot in "After taking his wallet, they waved a *gun* and threatened to *shoot* him if he reported it."; Van Petten, Weckerly, McIsaac & Kutas, 1997; Federmeier, Van Petten, Schwartz, & Kutas, 2003). In this case, even the presence of a single intervening content word seems to eliminate the effect of lexical association, and the authors argued that this is because bottom-up spread of activation from semantically related words is short-lived. However, no study has systematically looked at the interaction of predictability and local semantic relatedness during natural language comprehension.

Present Study

In order to address these questions, we used the naturalistic, paragraph-length stimuli from Chapter 3. These stimuli were self-paced and paired with questions after each paragraph that encouraged deep reading comprehension, with the predictability of each word assessed using a separate incremental cloze task.

Most studies of reading comprehension have used rapid serial visual presentation (RSVP), a technique in which words are presented sequentially at the center of a screen at a fixed presentation rate. However, having a fixed time per word is quite different from how reading unfolds in naturalistic

contexts, where certain words are fixated on for longer durations. In addition, the presentation rate of stimuli is known to affect the way our brain reacts to language input. Faster presentation rates (e.g. 200ms) bias comprehenders towards an increased influence of lexical factors, while lower rates bias towards an increased influence of context on the N400 (Camblin, Ledoux, Boudewyn, Gordon, & Swaab, 2007; Swaab, Camblin, & Gordon, 2004; discussed further in Ditman, Holcomb & Kuperberg, 2007). Fortunately, it is now possible to combine ERPs with self-paced reading (Ditman, Holcomb & Kuperberg, 2007; Wlotko & Federmeier, 2015). This allows participants to take the time they need to process each incoming word, and therefore more closely approximates the dynamics of natural reading.

A second goal of our design was to maximally engage the participant in the task of comprehension. Previously, ERP studies of language comprehension have either used explicit plausibility judgments, or infrequent, minimally engaging comprehension questions. Here, participants are randomly given one of three possible types of comprehension questions after each story, as discussed in Chapter 3. These appeared as three-answer multiple-choice questions, yes/no questions requiring participants to make some inference about the characters or situation, or open-response questions, where participants were asked to provide a plausible and specific continuation by speaking their answers aloud. Participants did not know ahead of time which type of question they would receive. These questions ensured that participants were engaging in deep comprehension of the stimuli.

Finally, we collected cloze probabilities for every word in each of our paragraph length stimuli. While multiple studies have now assessed the effect of predictability on the N400 in naturalistic contexts using probabilities derived from computational models, these measures do not correlate perfectly with cloze probability, which directly reflects online human prediction, and it is not a given that model-derived probabilities are fully capturing the lexical predictability of words.

Using this design, we addressed each of the open questions laid out above. First, we examine the independent contributions of frequency and predictability of words in natural contexts, and whether there is any interaction between the effects. Second, we examine the impact of semantic relatedness and again determine if the effect of semantic relatedness interacts with predictability on the N400. To accomplish

this, we combine measures of semantic relatedness derived from semantic vector distance between pairs of words, human ratings of cloze probability, and distance of related lexical items.

Following the Methods, results are presented for each question. Discussion of questions, and broader implications of this work, will be revisited in the General Discussion.

Methods

Participants

For each open question, we analyzed data from the same set of 22 participants (9 male; mean age = 21.6). Data from 2 additional ERP participants were excluded due to excessive EEG artifact ($\geq 30\%$ of trials rejected). All were right-handed by self-report and as measured by a modified version of the Edinburgh Handedness Inventory (Oldfield, 1971; White & Ashton, 1976). Further, all participants were native speakers of English, had no other significant language exposure before age 5, had normal or corrected-to-normal vision, were not taking any psychoactive medications, and had no reported diagnoses of psychiatric or neurological disorders. Written consent was obtained from all participants before participation according to the guidelines of the Tufts IRB.

Construction of Stimuli

For this study, we created 180 novel multisentence discourse scenarios. Each story was three to five sentences in length (mean length: 3.67, SD: 0.56), with each sentence containing 4 to 19 words (mean = 10.88, SD = 2.27) see Appendix X. These story contexts were short narrative texts that often described interactions between multiple characters in a particular setting. The stories were written to be engaging and to cover a wide range of topics and scenarios.⁶

⁶As these stimuli were expanded based on the stimuli in Kuperberg et al. (2007), some of the sentences contained either semantically or syntactically anomalous verbs (e.g. *Every morning for breakfast the eggs would eat...*). Participants were informed that some of the sentences might contain anomalies, but that they should continue to try to understand the story to the best of their ability. In this study, all ERP activity following these critical words was excluded from the present analyses (see Chapter 3 for additional details).

Each story was followed by one of three types of questions: yes/no questions, three-answer multiple choice questions, and open-response questions. Yes/no questions were designed to require some sort of inference on the part of the participant. Three-answer multiple choice questions focused on specific facts from the prior story. For open-response questions, a prompt was presented in the middle of the screen (“Continue the story”), and participants were asked to provide a plausible continuation of the story using a single sentence. There were 60 of each type of question. Of the 60 yes/no questions, 30 had ‘yes’ as a correct answer.

Operationalization of Lexical Characteristics

We first characterized all of our words by their position within sentences, and by their sentence’s position within the paragraph. For example, the word 3 in sentence 2 has a word-in-sentence position of 3, and a sentence-in-paragraph position of 2. Then, we characterized the individual properties of the words themselves: class of word (content or non-content), which stimulus they appeared in (1-180), word frequency, concreteness, orthographic neighborhood, word length, predictability from prior context, and semantic relatedness with prior context.

Sentence position was calculated using all words in the sentence (even when only content words were analyzed), with the first word of each sentence being assigned a value of 1. Word frequency was operationalized as the logarithm of the frequency per million words in the SUBTLEX-US corpus (Brysbaert & New, 2009). Concreteness values (1-5) were taken from Brysbaert, Warriner, & Kuperman (2014). Orthographic neighborhood density was calculated using the orthographic Levenshtein distance (OLD20) measure (Yarkoni, Balota, & Yap, 2008) with larger OLD20 values representing denser orthographic neighborhoods.

Operationalization of Lexical Predictability (Cloze Probability)

Predictability from prior context was measured using sentence continuations (cloze), which were collected offline in a separate set of participants. Cloze was collected for each word sequentially,

following the method outlined in Luke & Christianson (2018), described in detail below. Each participant scored one of 9 non-overlapping lists of 20 paragraphs.

Participants were instructed to “please type the word that you think will come next.” Beneath these instructions was part of the paragraph, as well as a space for participants to enter their response. The first trial for each paragraph displayed the first word alone, and subsequent trials displayed the first two words together and so forth until the end of the paragraph was reached until each participant scored 20 paragraphs. In total, we collected ratings from 477 participants. Of these, 4 were excluded for being less than 50% complete. Another 96 datasets were excluded for poor accuracy ($<10\%$). We determined this threshold by examining the distribution of accuracy scores across participants, which was strongly bimodal. Manual examination of the datasets with less than 10% accuracy revealed that few of these responses appeared to be made in good faith (rather, low accuracy tended to result from participants entering nonsense responses, e.g. random strings of letters or the same word over and over). An additional 6 outlier datasets were excluded for unusually high accuracy (accuracy $> 50\%$), which would be extremely unlikely unless participants had repeated the experiment more than once, and were thus familiar with the stimuli. Cloze for words was computed from the remaining 361 participants (mAge: 33.54, 168 male), resulting in an average of 40.06 cloze ratings for every individual word in our plausible stimuli (SD = 3.91, min = 28, max = 46). No cloze ratings were collected for the first word of each paragraph. On average, content words in our stimulus set had a cloze probability of 15.3% (SD = 24.5%), and function words had a cloze probability of 33.2% (SD = 32.3%). Cloze probability increased on average over the course of sentences and was higher for sentences later in the paragraph (see figure 6).

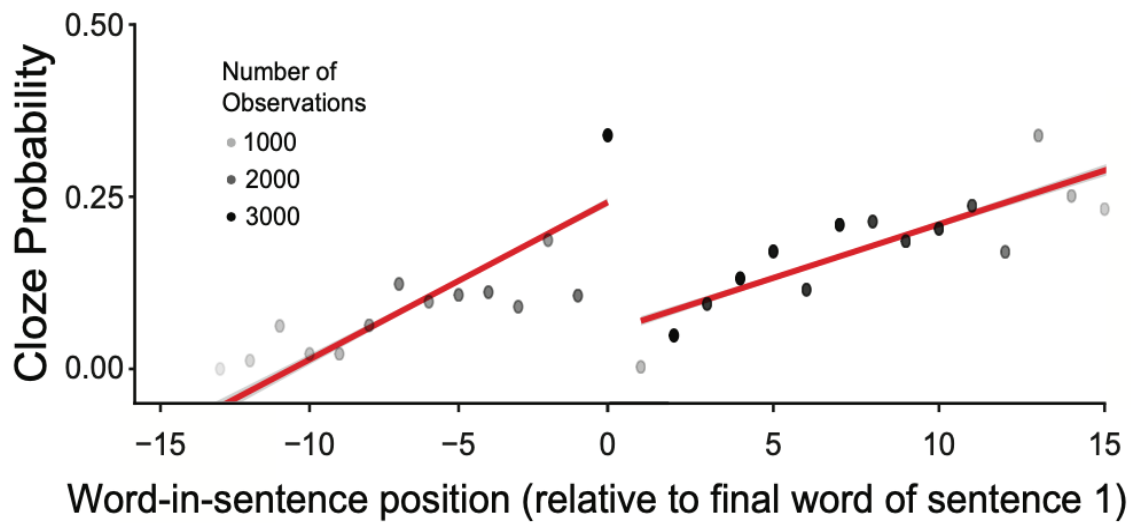


Figure 6. Averaged cloze probability to content words as a function of word-in-sentence position relative to the final word of sentence 1. Regression lines are calculated within each sentence individually, and gray regions represent standard error. Opacity of data points represents the number of data points for each word position.

ERP Procedure

In the main ERP experiment, participants read a series of 180 unrelated paragraphs one word at a time at their own pace. After receiving instructions and their response keyboard, each participant was given a four-trial practice to ensure that they understood how to proceed through the experiment. Participants were then randomly assigned to one of four lists used for counterbalancing. Participants sat in a comfortable chair in a dimly lit and quiet room. Stimuli were presented one word at a time displayed in the middle of a computer screen. When the participant was finished reading the word, they pressed a button on their keyboard to proceed to the next word of the stimulus.⁷ The final word of each sentence appeared with a period. Between each sentence of the story, a fixation cross was presented. After each

⁷ Although all words, fixations, and questions were self-paced, there was a minimum time for the presentation of each: words in the sentences could not be advanced before 250 ms had elapsed, fixation crosses could not be advanced before 700 ms, and questions could not be advanced before 1250 ms

story was presented in full, another fixation would appear, followed by one of the three question types. As some stimuli contained anomalies, participants were told that some sentences may “seem weird”, but that they should always try to pay attention to the gist of the story so that they can answer the follow-up questions. Between each word there was also an interstimulus interval (ISI) of 10 ms.

For each question type, the participant was directed to make some sort of response. For three-answer multiple-choice and yes/no questions, numbered options appeared on the screen and participants selected the number of their preferred choice on the keyboard in their lap. For open-response questions, a prompt was presented in the middle of the screen (“Continue the story”), and participants were asked to provide a plausible and specific continuation. Participants spoke their answers aloud, which were captured by a microphone situated on the opposite side of the room to minimize electrical interference.

In total, the experiment took approximately 1.5 hours. The experiment was divided into 4 blocks, separated by rest breaks.

EEG recording

EEG data was recorded using a BioSemi ActiveTwo system (biosemi.com) with 32 Ag/AgCl active electrodes in a modified 10/20 system montage. Signals were digitized at 512 Hz and a passband of DC - 104 Hz. Processing was carried out using EEGLab (Delorme & Makeig, 2004) and ERPLab (Lopez-Calderon & Luck, 2014) analysis packages in the MATLAB environment. After importing the data for processing, the EEG was referenced offline to the average of the right and left mastoids, and a 0.05-30 Hz bandpass 24 dB per octave filter was applied to the continuous data.

The EEG, time-locked to each word in the discourse, was segmented into -300 to 500 ms epochs. Independent Components Analysis (using the extended infomax algorithm implemented in EEGLab) was then applied to the epoched data to correct for blinks and other eye movement artifacts. Then, epochs containing residual artifacts such as blinks, eye-movements, or muscle artifacts were rejected prior to analysis. On average, there were 7164 trials per condition following artifact rejection (range: 6852-7207).

Data analysis

For each of our open questions, we analyzed trial-level ERP amplitudes using a series of linear mixed-effect regression models and restricted maximum likelihood estimation. Doing this allowed us to appropriately control for a number of partially correlated continuous item-level predictors. We further included random effects for items, referring to each of the discourse contexts, as well as participants.

All models were analyzed using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014) in R (R core team, 2014). With one exception, all models used an averaged spatiotemporal region of interest including all time points between 300 and 500 ms at 9 electrode locations (Cz, C3/4, CPz, CP1/2, Pz, P3/4) where previous studies have shown N400 effects to be maximal. For the remaining analysis assessing the effect of length, we selected an earlier, more posterior window (F3/z/4, C3/z/4, FC1/2) between 175 and 225 ms, where word length effects have been observed to be maximal in previous studies.

Predictors of variance for the words included word-in-sentence position, sentence-in-paragraph position, log word frequency, concreteness, orthographic neighborhood (operationalized as orthographic Levenshtein distance to the 20 closest words, or OLD20), word length (in letters), cloze probability, and semantic relatedness with prior context (calculated as Word2Vec distance between the word and each of the four preceding content words), as defined above. All variables were mean-centered and standardized. For stimuli that contained some form of anomaly, the anomaly and any subsequent words in that story were always excluded from our analysis.

For our statistical analyses, we specified a fully maximal random effects model with random intercepts and slopes for both subjects and items (defined as unique word tokens). Random slopes were specified for each fixed effect of theoretical interest. For any models that fail to converge, we simplified the random effects structure using the best path algorithm (Barr et al, 2013).

Results

1. Do frequency and lexical predictability interact in naturalistic contexts?

As discussed in the introduction, it is shown in both the naturalistic and experimental literature that the N400 is sensitive to a word's predictability given full prior context. In experimental studies using single sentences, there is also evidence that the effect of frequency on the N400 interacts with the effect of predictability. Specifically, the effect of frequency is smaller to words presented later in supportive sentences, which are presumably more predictable (Van Petten & Kutas, 1991; Payne, Lee & Federmeier, 2015). Here, we aim to replicate this finding during self-paced reading in naturalistic paragraphs.

Model 1: Is the effect of frequency influenced by the contextual predictability of words?

In Model 1, we aimed to replicate the interaction between the effects of frequency and predictability on the N400. We fit a model that predicted the amplitude of the N400 (averaged over C3/z/4, CP1/z/2, P3/z/4, 300-500ms), with fixed effects of Frequency, Cloze probability, and an interaction between Frequency and Cloze probability. If the effect of frequency on the N400 is attenuated to highly predictable words, we should observe a significant interaction between Frequency and Cloze probability, and follow up models within high and low Cloze probability subgroups should confirm a smaller effect of frequency within the high Cloze probability subgroup.

Consistent with prior findings, we observed an interaction between frequency and cloze ($b = -0.15$, $t = -2.84$, $p = 0.005$) and a main effect of cloze ($b = 0.77$, $t = 7.50$, $p < .001$), but no main effect of frequency ($b = 0.12$, $t = 1.45$, $p = .145$).

In order to follow up on the interaction with Cloze probability, we divided the data into groups with above median or below median cloze probability, and fit two additional models which were identical to the first, but without the inclusion of Cloze. As predicted, within the low Cloze subgroup, there was an effect of Frequency on the N400 ($b = 0.19$, $t = 2.28$, $p = 0.026$). However, there was no effect of

Frequency on the N400 in the high Cloze subgroup ($b = 0.07$, $t = 0.54$, $p = .588$), confirming that word frequency has a reduced impact on lexico-semantic access in words that are highly predictable.

Table 6. Effects of Frequency and Cloze on the N400

Model 1 (N400)		Estimates		
Predictor	b	SE	t -value	p -value
Frequency	0.12	0.08	1.45	0.145
Cloze	0.77	0.10	7.50	<0.001*
Frequency x Cloze	-0.15	0.05	-2.84	0.005*

Random Effects Structure: (Cloze*Frequency || subject) + (1 | item)

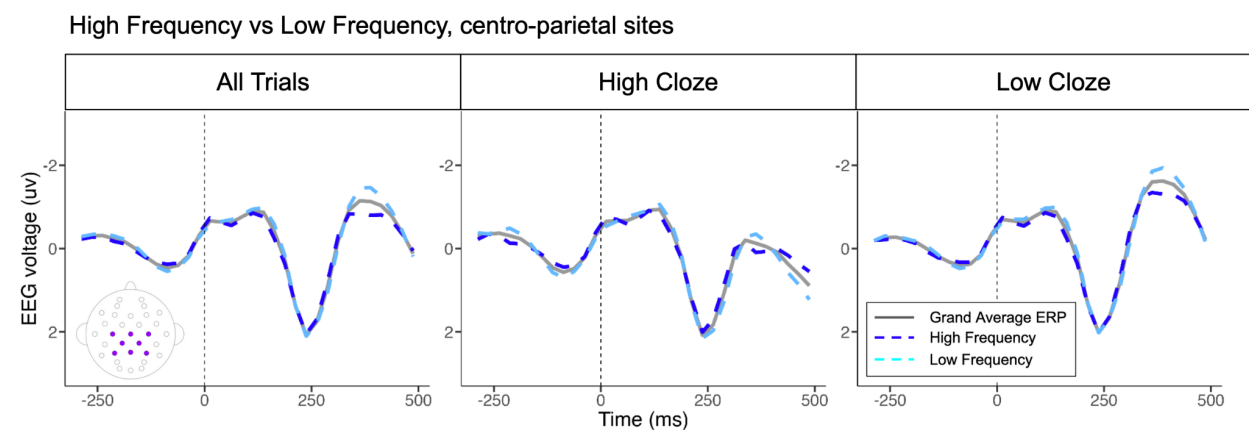


Figure 7. Averaged N400 amplitudes to content words in our spatial region of interest (P3/z/4, C3/z/4, CP1/2) plotted with all trials (left), only trials in the top third of values for cloze probability (center), and only trials in the bottom third of values for cloze probability (right). The gray line represents the grand average ERP within this spatial region of interest. To compute the effect of frequency, ERP data was binned temporally into 25ms time windows, and model 2 was fit individually at each time point. At each time point, the dotted line represents the estimated ERP amplitude to words that are 1 standard deviation

higher (dark blue) or 1 standard deviation lower (light blue) frequency than average, based on the beta weights extracted from the models.

Table 7. Effect of frequency, follow-up analyses

Model 1 (N400)		Estimates		
Predictor	<i>b</i>	<i>SE</i>	<i>t-value</i>	<i>p-value</i>
Frequency, high cloze trials	0.07	0.13	0.54	0.588
Frequency, low cloze trials	0.19	0.09	2.28	0.026*

Random Effects Structure: (Frequency || subject) + (1 | item)

2. What is the role of semantic relatedness in naturalistic contexts?

As discussed in the introduction, semantic relatedness has been shown to interact with predictability. However, this has not been rigorously demonstrated using cloze probability as a measure of predictability. Instead, prior studies have inferred predictability by contrasting effects of semantic relatedness in congruent sentences vs sentences which are either incongruent or scrambled (e.g. Van Petten, Weckerly, McIsaac & Kutas, 1997; Simpson, Peterson, Casteel, & Burgess, 1989; Boudewyn, Long & Swaab, 2015).

Here, we will assess the possibility of an interaction between predictability and semantic relatedness on the N400. We also ask how *long* the effects of semantic relatedness can influence incoming words. Specifically, we ask if semantic relatedness only affects immediately preceding content words as in Van Petten, Weckerly, McIsaac & Kutas (1997), or if the effect of semantic relatedness lasts longer in self-paced, naturalistic paragraphs.

Methods

In order to test the effect of semantic relatedness with prior words on the amplitude of the N400, we calculated the semantic similarity between each content word and each of the four preceding content

words. Preceding content words could cross sentence boundaries within paragraphs. Words that did not have at least four preceding content words were excluded from this analysis.

Semantic relatedness between words was assessed using word2vec (Mikolov, Chen, Corrado, & Dean, 2013). In word2vec, each word is assigned a vector by a neural network, such that words found in similar contexts have similar vectors. Similarity between words is measured by calculating the cosine distance between their respective vectors.

In addition to the semantic relatedness values for the previous four content words, we included cloze probability, as well as our standard control variables (length, frequency, orthographic neighborhood, and concreteness).

Results

Model 1: Are the effects of local Semantic Relatedness with prior words on the N400 distinct from cloze probability? How far back in sentences does semantic relatedness between words remain important?

To test the effects of Semantic Relatedness with prior words on the N400, we fit a model with fixed effects for the Semantic Relatedness between each word and each of the four preceding content words, as described in the Methods. We additionally included a fixed effect of cloze probability, and an interaction between cloze probability and the Semantic Relatedness between the current word and the immediately preceding content word (to which we assumed the effect of local Semantic Relatedness would be the strongest). As before, we also included our standard control variables (frequency, concreteness, orthographic neighborhood, and word length).

By measuring the Semantic Relatedness to each of the four preceding content words, we should be able to determine if the brain holds on to and uses semantic features from previous words in the access of new lexico-semantic information

When controlling for cloze probability, we observed a robust effect of semantic relatedness of the immediately preceding content word on the N400 ($b = 0.30$, $t = 5.13$, $p < .001$). However, content words further back in the sentence had no independent influence on the amplitude of the N400 (all $ps > .1$). As

expected, there was an independent effect of cloze probability ($b = 1.91$, $t = 4.66$, $p < .001$), as well as an interaction between cloze and the semantic relatedness for the immediately preceding content word ($b = -.17$, $t = -3.91$, $p < .001$).⁸

9

Table 8. Effects of Semantic Relatedness with prior words and Cloze on the N400

Model 1 (N400)		Estimates		
Predictor	<i>b</i>	<i>SE</i>	<i>t-value</i>	<i>p-value</i>
Cloze	0.71	0.11	6.61	<0.001*
Semantic Relatedness CW-1	0.30	0.06	5.13	<0.001*
Semantic Relatedness CW-2	0.07	0.05	1.23	0.22
Semantic Relatedness CW-3	0.08	0.05	1.63	0.10
Semantic Relatedness CW-4	-0.08	0.05	-1.51	0.13
Cloze:Semantic Relatedness CW-1	-0.17	0.04	-3.91	<0.001*

Random Effects Structure: (Cloze*Semantic Relatedness CW-1 || subject) + (1 | item)

Taken together, these findings suggest that there is an effect of the semantic relatedness with the immediately preceding word on the processing of the current word over and above the effect of predictability. The interaction suggests that this effect is greatest when words are relatively less predictable as measured by cloze probability. Further, these effects of semantic relatedness are transient and, once accounting for contextual probability, can only be seen to the immediately preceding content word, suggesting that if preactivation due to semantic relatedness plays a role in comprehending natural

⁸ In order to maximize power to detect an effect of two content words back, we fit one additional post-hoc model which removed the model terms for three and four words back. We observed the same pattern in this post-hoc model, with no effect of the semantic relatedness of two content words back ($p > .2$).

contexts, that effect is very short-lived. When words are highly predictable from context, this effect of semantic relatedness is further diminished.

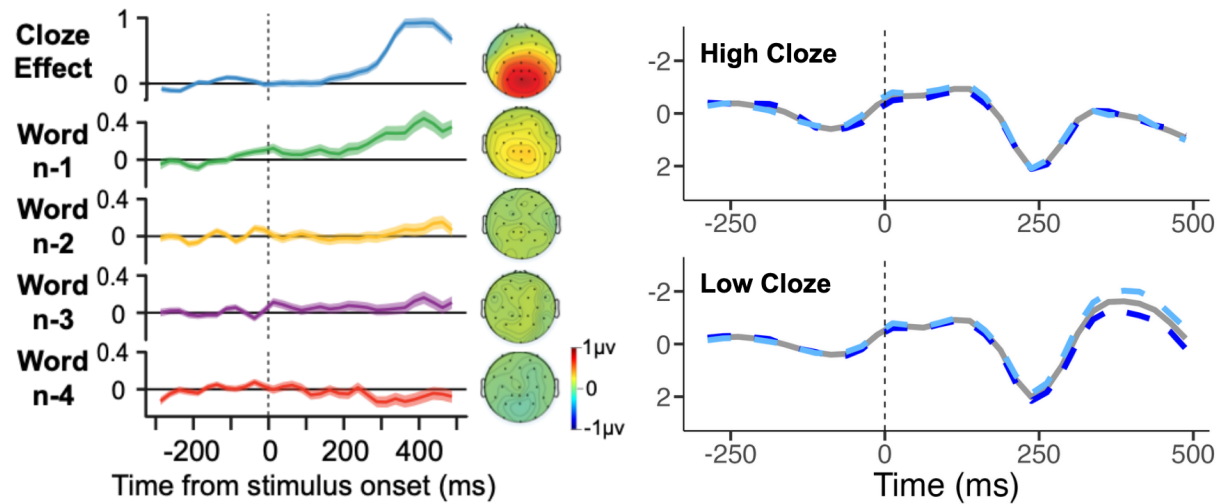


Figure 8. Left: Model-estimated partial effects (beta weights) for the effect of cloze probability and semantic relatedness with preceding content words over time. The semantically associated words that immediately preceded the content word (n-1) reduced the amplitude of the N400 response, but there were no significant effects of semantic relatedness for earlier sentence positions. Right: Averaged N400 amplitudes to content words in our spatial region of interest (P3/z/4, C3/z/4, CP1/2) plotted with trials with above median values for cloze probability (top right), and only trials with below median values for cloze probability (bottom right). The gray line represents the grand average ERP within this spatial region of interest. To compute the effect of semantic relatedness, ERP data was binned temporally into 25ms time windows, and model 1 was fit individually at each time point. At each time point, the dotted line represents the estimated ERP amplitude to words that are 1 standard deviation more semantically related to the preceding content word (dark blue) or 1 standard deviation less semantically related (light blue) than average, based on the beta weights extracted from the models.

General Discussion

In this study, we addressed two related questions about the functional significance of the N400 during naturalistic reading comprehension. These questions concerned effects previously described in the ERP literature: whether the effects of broader context can override the effects on the N400 of two local variables, word frequency and semantic relatedness. Specifically, we asked if these interactions would replicate in naturalistic contexts, where evidence from eye-tracking has been mixed. Below, we discuss each of these findings in more detail, and discuss future directions for naturalistic studies of the N400.

1. Interaction of frequency and predictability

We compared the effects of frequency and lexical predictability from cloze probability on the N400 during naturalistic reading. We found that the effect of frequency on the N400 is diminished to words that are highly predictable from context, observing independent and interacting effects of frequency and lexical predictability. Additionally, we found a main effect of frequency only in a subset of words which were not predictable from context, but no main effect of frequency in a subset of words that was highly predictable.

This is consistent with results found in non-naturalistic ERP studies. For example, words appearing later in short, single congruent sentences (which were presumed to be more predictable) have been shown to have smaller effects of frequency (Van Petten & Kutas, 1991; Payne, Lee & Federmeier, 2015). As noted in the introduction, this interaction has also been replicated using cloze probability in isolated sentences (Dambacher, Kliegl, Hofmann, & Jacobs, 2006), and in MEG using model-derived cloze probability in single sentences (Huizeling, Arana, Hagoort, & Schoffelen, 2022).

Our replication of the interaction between frequency and lexical predictability is notable because it represents the first time these have been shown to interact on the N400 in healthy adults using cloze probability in naturalistic paragraphs. This is particularly relevant, given that there is ample evidence from studies using naturalistic language and eye-tracking measures that these two factors do not interact.

In a metareview, Staub (2015; table 2) reports results from 11 studies showing effects of lexical frequency and predictability on gaze duration of words, which is shorter to both highly predictable and high frequency words. While one of these studies does not report whether an interaction between these factors exists and cannot be evaluated, the remaining ten report that there was no significant interaction. Similarly, in two studies using fixation-related potentials (FRPs), the authors failed to replicate interactions between frequency and predictability (Kretzschman, et al., 2015; Degno et al., 2018), and a recent fMRI study using model-derived cloze probability in larger discourse fragments found dissociable effects of both frequency and predictability with no evidence for an interaction (Shain et al., 2023).

It is not immediately clear why this discrepancy exists in the literature. It is possible that the underlying mechanisms responsible for eye movement and for the generation of the N400 operate at different levels of representation within the brain. This possibility is further discussed in Huizeling, Arana, Hagoort, & Schoffelen (2022), who suggest that frequency and lexical predictability have independent influences on early stages of word processing (<150ms), but interact during later stages of word processing between 150 and 250ms. This discrepancy between the ERP and eye-tracking literature underscores the importance of replicating effects using different techniques, which may not be sensitive to the same cognitive processes.

2. Effects of semantic relatedness and predictability

2.1. Interactive effects of Semantic relatedness and predictability

We found that semantic relatedness with the immediately preceding word attenuates the N400 response to words over and above the effect of cloze probability based on all prior context. Moreover, the degree to which the N400 was attenuated by semantic relatedness was greater when words were relatively less predictable from context, as indicated by the observed interaction between semantic relatedness and cloze probability.

While others have reported that the effects of word predictability and semantic relatedness on the N400 have similar spatial distributions and time windows (e.g. Van Petten, Weckerly, McIsaac & Kutas,

1997; Federmeier, Van Petten, Schwartz & Kutas, 2003), this study is the first to our knowledge to separate out the simultaneous influence of graded cloze probability and semantic relatedness on the N400 in a single study, and the first to test these effects during naturalistic reading. By including both factors independently in our models, we are able to better understand how they work together to influence the N400.

The dissociation we observed between semantic relatedness and global predictability is mirrored in other paradigms. For example, when stimuli are presented specifically in the left or right visual field, which is thought to partially localize processing of the stimulus to a single hemisphere, lexical association attenuates the N400 only when words are presented in the right visual hemifield, whereas effects of congruence can be seen to words in either hemifield (Coulson, Federmeier, & Kutas, 2005). Comparing groups of participants has also provided evidence for dissociation. N400s in participants with high working memory spans, for example, are relatively more affected by congruence than by lexical association (Van Petten, Weckerly, McIsaac & Kutas, 1997; Boudewyn, Long, & Swaab, 2013). A similar pattern has been observed in patients with schizophrenia, with patient N400s exhibiting greater sensitivity to lexical association (Swaab et al., 2013).

However, not all studies have found separable effects of predictability and semantic relatedness. In their ERP findings, Frank & Willems (2017)¹⁰ report that they cannot distinguish between predictability and semantic relatedness in their ERP results, as each have similar topography, “nearly identical onset and offset, and... effect sizes.” However, it is important to note that this paper used a relatively weak measure of predictability (n-gram models) and a measure of semantic relatedness which compared each word to its full prior context, rather than pairwise comparisons between individual words.

¹⁰ Of note, Frank & Willems (2017) also report evidence from fMRI that distinct brain regions are sensitive to semantic relatedness and predictability. Unfortunately, due to the temporal imprecision of fMRI recording, and the use of n-gram models to assess lexical predictability, it is difficult to interpret these findings.

2.2. Interactive effects of Semantic relatedness and predictability

In addition to finding that semantic relatedness and predictability exerted independent and interacting effects on the N400, we found that the effect of semantic relatedness on the N400 is short-lived. After accounting for contextual probability, only semantic relatedness with the immediately preceding content word affects the N400 to the current word. Despite a relatively quick average word-reading time of 385ms and robust effects of semantic relatedness for the previous content word, semantic relatedness with content words appearing two, three, or four positions further back in the stimulus did not have any effect on the N400 in our model.

The extreme sensitivity to word distance we observed in our study, where effects of semantic relatedness are only observed when comparing words to their immediately preceding content word, is supported by previous findings in traditional ERP paradigms. In two studies that manipulated the distance between lexical associates and downstream critical words (Van Petten, Weckerly, McIsaac & Kutas, 1997; Federmeier, Van Petten, Schwartz, & Kutas, 2003), effects of semantic relatedness were observed only when lexical associates were close to the measured word, but not when it was further away within the sentence (e.g. "She fixed the sticky drawer so that it *opened* and *closed* easily." vs "After taking his wallet, they waved a *gun* and threatened to *shoot* him if he reported it."). In this case, as in our results, even the presence of a single intervening content word seems to eliminate the effect of lexical association. Conversely, the effect of discourse congruence was unaffected by this manipulation of lexical associate distance.

This suggests that the residual semantic activation from single words is fragile during natural reading, with the exception of the immediately preceding content word. Semantically related context that is separated by more than one content word likely does not influence lexico-semantic access directly by semantic relationships between pairs of words; rather, lexico-semantic access is facilitated in these cases because semantically related context is associated with enhanced lexical predictability.

3. Hierarchical generative models and predictive coding can explain interactions between local and contextual variables

While frequency and semantic relatedness affect our lexico-semantic access in different ways, we found that both of their effects on the N400 could be at least partially overridden by predictability from prior context, and that each exert smaller influence when words are highly predictable from context. We believe that both of these interactions with predictability can be understood by viewing language comprehension as a hierarchical generative process, where comprehenders update an internal model consisting of multiple levels of representation (see Kuperberg & Jaeger, 2015 for further discussion). In this framework, predictions are generated at the highest, most abstract layer, which represents events and event sequences, and feed down progressively through layers to preactivate concepts and features relevant to the predicted event, where they interact with bottom-up information coming from the perceived stimulus.

Predictive coding is a biologically plausible mechanism describing how hierarchical generative models could be approximated within the brain. In predictive coding, which was first described in the visual system in Rao & Ballard (1999), these predictions take the form of ‘reconstructions’ — that is, each layer will attempt to predict the activation pattern of the layer below it. The difference between these reconstructions and the bottom-up input is termed the ‘prediction error’, which represents the new information from the bottom-up input that has not already been predicted (Rabovsky & McRae, 2014; Fitz & Chang, 2019). If these reconstructions are correct, then the prediction error is low, and the new bottom-up input does not encode any new information that has not already been preactivated at the layer above. If the bottom-up information does not match the reconstructions, the resulting prediction error is large.

More recently, we implemented a predictive coding model of language comprehension (for detailed description and discussion, see Nour Eddine et al., 2024). In this implementation, we consider the size of the prediction error at the lexico-semantic level of representation to correspond to the magnitude

of the N400. By using lexico-semantic prediction error as an index, we found that not only was this model able to explain the effects of predictability on the N400, but that it could additionally explain why the effect of frequency is smaller for highly predictable words.

In a predictive coding model, word frequency is ‘baked in’ as the prior probability in the default reconstruction. Top-down connections relating to more frequent words have stronger weights, and thus have a larger influence on reconstructions (see also Norris, 2006; Spratling, 2016; Delaney-Busch, Morgan, Lau & Kuperberg, 2019, which treat frequency as a prior belief for a particular lexical item in a Bayesian framework). When word predictability from context is low, this leads to higher frequency words receiving greater preactivation than low frequency words. When word predictability is high following constraining contexts, reconstructions override these default patterns of frequency, and the effect of frequency is diminished (Nour Eddine et al., 2024).

This pattern is exactly what we observed with frequency in this study, where the effect of frequency is diminished when words are highly predictable. Additionally, this framework can explain why the effect of frequency is lower at the ends of congruent sentences (Van Petten & Kutas, 1991; Payne, Lee & Federmeier, 2015; Payne & Federmeier, 2018), when words are repeated (Rugg, 1990), as these cases also allow for greater preactivation based on top-down influences.

While we have also simulated the effects of semantic relatedness in predictive coding (Nour Eddine et al., 2024), we have so far only simulated this effect on pairs of words in isolation, and not on words embedded in larger contexts. However, we believe that semantic relatedness may behave on similar principles: that is, the advantage of preactivating lexico-semantic features via any mechanism, whether from semantic priming or higher frequency, may become less prominent when contexts are highly constraining.

4. Future directions

We believe that ERP studies using naturalistic language have the potential to answer several open questions in the language comprehension literature.

One major open question concerns how the N400 changes over the course of sentences, over and above the impact of lexical predictability. As noted earlier, it has been well established that the N400 is progressively attenuated over the course of individual, congruent sentences (Van Petten & Kutas, 1991; Payne, Lee & Federmeier, 2015). This is generally attributed to the gradual accumulation of contextual constraints at various levels of representation over the course of a sentence, which leads later words to be more predictable.

However, it is possible that there are additional mechanisms other than lexical probability that are affecting the ease of lexico-semantic access at different parts of sentences. For example, because of the importance of these early words for establishing structure, readers may choose to devote more attentional resources to these words as described in Gernsbacher's structure-building framework (Gernsbacher, 1990).

Another open question concerns the effect of word position within paragraphs. In particular, while much is now understood about the dynamics of the N400 within individual sentences, no study has asked what happens to the N400 when readers transition between sentences within a single, larger discourse.

One possibility is that the N400 'resets' at sentence boundaries. This could occur due to difficulty bringing information between sentences, due to a corresponding decrease in contextual predictability, or due to changes in attentional engagement. A second possibility is that the amplitude of the N400 could continue to decrease monotonically beyond the first sentence of a connected paragraph, with only a limited rebound between sentences. Finally, it is possible that there is a limit to the degree to which lexico-semantic access can be facilitated in natural contexts. If this limit is reached by the end of a sentence, the size of the N400 may plateau after reaching the end of the first sentence and continue in the next sentence with a steady diminished amplitude.

Chapter 5: General Discussion

In this thesis, we report findings from three ERP studies examining two ERP components that have been influential in the language comprehension literature: the P600 and the N400. In Chapter 2, we used tasks which either encouraged or prohibited error correction, in order to clarify the relationship between the P600 and error processing. In Chapter 3, presented anomalies in varied contexts, and suggested that the error processing indexed by the P600 is directly relevant to task-free reading comprehension. In Chapter 4, we focused on an earlier ERP component, the N400, to every word in a maximally naturalistic task environment and found that highly constraining contexts can reduce the impact of frequency and semantic relatedness on the N400.

In this chapter, we first summarize and synthesize these results. Then, we discuss how these results relate to situation models, and why it is critically important that we understand when experimental paradigms encourage or discourage their establishment.

Summary of findings

In Chapter 2, we sought to determine whether the P600 seen after linguistic anomalies reflects some form of error correction, as is assumed in most accounts. We asked participants to read and make plausibility judgments about short, semantically attracted sentences which could be plausible or anomalous. Participants would then either mentally correct any errors and speak the corrected versions aloud (Correction task) or repeat sentences aloud verbatim (Repetition task). In both cases, participants had to judge the plausibility of the sentences. However, participants only engaged in error correction in the Correction task, as the need to quickly repeat sentences in the Repetition task blocked attempts at error correction.

We found that the P600 effect was unaffected by the task demands (Repetition vs. Correction). We therefore concluded that the P600 does not index the process of error correction, contrasting with previous theories (e.g. noisy channel; Levy, 2008; Levy, 2011; Gibson, Bergen & Piantadosi, 2013; Ryskin et al.,

2021). Rather, we argued that the P600 is more likely linked to reprocessing, possibly associated with a prolonged decision making process.

Importantly, however, while the task manipulation in Chapter 2 allowed us to influence whether participants engaged in error correction, the task demands and stimuli were artificial. It was therefore difficult to generalize these findings to understand the P600's relevance with regard to natural reading.

In Chapter 3, we discussed the inherent tension between examining language comprehension using tightly controlled tasks and stimuli, and using stimuli and tasks that more closely resemble naturalistic environments. We directly compared paradigms at the extremes: single sentences with a passive comprehension task (Experiment 1a), single sentences with plausibility judgments (Experiment 1b), and multi-sentence discourse fragments with engaging comprehension tasks related to naturalistic reading (Experiment 2).

We observed robust P600 effects in both Experiment 1b and Experiment 2. Further, naturalistic contexts and tasks appeared to enhance the amplitude of the P600, as P600 effects were larger in Experiment 2 compared to Experiment 1a. Together, we took these findings as clear evidence that the P600 response to linguistic errors is not simply an artifact of artificial task requirements. Instead, we suggested that the P600 effect was enhanced in Experiment 1b and Experiment 2 because, in comparison to Experiment 1a, they both required participants to engage in deeper comprehension.

Finally, in Chapter 4, we used these same naturalistic contexts and tasks to answer questions about an earlier ERP component, the N400, whose amplitude is inversely related to the predictability of a word given its prior context. While the N400 has a longer history in the psycholinguistic literature, its behavior specifically within longer naturalistic contexts is still poorly understood. In this chapter, we again used the naturalistic contexts and tasks from Chapter 3, Experiment 2, and collected measures of lexical predictability for every word. We then assessed whether the effects of frequency and semantic relatedness on the N400 would be diminished when words are highly predictable from context in naturalistic contexts by fitting a series of linear mixed effects models to assess the individual impact of each factor.

We found that effects of two surface-level factors, frequency and priming from semantically related words, were indeed reduced when words were highly predictable from global context. Moreover, we found that the effect of semantic relatedness on the N400 is short-lived, and that after accounting for lexical predictability, only semantic relatedness with the immediately preceding content word affects the N400 to the current word, and semantic relatedness with content words appearing two, three, or four positions further back in the stimulus did not have any effect on the N400.

Broader implications

One general theme that emerges from this work is that language comprehension goes beyond constructing a simple surface structure. Rather, successful understanding requires the comprehender to establish a high-level representation of the events, actions, and objects described in prior context, known in the reading comprehension literature as a situation model (Van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). This model is held in working memory and interacts with new language input as it arrives.

There are many factors which can encourage or discourage the establishment of a situation model in a given experimental paradigm (see Kuperberg, 2007 for review). These factors include embedding critical words within larger contexts, requiring participants to make plausibility judgments about the stimuli, and using critical words which are semantically attracted to preceding context (see Kuperberg, 2016 for discussion). When participants are more likely to have established a situation model, P600 effects are more reliably seen to anomalies (Kuperberg, 2007, Brothers, Wlotko, Warnke, & Kuperberg, 2020).

In Chapter 2, we used very short sentence contexts, which would ordinarily discourage the establishment of a situation model, and result in no P600 effects in response to anomalies. However, we did observe P600 effects in response to anomalies. In our discussion, we argued that strong semantic attraction between individual words within these contexts (e.g. “dark-caves-exploring” or “judge-gavel-pardon”) was sufficient to establish this type of model and motivate comprehenders to reprocess the input,

as semantically attracted words provided highly reliable cues for the retrieval of a canonical schema stored in long-term memory.

In Chapters 3 and 4, we used much longer discourse contexts, as well as comprehension questions and other tasks to encourage the establishment of a situation model. In Chapter 3 specifically, we showed that this made a critical difference to the amplitude and timecourse of the P600 as an index of prolonged error-based processing. Compared to single-sentence stimuli (Experiment 1a), anomalies embedded within longer stimuli (Experiment 2) resulted in P600 effects that were both larger in amplitude and longer lasting. P600s to anomalies in single sentences were also impacted by the addition of a plausibility judgment (Experiment 1b), but this effect appeared to diminish more rapidly than to the anomalies in Experiment 2.

In Chapter 4, we showed that top-down predictions from this higher-level situation model could override both the effect of frequency and the effect of simple semantic priming on another ERP component, the N400. As we discuss in that chapter, this pattern of results suggests that the comprehender's abstract mental representation of the context allows them to preactivate relevant lexicosemantic features ahead of encountering new input.

This has important general implications for understanding why some readers may struggle with reading comprehension, even despite understanding the individual words and surface structure (Cain, 2016). Such readers often fail to consciously detect even very clear global inconsistencies in text (Markman, 1979; Garner, 1980; Wagoner, 1983), suggesting that they are either not establishing or not maintaining robust situation models during reading.

These problems with deep comprehension with language are common and can have severe impacts on developing readers. Problems with reading comprehension are among the most frequently reported deficits in children with learning disabilities (Lyon, 1995; Vaughn, Levy, Coleman, & Bos, 2002). Younger readers will sometimes struggle with reading comprehension despite full understanding of the individual words (Deficits in the use of global context also affect populations with certain clinical

disorders such as schizophrenia (e.g. Kuperberg, McGuire, & David, 1998; Sitnikova, Salisbury, Kuperberg, & Holcomb, 2002).

While there are myriad mechanisms which may go awry in these cases, two are of particular relevance here. First, it is possible that poor comprehenders lack the ability to monitor for and detect errors and inconsistencies. As discussed above, this can only happen when readers are successfully establishing situation models as they read. If an error is encountered before a situation model can be established, we do not see P600 effects, and can consequently conclude that they are not engaging in reprocessing, and not accumulating additional evidence to determine if observed input is truly anomalous.

This inability to actively monitor comprehension during reading manifests in the behavior of many with reading difficulties. Poor readers are known to be less accurate at detecting linguistic anomalies (Garner, 1980; Hacker, 1997; Oakhill & Cain, 2012; Rubman & Salatas Waters, 2000), and less likely to reread words when comprehension difficulties arise (Ehrlich, Remond, & Tardieu, 1999; Hacker, Dunlosky, & Graesser, 1998). When readers cannot make regressive saccades, their comprehension performance suffers, particularly in sentences that require semantic or syntactic re-analysis (Schotter, Tran & Rayner, 2014; Metzner, von der Malsburg, Vasishth & Rosler, 2017).

Second, it is possible that poor comprehenders are suffering from a lack of ability to generate predictions about future input. In Chapter 4, we demonstrate that readers do generate predictions from the global context of multi-sentence naturalistic discourse fragments, and that using these predictions eases lexico-semantic access. Moreover, while highly frequent and semantically primed words were easier to access, this effect was reduced when global context was highly constraining. However, it is possible that readers vary in the degree to which they rely on these predictions from global context compared to surface level features such as semantic priming, possibly due to an inability to generate strong predictions online.

In order to better understand how to help those struggling with reading comprehension, it is critical that our experimental paradigms do not ignore the factors which might encourage or discourage the establishment of a situation model.

References

- Ainsworth-Darnell, K., Shulman, H. G., & Boland, J. E. (1998). Dissociating brain responses to syntactic and semantic anomalies: Evidence from event-related potentials. *Journal of Memory and Language*, 38(1), 112-130. doi: 10.1006/Jmla.1997.2537
- Altmann, G. T. (1998). Ambiguity in sentence processing. *Trends in Cognitive Sciences*, 2, 146-152.
- Audacity Team. (2019). Audacity(R): Free Audio Editor and Recorder (Version 2.32). Retrieved from <https://audacityteam.org/>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255-278. doi: 10.1016/j.jml.2012.11.001
- Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi: 10.18637/jss.v067.i01
- Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology*, 60, 343-355.
- Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience*, 35(8), 3478-3484. doi: 10.1523/JNEUROSCI.0797-14.2015
- Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2013). Effects of working memory span on processing of lexical associations and congruence in spoken discourse. *Frontiers in Psychology*, 4, 60. doi: 10.3389/fpsyg.2013.00060
- 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.
- Brothers, T., Greene, S., & Kuperberg, G. R. (2020). *Distinct neural signatures of semantic retrieval and event updating during discourse comprehension*. Paper presented at the 27th Annual Meeting of the Cognitive Neuroscience Society, Boston, MA.
- Brothers, T., Wlotko, E. W., Warnke, L., & Kuperberg, G. R. (2020). Going the extra mile: Effects of discourse context on two late positivities during language comprehension. *Neurobiology of Language*, 1(1), 135-160. doi: 10.1162/nol_a_00006
- Brothers, T., Zeitlin, M., Choi Perrachione, A., Choi, C., & Kuperberg, G. (2021). Domain-general conflict monitoring predicts neural and behavioral indices of linguistic error processing during reading comprehension. *Journal of Experimental Psychology: General*. doi: 10.1037/xge0001130
- Brouwer, H., & Crocker, M. W. (2017). On the proper treatment of the N400 and P600 in language comprehension. *Frontiers in Psychology*, 8, 1327. doi: 10.3389/fpsyg.2017.01327
- Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41, 977-990.
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904-911. doi: 10.3758/s13428-013-0403-5
- Cain, K. (2016). Reading comprehension development and difficulties: an overview. *Perspectives on Language and Literacy*, 42(2), 9.
- Chow, W. Y., & Phillips, C. (2013). No semantic illusions in the "Semantic P600" phenomenon: ERP evidence from Mandarin Chinese. *Brain Research*, 1506, 76-93. doi: 10.1016/j.brainres.2013.02.016
- Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42, 368-407.
- Coulson, S., Federmeier, K. D., Van Petten, C., & Kutas, M. (2005). Right hemisphere sensitivity to word- and sentence-level context: evidence from event-related brain potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(1), 129-147. doi: 10.1037/0278-7393.31.1.129

- Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event-related brain responses to morphosyntactic violations. *Language and Cognitive Processes*, 13(1), 21-58. doi: 10.1080/016909698386582
- Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, 1084(1), 89-103. doi: 10.1016/j.brainres.2006.02.010
- 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(June 2019), 10-20. doi: 10.1016/j.cognition.2019.01.001
- DeLong, K. A., Chan, W. H., & Kutas, M. (2019). Similar time courses for word form and meaning preactivation during sentence comprehension. *Psychophysiology*, 56(4), e13312. doi: 10.1111/psyp.13312
- DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, 61C, 150-162. doi: 10.1016/j.neuropsychologia.2014.06.016
- 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. doi: 10.1038/nn1504
- Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods*, 134(1), 9-21. doi: 10.1016/J.jneumeth.2003.10.009
- Desender, K., Murphy, P., Boldt, A., Verguts, T., & Yeung, N. (2019). A post-decisional neural marker of confidence predicts information-seeking in decision-making. *Journal of Neuroscience*, 39(17), 3309-3319. doi: 10.1523/JNEUROSCI.2620-18.2019
- Desender, K., Ridderinkhof, K. R., & Murphy, P. (2021). Understanding neural signals of post-decisional performance monitoring: An integrative review. *Elife*, 10. doi: 10.7554/eLife.67556
- Dien, J., Michelson, C. A., & Franklin, M. S. (2010). Separating the visual sentence N400 effect from the P400 sequential expectancy effect: cognitive and neuroanatomical implications. *Brain Research*, 1355, 126-140. doi: 10.1016/j.brainres.2010.07.099
- Ditman, T., Holcomb, P. J., & Kuperberg, G. R. (2007). An investigation of concurrent ERP and self-paced reading methodologies. *Psychophysiology*, 44(6), 927-935. doi: 10.1111/j.1469-8986.2007.00593.x
- Ehrlich, M.-F., Remond, M., & Tardieu, H. (1999). Processing of anaphoric devices in young skilled and less skilled comprehenders: Differences in metacognitive monitoring. *Reading and Writing*, 11(1), 29-63. doi: 10.1023/a:1007996502372
- Ericsson, E., Olofsson, J. K., Nordin, S., Rudolfsson, T., & Sandstrom, G. (2008). Is the P600/SPS affected by the richness of semantic content? A linguistic ERP study in Swedish. *Scandinavian Journal of Psychology*, 49(1), 1-9. doi: 10.1111/j.1467-9450.2007.00604.x
- Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41(4), 469-495. doi: 10.1006/Jmla.1999.2660
- Federmeier, K. D., Van Petten, C., Schwartz, T. J., & Kutas, M. (2003). Sounds, words, sentences: age-related changes across levels of language processing. *Psychology and Aging*, 18(4), 858-872. doi: 10.1037/0882-7974.18.4.858
- Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47, 164-203. doi: 10.1016/S0010-0285(03)00005-7
- Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1), 11-15. doi: 10.1111/1467-8721.00158
- Fields, E. C. (2017). Factorial Mass Univariate ERP Toolbox [Computer software]. <https://github.com/ericcfields/FMUT/releases>: Github.

- 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. doi: 10.1111/psyp.13468
- Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, 111, 15-52. doi: 10.1016/j.cogpsych.2019.03.002
- Fodor, J. D., & Inoue, A. (1994). The diagnosis and cure of garden paths. *Journal of Psycholinguistic Research*, 23(5), 407-434. doi: Doi 10.1007/Bf02143947
- Forstmann, B., Ratcliff, R., & Wagenmakers, E. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, 67, 641-666. doi: 10.1146/annurev-psych-122414-033645
- Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally-ambiguous sentences. *Cognitive Psychology*, 14, 178-210.
- Friederici, A., Hahne, A., & Saddy, D. (2002). Distinct neurophysiological patterns reflecting aspects of syntactic complexity and syntactic repair. *Journal of Psycholinguistic Research*, 31, 45-63.
- Friederici, A. D., Pfeifer, E., & Hahne, A. (1993). Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain Research*, 1(3), 183-192.
- Garner, R. (1980). Monitoring of understanding: An investigation of good and poor readers' awareness of induced miscomprehension of text. *Journal of Reading Behavior*, 12(1), 55-63.
- Geyer, A., Holcomb, P., Kuperberg, G. R., & Perlmutter, N. (2006). Plausibility and sentence comprehension. An ERP Study. *Journal of Cognitive Neuroscience, Supplement, Abstract*.
- 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, USA*, 110(20), 8051-8056. doi: 10.1073/pnas.1216438110
- Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3), 371-395. doi: 10.1037/0033-295x.101.3.371
- Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, 24(2), 95-112.
- Gregg, J., & Inhoff, A. W. (2016). Misperception of orthographic neighbors during silent and oral reading. *J Exp Psychol Hum Percept Perform*, 42(6), 799-820. doi: 10.1037/xhp0000193
- Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, 48(12), 1711-1725. doi: 10.1111/j.1469-8986.2011.01273.x
- Gunter, T. C., Friederici, A. D., & Hahne, A. (1999). Brain responses during sentence reading: Visual input affects central processes. *Neuroreport*, 10(15), 3175-3178. doi: 10.1097/00001756-199910190-00009
- Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, 8(4), 439-483. doi: 10.1080/01690969308407585
- Hahne, A., & Friederici, A. D. (2002). Differential task effects on semantic and syntactic processes as revealed by ERPs. *Cognitive Brain Research*, 13(3), 339-356. doi: 10.1016/S0926-6410(01)00127-6
- Hansen, J. C., & Hillyard, S. A. (1980). Endogeneous brain potentials associated with selective auditory attention. *Electroencephalography and Clinical Neurophysiology*, 49(3-4), 277-290. doi: 10.1016/0013-4694(80)90222-9
- Hanulikova, A., van Alphen, P. M., van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, 24(4), 878-887. doi: 10.1162/jocn_a_00103
- Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National*

- Academy of Sciences of the United States of America*, 119(32), e2201968119. doi: 10.1073/pnas.2201968119
- Hoeks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, 19, 59-73.
- Ito, A., Corley, M., Pickering, M., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, 86, 157-171. doi: 10.1016/j.jml.2015.10.007
- Jouravlev, O., Stearns, L., Bergen, L., Eddy, M., Gibson, E., & Fedorenko, E. (2016). Processing temporal presuppositions: an event-related potential study. *Language, Cognition and Neuroscience*, 31(10), 1245-1256.
- Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20(2), 137-194. doi: 10.1016/s0364-0213(99)80005-6
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1), 122-149.
- Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, 15(2), 159-201.
- Kaan, E., & Swaab, T. Y. (2003). Electrophysiological evidence for serial sentence processing: A comparison between non-preferred and ungrammatical continuations. *Cognitive Brain Research*, 17(3), 621-635.
- Kaan, E., & Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of Cognitive Neuroscience*, 15(1), 98-110. doi: 10.1162/089892903321107855
- Kaufman, R. C., & Obler, L. K. (1995). Classification of normal reading error types. In C. K. Leong & R. M. Joshi (Eds.), *Developmental and Acquired Dyslexia* (pp. 149-157). Dordrecht: Springer.
- Kiesel, A., Miller, J., Joligaeur, P., & Brisson, B. (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. *Psychophysiology*, 45(2), 250-274. doi: 10.1111/j.1469-8986.2007.00618.x
- Kim, A., Oines, L., & Miyake, A. (2018). Individual differences in verbal working memory underlie a tradeoff between semantic and structural processing difficulty during language comprehension: An ERP investigation. *J Exp Psychol Learn Mem Cogn*, 44(3), 406-420. doi: 10.1037/xlm0000457
- Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, 52(2), 205-225. doi: 10.1016/J.Jml.2004.10.002
- Kim, A., & Sikos, L. (2011). Conflict and surrender during sentence processing: an ERP study of syntax-semantics interaction. *Brain and Language*, 118(1-2), 15-22. doi: 10.1016/j.bandl.2011.03.002
- Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5), 363-394.
- Kolk, H. H. J., Chwilla, D. J., van Herten, M., & Oor, P. J. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain Lang*, 85(1), 1-36. doi: 10.1016/S0093-934X(02)00548-5
- Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, 12(5), 535-540. doi: 10.1038/nn.2303
- Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146, 23-49. doi: 10.1016/j.brainres.2006.12.063
- Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, 31(5), 602-616. doi: 10.1080/23273798.2015.1130233
- Kuperberg, G. R., Alexander, T., & Brothers, T. (in preparation). Decision making in deep comprehension: Why both task and discourse context enhance the P600 to linguistic errors.

- Kuperberg, G. R., Brothers, T., & Wlotko, E. (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. doi: 10.1162/jocn_a_01465
- Kuperberg, G. R., Caplan, D., Sitnikova, T., Eddy, M., & Holcomb, P. J. (2006). Neural correlates of processing syntactic, semantic, and thematic relationships in sentences. *Language and Cognitive Processes*, 21(5), 489-530. doi: 10.1080/01690960500094279
- Kuperberg, G. R., Holcomb, P. J., Sitnikova, T., Greve, D., Dale, A. M., & Caplan, D. (2003). Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies. *Journal of Cognitive Neuroscience*, 15(2), 272-293. doi: 10.1162/089892903321208204
- Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1), 32-59. doi: 10.1080/23273798.2015.1102299
- Kuperberg, G. R., Kreher, D. A., Sitnikova, T., Caplan, D. N., & Holcomb, P. J. (2007). The role of animacy and thematic relationships in processing active English sentences: evidence from event-related potentials. *Brain Lang*, 100(3), 223-237. doi: 10.1016/j.bandl.2005.12.006
- Kuperberg, G. R., Lau, E. F., & Clegg, L. (2011). *A Gratton effect on the syntactic P600: Evidence that syntactic processing is subject to a dynamic adjustment of executive control*. Paper presented at the 24th Annual CUNY Conference on Human Sentence Processing.
- Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, 17(1), 117-129. doi: 10.1016/S0926-6410(03)00086-7
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62, 621-647. doi: 10.1146/annurev.psych.093008.131123
- Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947), 161-163. doi: 10.1038/307161a0
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211-240. doi: 10.1037/0033-295x.104.2.211
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2-3), 259-284. doi: 10.1080/01638539809545028
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8), 707-710.
- Levy, R. (2008). *A noisy-channel model of rational human sentence comprehension under uncertain input*. Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08), Edinburgh, UK.
- Levy, R. (2011). *Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results*. Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT '11), Portland, OR.
- Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences, USA*, 106(50), 21086-21090. doi: 10.1073/pnas.0907664106
- Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, 8, 213. doi: 10.3389/fnhum.2014.00213
- Luck, S. J. (2014). *Online Chapter 13: The Mass Univariate Approach and Permutation Statistics (An introduction to the event-related potential technique)* (2nd ed.). MIT Press Online: MIT press.
- MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4), 676-703. doi: 10.1037/0033-295X.101.4.676
- Markman, E. M. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. *Child Development*, 50(3). doi: 10.2307/1128929

- Metzner, P., von der Malsburg, T., Vasishth, S., & Rosler, F. (2015). Brain responses to world knowledge violations: a comparison of stimulus- and fixation-triggered event-related potentials and neural oscillations. *Journal of Cognitive Neuroscience*, 27(5), 1017-1028. doi: 10.1162/jocn_a_00731
- Metzner, P., von der Malsburg, T., Vasishth, S., & Rosler, F. (2017). The importance of reading naturally: Evidence from combined recordings of eye movements and electric brain potentials. *Cognitive Science*, 41 Suppl 6, 1232-1263. doi: 10.1111/cogs.12384
- Michaelov, J. A., Coulson, S., & Bergen, B. K. (2021). So Cloze yet so Far: N400 amplitude is better predicted by distributional information than human predictability judgements. *arXiv preprint arXiv:2109.01226*.
- Munte, T. F., Heinze, H. J., Matzke, M., Wieringa, B. M., & Johannes, S. (1998). Brain potentials and syntactic violations revisited: No evidence for specificity of the syntactic positive shift. *Neuropsychologia*, 36(3), 217-226.
- Munte, T. F., Matzke, M., & Johannes, S. (1997). Brain activity associated with syntactic incongruities in words and pseudo-words. *Journal of Cognitive Neuroscience*, 9, 318-329.
- Murphy, P. R., Robertson, I. H., Harty, S., & O'Connell, R. G. (2015). Neural evidence accumulation persists after choice to inform metacognitive judgments. *Elife*, 4. doi: 10.7554/eLife.11946
- Nakatani, C. H., & Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *J Acoust Soc Am*, 95(3), 1603-1616. doi: 10.1121/1.408547
- Nieuwland, M. S., & Van Berkum, J. J. A. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, 24(3), 691-701.
- Nooteboom, S. G. (1980). Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In V. A. Fromkin (Ed.), *Errors in Linguistic Performance : Slips of the Tongue, Ear, Pen and Hand* (pp. 87–95). London: Academic Press.
- Norris, D. (2006). The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, 113(2), 327-357. doi: 10.1037/0033-295X.113.2.327
- Nour Eddine, S., Brothers, T., Wang, L., Spratling, M., & Kuperberg, G. R. (2024). A predictive coding model of the N400. *Cognition*, 246, 105755. doi: 10.1016/j.cognition.2024.105755
- Oakhill, J. V., & Cain, K. (2012). The precursors of reading ability in young readers: Evidence from a four-year longitudinal study. *Scientific studies of reading*, 16(2), 91-121.
- Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9(1), 97-113. doi: 10.1016/0028-3932(71)90067-4
- Osterhout, L., Allen, M. D., McLaughlin, J., & Inoue, K. (2002). Brain potentials elicited by prose-embedded linguistic anomalies. *Memory & Cognition*, 30(8), 1304-1312. doi: 10.3758/bf03213412
- Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6), 785-806. doi: 10.1016/0749-596x(92)90039-Z
- Osterhout, L., & Holcomb, P. J. (1993). Event-related potentials and syntactic anomaly: Evidence of anomaly detection during the perception of continuous speech. In S. M. Garnsey (Ed.), *Language and Cognitive Processes. Special Issue: Event-related brain potentials in the study of language* (Vol. 8 (4), pp. 413-437). Hove: Lawrence Erlbaum Associates.
- Osterhout, L., Kim, A., & Kuperberg, G. R. (2012). The neurobiology of sentence comprehension. In M. Spivey, M. Joannisse & K. McRae (Eds.), *The Cambridge Handbook of Psycholinguistics* (pp. 365-389). Cambridge: Cambridge University Press.
- Osterhout, L., & Nicol, J. (1999). On the distinctiveness, independence and time course of the brain responses to syntactic and semantic anomalies. *Language and Cognitive Processes*, 14, 283-317.
- Paape, D., & Vasishth, S. (2022). Is reanalysis selective when regressions are consciously controlled? *Glossa Psycholinguistics*, 1(1). doi: 10.5070/g601139
- Paczynski, M., & Kuperberg, G. R. (2011). Electrophysiological evidence for use of the animacy hierarchy, but not thematic role assignment, during verb argument processing. *Language and Cognitive Processes*, 26(9), 1402-1456. doi: 10.1080/01690965.2011.580143

- Paczynski, M., & Kuperberg, G. R. (2012). Multiple influences of semantic memory on sentence processing: Distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. *Journal of Memory and Language*, 67(4), 426-448. doi: 10.1016/j.jml.2012.07.003
- Payne, B. R., Lee, C. L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology*, 52(11), 1456-1469. doi: 10.1111/psyp.12515
- Peirce, J. W. (2007). PsychoPy--Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1-2), 8-13. doi: 10.1016/j.jneumeth.2006.11.017
- Pijnacker, J., Geurts, B., van Lambalgen, M., Buitelaar, J., & Hagoort, P. (2010). Exceptions and anomalies: an ERP study on context sensitivity in autism. *Neuropsychologia*, 48(10), 2940-2951. doi: 10.1016/j.neuropsychologia.2010.06.003
- Quante, L., Bolte, J., & Zwitserlood, P. (2018). Dissociating predictability, plausibility and possibility of sentence continuations in reading: evidence from late-positivity ERPs. *PeerJ*, 6, e5717. doi: 10.7717/peerj.5717
- Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition*, 132(1), 68-89. doi: 10.1016/j.cognition.2014.03.010
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2), 59-108. doi: 10.1037/0033-295x.85.2.59
- Royle, P., Drury, J. E., & Steinhauer, K. (2013). ERPs and task effects in the auditory processing of gender agreement and semantics in French. *The Mental Lexicon*, 8(2), 216-244. doi: 10.1075/ml.8.2.05roy
- Rubman, C. N., & Salatas Waters, H. (2000). A, B seeing: The role of constructive processes in children's comprehension monitoring. *J Educ Psychol*, 92(3), 503-514. doi: 10.1037//0022-0663.92.3.503
- Rugg, M. D. (1985). The effects of semantic priming and word repetition on event-related potentials. *Psychophysiology*, 22, 642-647.
- Rugg, M. D. (1990). Event-related potentials dissociate repetition effects of high and low frequency words. *Memory and Cognition*, 18, 367-379.
- Ryskin, R., Stearns, L., Bergen, L., Eddy, M., Fedorenko, E., & Gibson, E. (2021). An ERP index of real-time error correction within a noisy-channel framework of human communication. *Neuropsychologia*, 158, 107855. doi:10.1016/j.neuropsychologia.2021.107855
- Sanford, A. J., Leuthold, H., Bohan, J., & Sanford, A. J. S. (2011). Anomalies at the borderline of awareness: an ERP study. *Journal of Cognitive Neuroscience*, 23, 514-523.
- Sassenhagen, J., & Fiebach, C. J. (2019). Finding the P3 in the P600: Decoding shared neural mechanisms of responses to syntactic violations and oddball targets. *Neuroimage*, 200, 425-436. doi: 10.1016/j.neuroimage.2019.06.048
- Sassenhagen, J., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2014). The P600-as-P3 hypothesis revisited: single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain and Language*, 137, 29-39. doi: 10.1016/j.bandl.2014.07.010
- Schacht, A., Sommer, W., Shmulevich, O., Martienz, P. C., & Martin-Loeches, M. (2014). Differential task effects on N400 and P600 elicited by semantic and syntactic violations. *PloS One*, 9(3), e91226. doi: 10.1371/journal.pone.0091226
- Shetreet, E., Alexander, E. J., Romoli, J., Chierchia, G., & Kuperberg, G. R. (2019). What we know about knowing: Presuppositions generated by factive verbs influence downstream neural processing. *Cognition*, 184, 96-106. doi: 10.1016/j.cognition.2018.11.012
- Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(6), 1521-1543. doi: 10.1037/0278-7393.24.6.1521

- Spratling, M. W. (2016). A neural implementation of Bayesian inference based on predictive coding. *Connection Science*, 28(4), 346-383. doi: 10.1080/09540091.2016.1243655
- Szewczyk, J. M., & Schriefers, H. (2011). Is animacy special? ERP correlates of semantic violations and animacy violations in sentence processing. *Brain Res*, 1368, 208-221. doi: 10.1016/j.brainres.2010.10.070
- Van Berkum, J. J. A., Brown, C. M., & Hagoort, P. (1999). Early referential context effects in sentence processing: Evidence from event-related brain potentials. *Journal of Memory and Language*, 41, 147-182.
- Van Berkum, J. J. A., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of Cognitive Neuroscience*, 11(6), 657-671. doi: 10.1162/089892999563724
- van de Meerendonk, N., Kolk, H. H. J., Chwilla, D. J., & Vissers, C. T. W. M. (2009). Monitoring in language perception. *Language and Linguistics Compass*, 3(5), 1211-1224. doi: 10.1111/j.1749-818X.2009.00163.x
- van de Meerendonk, N., Kolk, H. H. J., Vissers, C. T. W. M., & Chwilla, D. J. (2010). Monitoring language perception: mild and strong conflicts elicit different ERP patterns. *Journal of Cognitive Neuroscience*, 22(1), 67-82. doi: 10.1162/jocn.2008.21170
- van Dijk, T. A., & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.
- van Gompel, R. P. G., Pickering, M. J., & Traxler, M. J. (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language*, 45(2), 225-258. doi: 10.1006/jmla.2001.2773
- van Herten, M., Chwilla, D. J., & Kolk, H. H. (2006). When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *Journal of Cognitive Neuroscience*, 18(7), 1181-1197.
- van Herten, M., Kolk, H. H., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cogn Brain Res*, 22(2), 241-255.
- Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory and Cognition*, 18, 380-393.
- Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Memory and Cognition*, 19(1), 95-112.
- Van Petten, C., Weckerly, J., McIsaac, H. K., & Kutas, M. (1997). Working memory capacity dissociates lexical and sentential context effects. *Psychological Science*, 8(3), 238-242. doi: 10.1111/j.1467-9280.1997.tb00418.x
- Vega-Mendoza, M., Pickering, M. J., & Nieuwland, M. S. (2021). Concurrent use of animacy and event-knowledge during comprehension: Evidence from event-related potentials. *Neuropsychologia*, 152, 107724. doi: 10.1016/j.neuropsychologia.2020.107724
- Vissers, C. T., Chwilla, D. J., & Kolk, H. H. (2006). Monitoring in language perception: The effect of misspellings of words in highly constrained sentences. *Brain Research*, 1106(1), 150-163.
- Wagoner, S. A. (1983). Comprehension monitoring: What it is and what we know about it. *Read Res Q*, 328-346.
- White, K., & Ashton, R. (1976). Handedness assessment inventory. *Neuropsychologia*, 14(2), 261-264.
- Wlotko, E. W., & Federmeier, K. (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex*, 68, 20-32. doi: 10.1016/j.cortex.2015.03.014
- Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, 30(6), 648-672. doi: 10.1080/23273798.2014.995679
- Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: a new measure of orthographic similarity. *Psychon Bull Rev*, 15(5), 971-979. doi: 10.3758/PBR.15.5.971
- Zheng, X., & Lemhofer, K. (2019). The "semantic P600" in second language processing: When syntax conflicts with semantics. *Neuropsychologia*, 127, 131-147. doi: 10.1016/j.neuropsychologia.2019.02.010

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2), 162-185. doi: 10.1037/0033-2909.123.2.162