



Grape Expectations: A collection of EEG stories on form-based prediction in natural language contexts

Citation

Yacovone, Anthony. 2023. Grape Expectations: A collection of EEG stories on form-based prediction in natural language contexts. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

Permanent link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37375749>

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

HARVARD UNIVERSITY
Graduate School of Arts and Sciences



DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Psychology

have examined a dissertation entitled

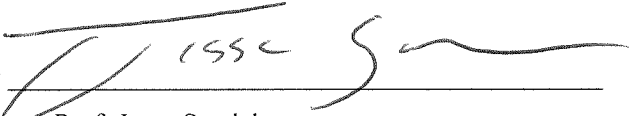
"*Grape Expectations*: A collection of EEG stories on form-based prediction in natural language contexts"

presented by Anthony Yacovone,

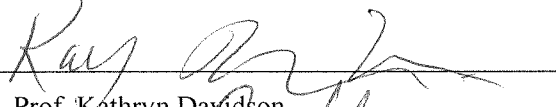
candidate for the degree of Doctor of Philosophy and hereby

certify that it is worthy of acceptance.

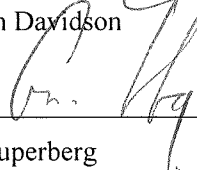
Signature


Typed name: Prof. Jesse Snedeker

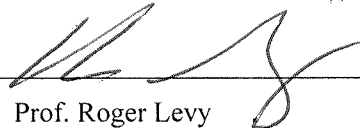
Signature


Typed name: Prof. Kathryn Davidson

Signature


Typed name: Prof. Gina Kuperberg

Signature


Typed name: Prof. Roger Levy

Date: April 11, 2023

***Grape Expectations: A collection of EEG stories on form-based prediction
in natural language contexts***

A dissertation presented

by

ANTHONY YACOVONE

to

THE DEPARTMENT OF PSYCHOLOGY

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in the subject of

PSYCHOLOGY

Harvard University

Cambridge, Massachusetts

April 11, 2023

© 2023 Anthony Yacovone

All rights reserved.

Grape Expectations: A collection of EEG stories on form-based prediction
in natural language contexts

Abstract

It is no longer controversial to say that language comprehension involves prediction. Decades of psycholinguistic research have demonstrated that comprehenders reliably anticipate both the meaning and the form of upcoming words (DeLong et al., 2005; Federmeier & Kutas, 1999; see Kuperberg & Jaeger, 2016; Kutas & Federmeier, 2011). The next frontier in predictive processing research is understanding the mechanisms by which prediction arises and how those mechanisms develop in the world's many languages. To do this, the field must begin to characterize how prediction works in more naturalistic contexts and across a wider range of modalities and populations.

In this dissertation, I present three EEG experiments that aim to understand the nature of predictive processing in ordinary storytelling contexts. In each experiment, we used a novel naturalistic story paradigm in which participants simply comprehend rich, naturally produced narratives that have experimental manipulations injected into them. As participants comprehended these narratives, we recorded their neural responses to a set of manipulated target words, allowing us to assess the nature of their linguistic predictions.

In Paper 1, we investigated prediction in Spanish-English bilinguals while they listened to short stories that were presented in English with occasional Spanish words. We found that bilinguals' predictions were lexically specific, meaning that they had generated expectations for

a particular word in a particular language. This work provided initial evidence that predictions in naturalistic settings are form-based and move beyond just anticipating the gist of upcoming linguistic material (see Yacovone, Moya, & Snedeker, 2021).

In Paper 2, we further assessed the nature of form-based predictions by asking whether English-speaking adults predict the sounds of upcoming words during comprehension. In this experiment, participants watched a cartoon narration of a children’s book, which had a set of manipulations spliced into it. Specifically, we identified highly predictable and unpredictable words, and then replaced them with nonwords that sounded similar or dissimilar to the original word (e.g. *cake*, *ceke*, *vake*). Results indicated smaller neural responses to baseline words (*cake*) and similar nonwords (*ceke*) relative to dissimilar nonwords (*vake*). This reduction in neural responses, however, was only observed for predictable words, replicating the prior findings from similar studies with more tightly controlled experimental designs. Thus, Paper 2 demonstrated that listeners are able to predict the sounds of upcoming words in natural language contexts.

In Paper 3, we explored the nature of form-based prediction in a different modality—namely, sign language. To do this, we had deaf signers of American Sign Language (ASL) watch a narrative in which we had manipulated a set of target signs. Following the logic of Paper 2, we identified both predictable and unpredictable signs, and then replaced them with similar and dissimilar non-signs. We found tentative evidence that signers anticipate the handshape of upcoming signs during naturalistic comprehension, mirroring the findings from Paper 2. Taken together, these studies present a body of work that highlights the importance of understanding prediction in natural language contexts and across different modalities and populations.

Table of Contents

Title page	i
Copyright	ii
Abstract	iii
Table of Contents	v
Acknowledgements	x
Chapter 1	1
1.1. Introduction	1
1.1.1. The adult language comprehension system is incremental, interactive, and predictive ...	2
1.1.2. What is linguistic prediction, and how have studies demonstrated this phenomenon?	4
1.1.2.1. Key findings from studies using reading paradigms.....	5
1.1.2.2. Key findings from studies using the visual world paradigm	7
1.1.2.3. Key findings from studies using electroencephalography	10
1.1.3. Studying linguistic prediction of meaning and form in natural language contexts	22
1.1.4. Summary of the chapters in this dissertation	26
Chapter 2	29
2.1. Introduction	29
2.1.1. The cost of code-switching in bilingual comprehension: Evidence from ERPs.....	30
2.1.1.1. Switch-related negativities and their variability in the literature	31
2.1.1.2. Switch-related LPCs and their variability in the literature	36
2.1.2. Two theories about the costs of code-switching	37
2.1.3. Testing the one-cost and two-cost hypotheses.....	40
2.1.4. The present study	43
2.2. Method	44
2.2.1. Participants.....	44
2.2.2. Stimuli.....	46
2.2.2.1. Oral story selection	47
2.2.2.2. Target sentence and English noun selection	48
2.2.2.3. Weak-fit English noun selection	49
2.2.2.4. Spanish noun selection.....	51
2.2.2.5. Assessing our critical manipulations within our study population	52

2.2.2.6. Other properties of the stimuli: word frequency, length, and sentence position	53
2.2.2.7. Audio stimulus creation	56
2.2.3. Procedure	58
2.2.4. EEG Recording	59
2.2.4.1. EEG Pre-processing	59
2.2.5. Statistical Analyses	60
2.2.5.1. Pre-registered mean amplitude analyses (300–500 ms post-stimulus onset).....	60
2.2.5.2. Exploratory permutation-based cluster mass analyses (0–2000 ms post-stimulus onset).....	61
2.3. Results and Discussion.....	63
2.3.1. Averaged waveforms and topographic voltage maps	63
2.3.2. Mean Amplitude Analysis at Fz, Cz, and Pz (300–500 ms): The N400.....	64
2.3.3. Permutation-based cluster mass analysis across all electrodes.....	67
2.4. General Discussion.....	73
2.4.1. Previous factorial manipulations of contextual fit and language switching	75
2.4.2. Understanding variability in switch-related negativities	81
2.4.2.1. The LAN and N400 as functionally distinct components.....	83
2.4.2.2. The LAN and the N400 as a unitary phenomenon	88
2.4.2.3. The LAN as an epiphenomenon resulting from component overlap.....	89
2.4.3. When should we expect reduced or absent switch-related N400 effects?	92
2.4.4. What does this tell us about the functional significance of the N400?.....	96
2.4.5. What does the present study say about LPC effects?	99
2.4.6. Methodological advantages and limitations to the Storytime paradigm.....	101
2.5. Conclusion	104
Chapter 3	106
3.1. Introduction.....	106
3.1.1. Reviewing the EEG evidence for form-based prediction during comprehension	108
3.1.1.1. The N400 as an effect of lexicosemantic pre-activation during comprehension...	109
3.1.1.2. The posterior P600 as an effect of reprocessing strong violations of expectation.	112
3.1.1.3. Early negativities as evidence for form-based prediction in spoken language contexts	113

3.1.2. Reviewing the typical paradigms in form-based prediction studies and their limitations	117
3.1.3. The present study	119
3.2. Method	121
3.2.1. Participants.....	121
3.2.1.1. Sample size calculation and power analyses for mixed-effects models	121
3.2.2. Stimuli.....	122
3.2.2.1. Selecting our story substrate	123
3.2.2.2. Characterizing the predictability of every word in the story	124
3.2.2.3. Selecting the predictable and unpredictable target words.....	125
3.2.2.4. Creating the nonword violations	126
3.2.2.5. Creating the spliced recordings.....	126
3.2.2.6. Creating the cartoon for the story	127
3.2.2.7. Describing the filler structure and the presentation of our target words.....	128
3.2.3. Procedure	129
3.2.3.1. Experimental set-up	129
3.2.3.2. EEG recording	129
3.2.4. Data pre-processing steps and data exclusion criteria	129
3.2.5. Statistical analyses	130
3.2.5.1. Determining our spatial and temporal regions of interest.....	130
3.2.5.2. Linear mixed effects model specifications	131
3.2.5.3. Outline of our analyses	132
3.3. Results and Discussion.....	133
3.3.1. Visualizing grand average waveforms and topographic maps	133
3.3.2. Are N400 effects reduced for form-similar errors in predictable contexts?	135
3.3.3. Do the P600 effects differ across error type and predictability?	137
3.3.4. Investigating how the N400 and P600 effects vary with word-level predictability	137
3.3.4.1. Visualizing grand average waveforms across three cloze probability bins	138
3.3.4.2. Do baseline N400 responses become smaller as predictability increases?	139
3.3.4.3. How do the N400 effects for our two error types change across predictability?... ..	140
3.3.4.4. How do the P600 effects for our two error types change across predictability?	141
3.4. General Discussion.....	142

3.4.1. Should form-based prediction be expected during naturalistic comprehension?	143
3.4.1.1. How predictable was our story?.....	144
3.4.1.2. How slowly was our story read?	147
3.4.1.3. Getting a head start on form-based prediction.....	149
3.4.2. How do our findings relate to prior work on phonological mismatch effects?	154
3.4.3. How do our findings advance understanding of late posterior positivities?.....	159
3.4.4. What are the open questions and what should we (collectively) do next?	163
3.5. Conclusion	168
Chapter 4	169
4.1. Introduction.....	169
4.1.1. How do hearing and deaf signing populations learn to read?.....	171
4.1.2. Predictive processing in written and spoken languages—evidence from EEG.....	174
4.1.3. Using EEG to assess signers’ sensitivity to form features during comprehension.....	177
4.1.4. The present study	179
4.2. Method	181
4.2.1. Participants.....	181
4.2.2. Stimuli.....	181
4.2.3. Procedure	183
4.2.3.1. EEG recording and pre-processing procedure	183
4.2.4. Analysis plan.....	184
4.2.4.1. Determining our spatial and temporal regions of interest.....	184
4.2.4.2. Linear mixed effects model specifications	185
4.3. Results	185
4.3.1. Visualizing the grand average waveforms and topographic maps.....	186
4.3.2. Are N400 effects reduced for handshape-change non-signs in higher cloze contexts?	189
4.3.3. Are the P600 effects sensitive to the non-signs and predictability manipulations?.....	190
4.3.4. Do the baseline N400s change as a function of the sign’s predictability?.....	192
4.4. Discussion	193
4.4.1. Two hypotheses about the emergence of form-based prediction in spoken and sign languages.....	194
4.4.2. Addressing the moderate predictability of our target signs	198

4.4.3. Addressing the variability in handshape similarity across our non-sign manipulations.	199
4.5. Conclusion and future directions	204
Chapter 5	207
5.1. Conclusion	207
5.2. Summarizing the key findings from Papers 1–3	207
5.2.1. Summary of Paper 1	210
5.2.2. Summary of Paper 2	211
5.2.3. Summary of Paper 3	213
5.3. Concluding summary	214
References	216

Acknowledgements

It has been the privilege of a lifetime to have met, worked, and learned alongside such a talented group of scholars, collaborators, and friends. Thank you all for your continued support.

To **Jesse Snedeker**: Thank you for this incredible opportunity. I am so grateful to have been surrounded by your endless enthusiasm, warmth, and wit. I am a better thinker, writer, mentor, and person because of you. I will also never forget how you helped me navigate some of the hardest realities of being human. You saw me as a person first, and I will always be thankful for that.

To **Kate Davidson, Roger Levy, and Gina Kuperberg**: You are the epitome of how scholars should interact, challenge, and support one another. This work has been transformed in your hands, and I am excited to see where it leads us next. It was an honor to have you on my committee.

To **Brian Dillon**: This journey would not have happened without your mentorship. It is not an exaggeration to say that I rely on skills that you taught me every day. I am truly lucky to have your continued guidance, encouragement, and friendship.

To **Akira Omaki**: My life and research are better for having known you. I often think about what we would have accomplished together. You taught me more than I can express, so thank you.

Thank you to all of my past mentors and academic role models, especially Lyn Frazier, Chuck Clifton, Amanda Rysling, Shayne Sloggett, Alfonso Caramazza, Patrick Mair, and Susan Carey.

To **Jayden Ziegler**: Thank you for taking me under your wing. I could not have asked for a better mentor and friend during my time at Harvard. The lab never fully recovered after you left.

To **Tanya Levari** and **Annemarie Kocab**: I have learned so much from you both over the years, and I would not be the researcher I am today without having met you. I am truly thankful for our conversations about teaching, science, and life.

To **Margaret Kandel**: Thank you for being a close friend and the best office mate. I will miss distracting you from your work in order to talk about a new agreement attraction study. You are thoughtful and caring, and your attention to those around you does not go unnoticed.

To my many other lab mates (past, present, and future) in the Snedeker lab, the Meaning and Modality lab (Harvard Linguistics), and the NeuroCognition of Language lab (Tufts Psychology), thank you for the wealth of support, the countless feedback, and the wonderful memories.

Thank you to my co-authors, collaborators, and research assistants, especially Emily Moya, Jessica Moore, Jessica Tanner, Barbora Hlachova, Beatriz Leitao, Ellie Muir, Harita Koya, Jenna Hughes, Karen Andres, Kate Kaufman, Madeleine Presgrave, Maribelle Dickins, Moshe Poliak, Paulina Piwowarczyk, Praneetha Inampudi, Siva Zhou, and Tim Guest.

Thank you to my cohort, my LDS community, and my many friends at Harvard (and beyond), especially Brandon Woo, Brian Leahy, Dan Janini, Durgesh Rajandiran, Emma Van Beveren, Mieke Slim, Ruben Van Genugten, Sarah Raulston, Shari Liu, Shirley Wang, and Tobi Abubakare.

To **Laura Thurlow**: Thank you for a decade of memories and advice. Whether we are traveling the world, or picking pumpkins in Amherst, I know that we will always be there for each other.

To **Arunima Sarin** and **Varun Gupta**: Thank you for showing me what it means to have a home away from home. Over these past years, your love and support has been unwavering. I can never repay you for what you have given me. I cannot wait to see what comes next for you both.

To **Briony Waite**: You always seem to understand how my mind works for (and against) me. Thank you for helping me rediscover all of the things that I enjoy in life. I am so proud of all your accomplishments, and I hope that you are too.

To **my family**: Your unconditional love is my most cherished gift. Each and every one of you inspire me with your selflessness, tenacity, and loyalty to the ones you love. You all gave me a safe space to explore and grow, and I will be forever grateful. I love you all to the moon and back.

To **Pogo**: You will never understand what is written here because you are a cat. But adopting you at the start of this program was the best decision I could have made. Thank you for bringing much needed joy, laughter, and warmth to this journey.

To **Tim Conklin**: Without you, I would have never found the strength to pursue academia, so this accomplishment is as much yours as it is mine. I am constantly in awe of your patience, kindness, intellect, and determination to be better than you were yesterday. I am incredibly lucky to share my life with you, and I am excited for our next adventure. I love you.

Chapter 1

[Introduction]

1.1. Introduction

Language comprehension is the process of using sounds, texts, or gestures to infer the intended message of the person who produced them. To do this, we must rapidly identify the linguistic forms being conveyed, retrieve their unique meanings, link them together into meaningful phrases, and then figure out why someone produced that particular message in the first place. Understanding how each of these processes is carried out, and how they interact with one another during comprehension, has been a central goal of cognitive science for over 60 years.

This dissertation focuses on a particular aspect of language comprehension that has received a lot of attention over the last two decades: linguistic prediction. Linguistic prediction is the process by which we anticipate upcoming words (and their features) before encountering them in the input (see Kuperberg & Jaeger, 2016, Section 3, p. 39). We have long known that linguistic prediction is possible—after all, everyone has had the satisfying experience of finishing someone else’s sentence. What has changed in recent years is that prediction is no longer seen as a rare or marginal phenomenon, but rather as a core aspect of language processing and development (e.g. Chang et al., 2006; Hale, 2001; Levy, 2008; Pickering & Garrod, 2007; but see, Huettig & Mani, 2016; Rabagliati et al., 2016), as well as cognition more broadly (e.g. Bar, 2007; Clark, 2013; Friston, 2005, 2010).

In the remainder of this Introduction, I do four things: First, I outline what we know about the adult language comprehension system, describing the ways in which comprehension unfolds in an incremental, interactive, *and* predictive manner. Second, I explicitly define prediction and review the empirical evidence for predictive processing at various levels of representation (e.g. meaning and form). Third, I discuss the claims that prediction of form (relative to meaning) is less likely to emerge in natural language contexts, motivating the need to study form-based prediction in more ordinary settings. Finally, I state the goals of this dissertation, foreshadowing how each chapter contributes to our understanding of predictive processing in adult comprehension.

1.1.1. The adult language comprehension system is incremental, interactive, and predictive

Adult language comprehension involves the coordination of information from both low and high levels of linguistic representation—the lowest levels of representation are those closest to perception, e.g., perceiving the sensory input (sounds, texts, gestures) and using it to recognize individual words. In contrast, the highest levels of representation are those necessary for understanding an interlocutor's intended messages and goals. Both low-level and high-level representations contribute to the process of understanding how an incoming message relates to the world around you, how that message could (or should) be interpreted, and even how the rest of the conversation is likely to continue (i.e. predictions about upcoming linguistic material). And because these representations are distributed across various levels of the language network, information must be able to flow between them during comprehension.¹

¹ Language production and comprehension are often argued to involve the same cognitive system and network of representations. The main difference, however, is the way that information flows through the system. In comprehension, the dominant flow of information is from linguistic input to messages, whereas in production, the dominant flow is from messages to linguistic output (see Dell & Chang, 2014; Meyer et al., 2016; Pickering & Garrod, 2013).

Decades of psycholinguistic research have demonstrated that this flow of information occurs *incrementally* and *interactively* (e.g. Garnsey et al., 1997; MacDonald et al., 1994; Tanenhaus et al., 1995; Trueswell & Tanenhaus, 1994). For language comprehension, incrementality means that the system passes information from lower levels up to higher ones without waiting for the lower-level processes to finish, and interactivity simply refers to the fact that many distinct sources of information are used to arrive at an interpretation (Allopenna et al., 1998; Altmann & Steedman, 1988; Boland et al., 1995; Eberhard et al., 1995). When information flows from a lower level to a higher one, this process is called *bottom-up processing*. Alternatively, when information flows from a higher level to a lower one, this process is called *top-down processing*.

To illustrate these processing dynamics, imagine someone hearing the following utterance: “Shelley put the birthday cake on the tray and lit the *can*....” After hearing the initial [k] of the incoming word, all of the words with the same phonological onset will become activated to some degree (e.g. *can*, *canary*, *candy*, *candle*, *canon*, *candelabra*, *cancel*, *core*, *cool*, *kick*, *kelp*, *key*; see Allopenna et al., 1998). As more sounds of the incoming word are heard, this activated cohort is incrementally winnowed down to a smaller set of candidate words that are consistent with the bottom-up input (e.g. *can*, *canary*, *candy*, *candles*, *candelabra*, *canon*, *cancel*). At the same time, the listener may also use whatever top-down information is available to constrain their interpretation and help them select the correct word. For example, they may upweight all nouns that refer to inanimate objects (e.g. *can*, *candy*, *candles*, *candelabra*, *canon*); and of those, the ones that can be lit (e.g. *candles*, *candelabra*, *canon*); and of those, the ones that typically co-occur with birthday cakes (e.g. *candles*). Thus, when taken together, these top-down constraints interact with

the bottom-up input and allow the listener to activate *candles* to the exclusion of all other lexical competitors before the word is fully produced.

In this example, the listener tackled the process of identifying the incoming word by using both *incremental* and *interactive* processes. However, there are good reasons to believe that the listener may have also engaged in *predictive* processes before encountering those initial phonemes. For example, if the listener was familiar with an American birthday celebration, they may have inferred a particular sequence of events given the context: *a birthday cake is baked* → *candles are placed into it* → ***candles are lit*** → *people sing “Happy Birthday”* → *candles are blown out* → *the cake is eaten* (for discussion, see Kuperberg, 2021). Given this knowledge, the listener may hear “Shelley put the birthday cake on the tray and lit the...” and then expect to hear *candles* next without receiving any additional bottom-up input. Under this conceptualization, top-down predictions are largely akin to the constraints that aided the listener in recognizing the incoming word on the basis of its initial phonemes. As I discuss below, the main difference between top-down predictions and other kinds of top-down constraints is *when* they emerge.

1.1.2. What is linguistic prediction, and how have studies demonstrated this phenomenon?

In this dissertation, I define linguistic prediction as a process in which top-down information activates a particular linguistic representation *before* it appears in the input (for similar definitions, see DeLong et al., 2014, 2021; Huettig et al., 2022; Kuperberg & Jaeger, 2016, Section 3, p. 39; Kutas & Federmeier, 2011; Pickering & Gambi, 2018). On this definition, linguistic prediction can occur at multiple levels of representation: from the highest levels related to event and argument structures, to the lowest levels related to grammatical, phonological, and perceptual features (e.g. Altmann & Mirković, 2009; McRae & Matsuki, 2009). In the psycholinguistic

literature, there is ample evidence that comprehenders make predictions at multiple levels of representation. Some of the clearest evidence for predictive processing of this kind comes from reading time studies, visual world eye-tracking studies, and studies that use various neuroimaging techniques. In the remainder of this section, I review some of the key findings from these studies, demonstrating how these techniques have been used in our pursuit of understanding linguistic prediction.

1.1.2.1. Key findings from studies using reading paradigms

Some of the earliest evidence for predictive processing came from research into the structural biases the comprehenders have when reading sentences. These biases (or parsing heuristics) may not seem like prediction per se—however, it became clear that readers had expectations about the likely (and unlikely) ways in which a sentence could continue, and those expectations led to processing difficulties when disconfirmed.

For example, the classic syntactic ambiguity involving direct objects and sentential complements is one such case. When a comprehender reads a sentence like “The man accepted the prize was not going to him,” they are likely to slow down upon reading “was not” (e.g. Ferreira & Henderson, 1990; Holmes et al., 1989). This slowdown is often attributed to readers’ expectations for a direct object following the verb *accepted*, which led them to incorrectly interpret *the prize* as a direct object instead of the beginning of a sentential complement. This interpretation of the data is supported by the fact that readers do not show similar slowdowns on “was not” when the sentence is disambiguated as in “The man accepted *that* the prize was not going to him.”

These structural expectations come from readers’ sensitivity to the kinds of syntactic environments that verbs typically occur in. Some verbs like *saw*, *read*, and *wrote* largely precede

direct objects like *the accident, the book, or the letter*. Other verbs like *realized, learned, and believed* largely precede sentential complements like *the answer was not that simple*. Thus, upon encountering a verb, the reader is likely to make a prediction about the syntactic category of the next word. At the simplest level, these findings demonstrate that comprehenders generate some structural expectations about the most probable ways for a sentence to continue during comprehension (see Hale, 2001; Levy, 2008).

Additional evidence from the reading literature suggests that readers make even narrower predictions about the specific meanings and words that are likely to appear next. Decades of research have shown that readers spend less time fixating on words (and skip them more often) when they are highly predictable in their given contexts (Balota et al., 1985; Boston et al., 2008; Ehrlich & Rayner, 1981; Frisson et al., 2017; Luke & Christianson, 2016; Rayner et al., 2001; Smith & Levy, 2013; for review, see Staub, 2015).

For example, Frisson et al. (2017, Experiment 1) had participants read sentences like those in (1). Critically, these sentences either strongly constrained for a particular target word (e.g. *liver* in 1a and 1b) or did not constrain for any word at all (1c and 1d). In the constraining contexts, the authors presented the highly predictable word (*liver*) or replaced it with a semantically plausible but unexpected word (*heart*). They used the same word pairs as controls in the non-constraining contexts. To assess predictability, Frisson et al. conducted a cloze task (Taylor, 1953) in which participants read the beginning of each sentence (up to the target word) and then guessed what word would come next. With these data, the authors calculated a cloze probability for each target word by finding the proportion of all participants that provided the target word in that particular context (e.g. if 7 out of 10 people said *liver* in 1a, then that word would have a cloze probability of 70% in that context).

- (1)
 - a. The doctor told Fred that his drinking would damage his *liver* very quickly.
 - b. The doctor told Fred that his drinking would damage his *heart* very quickly.
 - c. Yesterday Fred told his friend that they will look at his *liver* very thoroughly.
 - d. Yesterday Fred told his friend that they will look at his *heart* very thoroughly.

Frisson et al. found that the predicted target words in constraining contexts (*liver* in 1a, mean target cloze = 70.2%) were read faster (and skipped more often) than the identical target words in non-constraining contexts (*liver* in 1c, mean target cloze = 0.5%). In contrast, there were no differences in fixation times for the unpredictable words in constraining contexts (*heart* in 1b, mean target cloze = 1.8%) and the identical words in non-constraining contexts (*heart* in 1d, mean target cloze = 1.2%). Similar evidence for predictive processing during natural reading has emerged in recent studies using longer texts. In these studies, researchers demonstrated that reading times correlate with various measures of word-level predictability (e.g. *n*-gram, surprisal, human cloze) in more naturalistic reading environments (Lowder et al., 2018; Luke & Christianson, 2016; Smith & Levy, 2013). Essentially, the more predictable a word is in a particular context, the easier and faster it is to process.

1.1.2.2. Key findings from studies using the visual world paradigm

Some of the strongest evidence for predictive processing comes from studies using the visual world paradigm (Allopenna et al., 1998; Eberhard et al., 1995; Tanenhaus et al., 1995). In these studies, participants look at visual displays, and sometimes perform a behavioral task, while they listen to spoken sentences (for review, see Huettig et al., 2011). Researchers use the visual world paradigm to see where people fixate as they incrementally process the incoming speech.

The inspiration for this technique came from Roger Cooper (1974) after he noticed that, while listening to short narratives, people would fixate on objects being mentioned and on objects related to the unfolding context (e.g. looking at lions after hearing *lion*, and looking at jungle animals after hearing *Africa*). Using this paradigm, researchers have uncovered that listeners' eye-movements can be driven by semantic properties (e.g. looking to trumpets after hearing *piano*), and even physical properties like shapes and colors that are related to the words being produced (e.g. looking to twisty cables after hearing *snake*, and looking to red things after hearing *strawberry*; see Huettig & Altmann, 2004, 2005, 2011).

Critically, for research on predictive processing, this technique allowed researchers to assess comprehenders' predictions *before* the target representation appeared in the input. For example, in a foundational study, Altmann and Kamide (1999) presented participants with displays containing objects like a boy, a cake, and a set of toys. Then, participants heard sentences like "The boy will **move** the cake" or "The boy will **eat** the cake." If the verb was *eat*, participants would look to the cake earlier (as it was the only edible object in the display) relative to when the verb was *move*. This evidence suggests that comprehenders can use the selectional restrictions of the verb to anticipate an upcoming referent, i.e., the inanimate, edible object in the scene (for similar findings with children, see Mani & Huettig, 2012). These predictions, however, are not simply restricted to the verb—rather, it seems that comprehenders can rely on the entire context to generate their expectations. Kamide et al. (2003) demonstrated that hearing "The man will ride..." vs. "The girl will ride..." resulted in differential looks to a motorcycle and a carousel respectively (for similar findings with children, see Borovsky et al., 2012).

These visual world studies are often argued to demonstrate prediction at higher levels of representation (e.g. event structures, argument roles, and semantic features of upcoming words).

There is, however, another set of clever studies that demonstrate predictions at even lower levels related to the form of upcoming words (i.e. grammatical, phonological, and orthographic features). For example, Lew-Williams and Fernald (2007) found that Spanish-speaking adults and children can use grammatical features like the gender of pre-nominal articles to guide their expectations during comprehension. In this study, participants saw displays containing two objects. These objects either had the same grammatical gender (*la pelota*, “the_{FEM} ball_{FEM}” and *la galleta*, “the_{FEM} cookie_{FEM}”) or different grammatical gender (*la pelota*, “the_{FEM} ball_{FEM}” and *el zapato*, “the_{MASC} shoe_{MASC}”). While viewing these scenes, participants heard a sentence that directed them towards one of the objects (e.g. *Encuentra la pelota*, “Find the ball”). The authors found that participants were faster to look at the target object in trials with objects of *different* genders than in trials with objects of the same gender. These findings suggest that, in trials where the article matched the gender of only one object, both adults and children were able to anticipate which object would be mentioned next upon hearing the gender-marked article.

More direct evidence for the prediction of word-form features comes from a recent visual world study by Li et al. (2022). In this study, Mandarin-speaking adults saw displays containing four objects while listening to sentences with highly predictable target words, e.g., “After school, I put my pencil case and notebooks into my **schoolbag** and get ready to go home.” In each trial, the displays contained at least three objects that were unrelated to the predictable target word. The fourth object was manipulated to either be the predicted object (a schoolbag), another unrelated distractor (a funnel), or an object with similar semantic features (an eraser) or phonological form (a comb, as the word *comb* in Mandarin has the same first syllable and tone as *schoolbag*).

Li et al. (2022) found that participants began to look at the semantic and phonological competitors (over the distractors) well before hearing the target word in the sentence. These

increased anticipatory looks began roughly 1400 ms (or the duration of two words in their experiment) before the onset of the predicted target word (*schoolbag*). The authors interpreted these patterns as clear evidence that comprehenders are able to anticipate both the meaning *and* form of upcoming words before hearing them in the input (for similar findings in Japanese, see Ito, 2019).

1.1.2.3. Key findings from studies using electroencephalography

Evidence for prediction during language comprehension comes not only from visual world and reading paradigms but also from electroencephalography (EEG). Researchers use EEG to record changes in voltage on the scalp as participants perceive linguistic stimuli. By time-locking these changes to the onset of a particular word, researchers can generate event-related potentials (ERPs). ERPs vary systematically in their amplitudes, latencies, and scalp distributions, making them useful for characterizing when and to what degree different variables affect cognitive processes like language comprehension (see Kappenman & Luck, 2011; Luck, 2014). For this reason, EEG can potentially provide information about when predictions occur and at which level(s) of representation they are being made. Some EEG studies assess prediction *before* the critical word appears in the input (similar to visual world studies), while other studies look for effects of prediction on the critical words themselves (similar to most reading studies).

In this section, I review each of these approaches in turn: First, I describe the handful of studies that investigate the neural responses to pre-nominal articles and adjectives that match or mismatch the predicted features of an upcoming noun. Second, I briefly explain how researchers can analyze the continuous EEG data evoked *before* the presentation of target words to assess when (and at which levels) predictions are being made. Finally, I discuss the numerous studies that

look for effects of prediction on the critical words themselves. Specifically, I highlight studies that find predictive effects on a set of early (but somewhat illusive) components, as well as the more established N400 component.

1.1.2.3.1. Predictive effects on pre-nominal articles and adjectives

There are only a handful of studies demonstrating effects of prediction on pre-nominal articles and adjectives (for a recent overview, see Nicenboim et al., 2020). Most of these studies rely on languages with explicit gender marking or phonological constraints on words that introduce or modify an upcoming noun (e.g. Spanish, Italian, German, Dutch, and English). Specifically, in these studies, researchers construct sentences that strongly constrain for a particular noun, and then manipulate pre-nominal words to match or mismatch the features of that noun. Researchers then record the ERP responses evoked by these matched or mismatched pre-nominal words and compare them across conditions to detect any differential effects.

For example, Wicha et al. (2004) investigated form-based prediction in Spanish by manipulating the gender of pre-nominal articles. They had Spanish-speaking adults read sentences that constrained for an upcoming noun (mean target cloze = 80%). Then, they either introduced that predictable noun with an article that matched in gender (*la corona*, “the_{FEM} crown_{FEM}”) or mismatched (*el corona*, “the_{MASC} crown_{FEM}”). Results indicated that mismatched articles evoked a late posterior positivity relative to matched articles, suggesting that readers were sensitive to the mismatch between the features of the incoming article and the *predicted* features of the upcoming word.

Other studies using similar designs in both Spanish and Dutch have either found qualitatively different ERP effects (e.g. larger *negativities* in response to pre-nominal mismatches,

Foucart et al., 2014; Martin et al., 2018; Otten et al., 2007; Otten & Van Berkum, 2009) or failed to find any effects at all (Kochari & Flecken, 2019; Otten & Van Berkum, 2009). Although, Nicenboim et al. (2020) conducted a recent Bayesian meta-analysis of the data from these failed replications (as well as other null pre-nominal findings in German and English) and found an extremely small—but reliable—effect of prediction on pre-nominal words across studies.

In addition to these studies on gender, there is also a small (and contentious) literature on whether comprehenders can use the phonological features of a predicted noun to anticipate the form of a pre-nominal article. For example, there are phonological rules in English that dictate which article should be used before words that start with a consonant or vowel sound. If a word begins with a consonant sound like *kite*, it must be preceded by the article *a*. If a word begins with a vowel sound like *airplane*, it must be preceded by the article *an*.

DeLong et al. (2005) cleverly used these rules to assess whether comprehenders anticipate the form of upcoming nouns in English. To do this, they had English-speaking adults read sentences that constrained for a particular noun like *kite* (e.g. “The day was breezy so the boy went outside to fly...”). Then, they either provided the expected article-noun pair (*a kite*) or an unexpected article-noun pair (*an airplane*). If comprehenders predicted the form of *kite*, they should show differential ERPs to expected and unexpected articles before seeing the target noun. The results of this study were not as clear-cut as prior work: there was no overall difference between the ERPs evoked by matched and mismatched articles—however, there seemed to be a graded effect based on the predictability of the article itself (ascertained by human cloze probabilities). The more predictable the article, the smaller the evoked negativity (for similar findings on the N400, see Section 1.1.2.3.4 below).

These findings have been the center of a recent debate in the psycholinguistic literature. Some EEG studies replicate this graded effect, as well as the expected categorical effect, on pre-nominal articles in English (Martin et al., 2013). Other studies have used different behavioral tasks and response measures to demonstrate similar effects of prediction at the phonological level in English (e.g. Husband, 2022). However, while there are a handful of studies replicating or supporting these findings, there are also studies failing to replicate them (Ito et al., 2017a, 2017c; Nieuwland et al., 2018). The most well-known failure to replicate these effects comes from Nieuwland et al. (2018). In this study, a team of researchers spanning nine laboratories aimed to replicate the original findings from DeLong et al. (2005). Despite recruiting over 350 participants, they were unable to replicate the graded effect of predictability on the negativity evoked by pre-nominal articles.

There are, however, some issues surrounding this particular replication attempt (Ito et al., 2020; Nicenboim et al., 2020; Yan et al., 2017). For example, Yan et al. (2017) highlighted a few key differences between the original DeLong study and the Nieuwland replication: First, the two studies had different procedures for processing and analyzing their ERP data (e.g. baselining differences, by-trial statistical approaches). Second, unlike the original DeLong study, the replication studies did not include any filler trials, which meant that there was a greater proportion of trials in which the indefinite article strongly disconfirmed comprehenders' predictions for the upcoming word. Third, Yan et al. noted that human cloze probabilities may not be the best measure of predictability for pre-nominal articles, as they lack precision for very low probability events relative to other information-theoretic measures like surprisal. When reanalyzing the data using surprisal values, the authors found evidence for the original graded effect on pre-nominal articles in eight out of the nine total datasets. Since then, additional re-analyses of these data have found

similar evidence for a reliable—albeit small—effect of predictability on pre-nominal articles in English (see Nicenboim et al., 2020; Urbach et al., 2020).

While the evidence from English is somewhat inconsistent, converging evidence for phonological prediction has been found in other languages. For example, similar to most Romance languages, Italian marks the gender of upcoming words on the pre-nominal articles (e.g. *un libro*, “a_{MASC} book_{MASC}” vs. *una mela*, “an_{FEM} apple_{FEM}”). There are also additional rules regarding the form of pre-nominal articles depending on the phonological onset of the noun. If the noun begins with a vowel *and* it is masculine, the article remains unchanged (e.g. *un incidente*, “an_{MASC} accident_{MASC}”). If the noun begins with a vowel *and* it is feminine, the final vowel of *una* is elided (e.g. *un'inondazione*, “a_{FEM} flooding_{FEM}”). Finally, if the noun is masculine and begins with a particular consonant or vowel cluster, the masculine article becomes *uno* (e.g. *uno scontro*, “a_{MASC} collision_{MASC}”). Thus, in a language like Italian, it is possible to manipulate both the gender and the phonological features expressed on pre-nominal articles.

Ito et al. (2020) made use of these rules to assess the nature of readers' predictions during comprehension. They had native Italian speakers read highly constraining sentences like “*Il traffico in autostrada è rimasto bloccato a causa di...*” (“The traffic on the motorway came to a standstill because of...”). Then, the sentence continued with either the expected article-noun pair (*un incidente*, “an_{MASC} accident_{MASC}”), an article-noun pair with unexpected gender features (*un'inondazione*, “a_{FEM} flooding_{FEM}”), or an article-noun pair with unexpected phonological features (*uno scontro*, “a_{MASC} collision_{MASC}”). Similar to the studies above, Ito et al. found larger negativities on pre-nominal articles for gender mismatches (250–800 ms) and phonological mismatches (450–800 ms). Interestingly, the effects for phonological mismatches were delayed relative to those for gender mismatches, which the authors interpreted as evidence that people may

be faster (and more likely) to predict gender features relative to phonological features on pre-nominal articles. Evidence in support of this interpretation—or at least the claim that it takes longer to predict at lower levels relative to higher levels of representation—comes from a set of additional EEG studies reviewed below (e.g. Ito et al., 2016; Wang et al., *under revision*).

1.1.2.3.2. *Predictive effects demonstrated in pre-target continuous EEG*

Within the last decade, many neuroimaging studies of language comprehension have begun to use Representational Similarity Analyses (RSA) to assess the nature of comprehenders' predictions (e.g. Hubbard & Federmeier, 2021; Wang et al., 2018). The general assumption in RSA is that particular representations evoke distinct spatiotemporal patterns when activated, and these distinct patterns are captured in our EEG recordings (for similar approaches using MEG, see Dikker & Pykkänen, 2013; Sohoglu et al., 2012). For example, the process of activating the semantic and form features associated with a concept like <apple> will result in a distinct spatiotemporal pattern. This pattern is consistent within an individual, and it is argued to emerge whenever <apple> becomes activated. This activation can occur via new bottom-up input, e.g., after hearing [æpəl] in the speech stream, or via comprehenders' top-down predictions in a given context, e.g., “From the tree, I picked a juicy, bright red....” As a result, researchers can analyze the distinct spatiotemporal patterns from continuous EEG (as well as the patterns evoked by critical words themselves) to assess when and at which level predictions are being made.

To illustrate this, Wang et al. (*under revision*) created triads of sentences like those in (2): The first sentence (2a) strongly constrained for a predictable noun (e.g. *bank*). The authors created two additional sentences that constrained for a homograph of the target word (financial *bank* vs. river *bank*, form-overlap in 2b) or a semantically related word with no overlap in form (financial

bank vs. financial *loan*, semantic-overlap in 2c). These target sentences were presented word-by-word at a rather slow rate. Each word appeared for 300 ms, and then a blank screen was presented for another 400 ms between words. During the sentence presentation, the authors recorded participants' continuous EEG data.

- (2) a. "I went to deposit the check at the...*bank*."
- b. "The muddy sides of the river are called the river...*bank*."
- c. "His college was very expensive, so he had to take out a student...*loan*."

In their analysis, Wang et al. extracted the spatiotemporal patterns from the continuous EEG recordings for each trial across their entire sampling window. Then, they calculated the similarity between the patterns evoked by the sentence contexts in each condition (up until the presentation of the target word). As a control condition, they also calculated the similarity between the patterns from the target sentence (2a) and a constraining sentence from a different triad (e.g. "The incident taught us a valuable life...*lesson*"). If these contexts activated shared semantic and form features, and the spatiotemporal patterns reflected those activated features, then the similarity should be greater in comparisons *within* a particular triad (2a and 2b; 2a and 2c) relative to the unrelated control comparison.

Results indicated that the patterns evoked by contexts predicting similar semantic features (2a and 2c) had a greater similarity than those from the control conditions. This representational similarity emerged around 300 ms after the onset of the pre-target word (or 400 ms before the onset of the target word). The authors also found that the patterns evoked by contexts predicting the same form features (2a and 2b) had a greater similarity than those from the control conditions.

This form-based effect, however, emerged a bit later than the semantic effect at around 600 ms following the pre-target word (or roughly 100 ms before the onset of the target word). In line with prior work, this study demonstrated that comprehenders can predict the meaning and form of upcoming words before encountering them in the input. Moreover, these findings also support the claim that predictions about upcoming meanings emerge earlier than those about upcoming word-forms. I discuss this point further in Section 1.1.3.

1.1.2.3.3. Predictive effects on the target words: Evidence from early components

In the sections above, I reviewed effects of prediction that emerge *before* the target words appear in the input. There are, however, additional studies that address predictive processing by comparing ERP effects on the target words themselves (or on the unexpected words that replaced them). In these studies, researchers typically compare specific ERP components that have been previously linked to particular aspects of predictive processing.

For example, in the form-based prediction literature, there are various (early emerging) ERP components that have been linked to comprehenders' expectations about the orthographic and phonological forms of upcoming words. These early components are comprehensively reviewed in Nieuwland (2019)—however, as the author notes, most of them are not easily replicated. In his review, Nieuwland focuses on a set of 10 early components that have been argued to reflect the detection of mismatches between the form of a predicted word and the form of the word in the input (e.g. ELAN, P130, N1/P2, N200, PMN, N250). These components have been found across 16 studies. Nine of these components, however, either failed to replicate in at least one study or have not yet been subject to direct replication. The only component in his review that was consistently replicated was the N250 (Brothers et al., 2015). Although, the N250 is often

difficult to disentangle from another well-known negativity, the N400—largely because both components have the same polarity and similar scalp distributions, and the peak of the N250 often coincides with the typical time window used to analyze the N400. Given the inconsistent nature of these early components (and the possibility for component overlap with the N400), there are few conclusions that one can draw from them regarding the nature of prediction. Nevertheless, for the sake of completeness, I review one component that is particularly relevant to speech processing (and two of the three studies in this dissertation): the Phonological Mismatch Negativity (PMN).

Connolly and Phillips (1994) first coined the term PMN to describe the early negativities evoked by replacing highly predictable target words with violations that had unexpected phonological onsets. Specifically, in this study, English-speaking adults heard sentences that strongly constrained for a particular sentence-final word like those in (3). Then, the authors either kept the predicted word (*door*, 3a) or replaced it with a violation that shared semantic features (*sink* → *kitchen*, 3c), phonological onsets (*eyes* → *icicles*, 3b), or neither (*milk* → *nose*, 3d).

- (3) a. At night, the old woman locked the ***door***. *Onset match, Semantic match* (door)
b. Phil put some drops in his ***icicles***. *Onset match, Semantic mismatch* (eyes)
c. They left the dirty dishes in the ***kitchen***. *Onset mismatch, Semantic match* (sink)
d. Joan fed her baby some warm ***nose***. *Onset mismatch, Semantic mismatch* (milk)

Evidence from their earlier work suggested that this particular design should evoke two distinct negativities: one related to unexpected semantic features at around 400 ms (the N400, see Section 1.1.2.3.4) and the other related to encountering unexpected initial phonemes at around 200–300 ms (the PMN). Results confirmed these predictions, as the authors found typical N400

effects for conditions with mismatched semantic features (3b and 3d) relative to those with more congruent meanings (3a and 3c). For the phonological mismatches, the conditions with unexpected onsets (3c and 3d) evoked larger N400s than conditions with expected onsets (3a and 3b)—but note, the PMN in the double violation condition (3d) was unexpectedly larger than the PMN in the condition with only a mismatched onset (3b). In fact, the PMN and the N400 actually blurred together in this condition, forming one large, broadly distributed negativity between 200–600 ms.

Connolly and Phillips interpreted these PMNs as reflecting the influence of top-down expectations on the earliest stages of word recognition. Specifically, they argued that the sentences in their study gave rise to strong expectations about sentence-final words. Then, upon hearing the initial phonemes of the sentence-final word, listeners activated a cohort of words with the same phonological onset (see also, Section 1.1). If the incoming word had an onset that matched the predicted word, then the listener was able to activate a cohort containing the expected continuation (e.g. the onset of *icicles* in 3b can activate a cohort containing the predicted word, *eyes*). If the incoming word mismatched the onset of the predicted word, then the listener activated a cohort *without* the expected continuation (e.g. the onset of *kitchen* in 3c cannot activate a cohort containing the predicted word, *sink*). This initial conflict between the bottom-up input and the listeners' top-down expectations is what arguably gives rise to the PMN.

There are some difficulties, however, in accepting this interpretation. For example, their account does not predict that the PMN should be larger in the double mismatch condition (3d) relative to the condition with only mismatched onsets (3c). Thus, it seems like information beyond phonology is influencing the amplitude of this early evoked negativity. One possible explanation for this pattern could be that the PMN and the N400 are not as functionally distinct as the authors proposed, and the PMN simply reflects an early emerging congruity effect (e.g. Van Petten et al.,

1999). Recently, Lewendon et al. (2020) argued that, even after 30 years of research, there is not sufficient evidence to suggest that the PMN can be reliably differentiated from the N400. For these reasons, throughout this dissertation, I will primarily rely on the N400 response to assess the degree to which comprehenders make predictions about the meaning and form of upcoming words.

1.1.2.3.4. Predictive effects on the target words: Evidence from the N400

In contrast to these early and illusive components, the N400 is well-understood and has been consistently replicated in the psycholinguistic literature. The N400 was first reported by Kutas and Hillyard (1980) in sentence contexts where a highly predictable word was replaced by something unexpected, e.g., “He spread the warm bread with...*socks*.” Since then, numerous studies have demonstrated that this negativity is not an index of violation, and that its amplitude is inversely correlated with the predictability of a word given its preceding context—that is, the more predictable the word, the smaller (or more positive) the N400 response (e.g. DeLong et al., 2005; Kutas & Hillyard, 1984). It has also been observed that N400s become smaller as a context unfolds, likely due to the accumulation of contextual constraints that make each new word increasingly more predictable (e.g. Payne et al., 2015; Van Petten, 1993).

Taken together, these properties of the N400 are consistent with a framework in which top-down processes generate predictions about the features of upcoming words, making it easier to access incoming words if they are associated with those predicted features (for discussions, see Federmeier, 2007, 2022; Kuperberg et al., 2020; Kutas & Federmeier, 2011). Prior work has demonstrated that the N400 is sensitive to predictions of semantic features (e.g. Federmeier et al., 2002; Federmeier & Kutas, 1999; Kuperberg et al., 2020; Thornhill & Van Petten, 2012). For example, Federmeier and Kutas (1999) had participants read sentences like “The yard was

completely covered with a thick layer of dead leaves. Erica decided it was time to get out the (*rake / shovel / hammer*).” This particular context generates strong expectations for the word *rake*, presumably causing comprehenders to predict the lexicosemantic features associated with *rake*. Given the conceptualization of the N400 above, we should expect reduced N400s to *shovel* relative to *hammer*, as *shovel* shares more semantic features with the predicted word *rake*. Results confirmed this prediction, as the N400 responses for semantically similar words (*shovel*) were indeed reduced relative to less similar words (*hammer*). These findings support the idea that the N400 response to incoming words reflects the degree of mismatch between its own semantic features and the set of features that were pre-activated by the prior context (see also Kuperberg et al., 2020).

Additional studies using similar designs have shown that the N400 is also sensitive to predictions about the form of upcoming words (e.g. DeLong et al., 2005, 2019, 2021; Ito et al., 2016; Kim & Lai, 2012; Laszlo & Federmeier, 2009; Wicha et al., 2004). For example, Laszlo and Federmeier (2009) had English-speaking adults read sentences with highly predictable endings like those in (4). Then, the authors either kept the predictable endings or replaced them with violations that had similar or dissimilar forms (i.e. *orthographic neighbors* in 4a and *non-neighbors* in 4b). They also manipulated whether each of these violations were actual words (*bark*, *clam*), nonwords (*pank*, *horm*), or illegal letter strings in English (*bxnk*, *rqck*).

- (4) a. Before lunch, he had to deposit his paycheck at the (*bank / bark / pank / bxnk*).
b. The genie was ready to grant his third and final (*wish / clam / horm / rqck*).

Consistent with the prior findings of semantic prediction, the authors found reduced N400s to form-similar conditions (*bank* → *bark* or *pank*) relative to form-dissimilar conditions (*wish* → *clam* or *horn*). This pattern supports the claim that, when given strong contextual constraints, readers can make predictions about the form features of upcoming words. These findings have been replicated in numerous studies, and critically, they have been linked to predictive processes, as the reduced N400s to unexpected but form-similar words disappears in less predictable contexts (see Ito et al., 2016).

1.1.3. Studying linguistic prediction of meaning and form in natural language contexts

The majority of evidence for prediction of both meaning and form has come from studies that use clever and carefully controlled experimental designs. In these studies, researchers tend to create optimal conditions for making predictions about which words (and features) are likely to appear next in a sentence. For example, they construct a set of highly constraining, single sentence contexts that generate expectations for one particular target word. Then, they present these sentences at relatively slow rates, providing comprehenders with ample time to process the input and make inferences about upcoming linguistic material.

In a recent study, Ito et al. (2016) demonstrated how prediction can be influenced by these two properties—namely, the predictability of words and the rate of presentation. To do this, they had English-speaking adults read sentences with either highly predictable target words (e.g. “Nobody knows the time as this room has no...*clock*,” mean target cloze = 93.5%) or moderately predictable target words (e.g. “Paul is trying to stand on one...*leg*,” mean target cloze = 65.1%). Similar to prior work, the authors either kept the original words or replaced them with one of three violations: an unexpected word with a similar meaning (*clock* → *alarm*, *leg* → *hip*); an unexpected

word with a similar form (*clock* → *clerk*, *leg* → *lag*); or an unexpected word with neither a similar meaning nor form (*clock* → *scarf*, *leg* → *kid*). In addition, they manipulated the rate of presentation by conducting the experiment twice: once at a rate of 2 words per second, and then again at a rate of 1.5 words per second.

The authors found clear N400 effects for all unexpected conditions in both experiments. For violations with similar meanings (*clock* → *alarm*, *leg* → *hip*), the N400 responses were reduced in highly predictable but not moderately predictable contexts (regardless of presentation rate). For violations with similar forms (*clock* → *clerk*, *leg* → *lag*), the N400 responses were also reduced in highly predictable contexts—but only when sentences were presented at the slower rate. These findings provide evidence that prediction of semantic features may occur earlier than predictions of form, and moreover, that form-based prediction seems to require optimal environments to emerge during comprehension (but see, DeLong et al., 2021). For these reasons, many theorists have argued that form-based prediction is unlikely to occur in natural language, as it is both less predictable and produced faster than the sentences in our typical psycholinguistic experiments (Freunberger & Roehm, 2016, 2017; Indefrey & Levelt, 2004; Ito et al., 2016; Pickering & Garrod, 2007).

For example, Luke and Christianson (2016) demonstrated that the majority of words in natural texts are relatively unpredictable. In this study, they characterized the predictability of every word in a collection of 55 naturally written discourses. To do this, they had participants read these discourses and guess entire passages word-by-word. The authors used these guesses to ascertain the cloze probabilities for all of their 2,689 content and function words. They found that the content words had relatively low predictability on average (mean target cloze = 13%). Moreover, only 5% of these words could be labeled as highly predictable (i.e. having cloze

probabilities above 65%, see also Lowder et al., 2018; Smith & Levy, 2013). In contrast, most studies on prediction have entire *sets* of target words with cloze probabilities above 65%, highlighting the stark contrast between the predictability of language used inside and outside of the lab.

It has also been observed (and it is somewhat intuitive) that natural language unfolds at faster rates than those used in typical research (DeLong et al., 2005, 2019; Ito et al., 2016; Kim & Lai, 2012; Laszlo & Federmeier, 2009; but see, DeLong et al., 2021). Prior work has estimated that natural listening and reading occurs at rates of roughly 3–5 words per second (Brysbaert, 2019; Tauroza & Allison, 1990). For context, the Ito study above used presentation rates that were nearly twice as slow at roughly 1.5–2 words per second. If form-based prediction requires additional time to emerge, as well as highly predictable language, how widespread can this phenomenon be in more naturalistic settings?

To begin addressing this question, I present three studies in this dissertation that pull together lines of research using the clever and careful experimental designs from prior work, as well as more recent, innovative designs that use naturalistic stimuli to study language processing. In the remainder of this section, I first discuss some of this naturalistic research and then motivate the use of our novel naturalistic technique, the *Storytime* paradigm.

Much of the psycholinguistic research on natural language comprehension comes from studies using functional magnetic resonance imaging (fMRI, see Brennan, 2016; Hasson & Egidi, 2015). Historically, this body of work has focused on answering questions about language processing at the broadest levels, e.g., which brain regions are implicated in the comprehension of written and spoken narratives (see Speer et al., 2007; Wehbe et al., 2014; Xu et al., 2005; Yarkoni et al., 2008). More recently, researchers have begun to construct computational models that predict

how language might be processed, and then assess how well those models map onto the patterns of activation evoked by the brain during comprehension (for discussion, see Brennan, 2016; Willems et al., 2016).

For example, Willems et al. (2016) used neuro-computational data to investigate the neural bases of prediction during spoken language comprehension. To do this, the authors found three naturalistic stories and then ascertained various measures of predictability for every word in these stories. Specifically, they used a language model to determine *entropy* (a measure of how uncertain the model is about the next word in a sentence) and *surprisal* (a measure of how unexpected the observed word is given the preceding context). Then, they collected fMRI data from English-speaking adults while they listened to the same three stories without any experimental manipulations. They also had control conditions in which participants heard the same three stories in reverse. Willems et al. found that particular brain regions showed greater sensitivity to entropy and surprisal measures relative to the reversed control conditions, suggesting that these measures of prediction can be mapped to patterns of activation during comprehension. Interestingly, they also found that the brain regions sensitive to surprisal measures spanned regions that are implicated in both high and low-level processes (e.g. visual word form area). The authors interpreted this finding as evidence that prediction occurs at the level of meaning and form during naturalistic comprehension of speech.

One challenge, however, in using fMRI to study linguistic prediction is the poor temporal resolution of this imaging technique. As a result, many researchers have adapted this naturalistic paradigm for use in EEG, as it has far better temporal resolution than fMRI and has been instrumental to our understanding of how (and when) predictions are made during comprehension (e.g. Aurnhammer & Frank, 2019; Bhattasali et al., 2020; Brennan & Hale, 2019; Broderick et al.,

2018; Heilbron et al., 2022; Levani & Snedeker, 2018; Payne et al., 2015). Similar to the prior work in fMRI, these studies explore how naturally occurring variation in written and spoken discourses affect the ERP signals at each word. Single-trial ERP designs like these often ask participants to simply read or listen to natural language while recording neural responses to every word. Then, they correlate the by-word ERP responses like N400 amplitudes with the lexical properties of those words (e.g. frequency, concreteness, cloze probability, surprisal, entropy, word length).

One limitation to this approach, however, is that the findings from these studies are largely correlational and may be confounded with a wide range of unknown, unmeasured, and uncontrolled variables (for discussion, see Brothers & Kuperberg, 2021). Thus, in this dissertation, we take an alternative approach that, to the best of our knowledge, has not been used in prior psycholinguistic research. Rather than simply collecting neural responses to every word in a natural discourse and correlating them to lexical properties, we actually select a set of target words (with matched lexical properties) and directly manipulate them within the discourse. This approach allows us to graft tightly controlled experimental designs onto natural language, providing the conditions necessary for testing our hypotheses while reaping the benefits (and gaining the ecological validity) associated with natural language comprehension. In all three studies described below, we refer to this novel naturalistic technique as the *Storytime* paradigm.

1.1.4. Summary of the chapters in this dissertation

The goal of this dissertation is to investigate if (and when) prediction goes beyond the level of meaning during naturalistic comprehension. As I discussed above, form-based prediction has been argued to be an edge case in the predictive processing literature, requiring the most optimal

environments to emerge during comprehension. Numerous studies have tested the limitations of prediction using clever manipulations of contextual constraint and presentation rate—however, none of these studies have directly manipulated words within a larger discourse to actually test this hypothesis. As a result, the three studies in this dissertation use the *Storytime* paradigm to characterize whether form-based prediction can occur in natural language.

In Paper 1, we investigated the degree to which bilingual speakers make form-based predictions about upcoming words while listening to code-switched narratives. By definition, bilingual populations can express a single (lexical) concept using two words from different languages. Thus, there is an interesting question about whether bilinguals predict a specific word-form in a particular language during comprehension. To address this question, we compared N400 responses to switched and non-switched words within two naturally spoken stories. We also further manipulated these words such that some strongly fit within their contexts while others did not. This particular experimental design allowed us to test the nature of bilinguals' predictions in more natural linguistic environments.

In Paper 2, we further assessed form-based prediction in spoken language comprehension. Specifically, we investigated whether English-speaking adults make predictions about the phonological forms of upcoming words while listening to a 30-minute children's story. Following prior studies on prediction, we had a 2×3 design crossing predictability (higher cloze, lower cloze) and word type (baseline word, form-similar violation, or form-dissimilar violation). Similar to Ito et al. (2016), we expected to find reduced N400 responses to form-similar violations but only in predictable contexts. To foreshadow our findings, we indeed replicate this pattern and demonstrate that prediction at the phonological level can occur in natural language settings.

In Paper 3, we began to ask how widespread form-based prediction is across linguistic modalities. Nearly all of the prior work on prediction has focused on written and spoken languages—thus, we wanted to characterize prediction in another modality. To do this, we conceptually replicated Paper 2 in American Sign Language (ASL) by assessing the degree to which deaf signers anticipate the visual-manual features of upcoming signs during comprehension. Specifically, we translated the children’s story from Paper 2 into ASL and then implemented a similar 2×3 design crossing predictability (higher cloze, lower cloze) and sign type (baseline sign, form-similar sign, form-dissimilar sign). To make our violations, we manipulated the three main parameters of a sign: the handshape, the location, and the movement. The form-similar signs only had one parameter change (the handshape), whereas the form-dissimilar sign had all three parameters changed. In this study, we find tentative evidence that deaf signers indeed predict the handshape of upcoming signs during natural comprehension.

Across these three studies, we consistently demonstrate that prediction can go beyond the meaning of upcoming words in natural language contexts. In addition, we provide a proof of concept that typical psycholinguistic manipulations can be successfully injected into naturalistic stimuli, allowing for research that is both ecologically valid *and* tightly controlled. As we discuss in each chapter, this body of work provides exciting avenues for future research and presents clear demonstrations of how prediction works in the wild. This work, however, is not without its mysteries, and so we take the time to explore these findings in more detail throughout the dissertation.

Chapter 2

[Paper 1]

Unexpected words or unexpected languages?

Two ERP effects of code-switching in naturalistic discourse

Anthony Yacovone, Emily Moya, & Jesse Snedeker

Published 2021 in Cognition

[\[https://doi.org/10.1016/j.cognition.2021.104814\]](https://doi.org/10.1016/j.cognition.2021.104814)

2.1. Introduction

When bilinguals speak to one another, they often shift between their languages, producing utterances like “Can you get me *un café con leche y azúcar* [a coffee with milk and sugar]?” This flexible use of languages, or *code-switching*, occurs frequently in natural discourse (Poplack, 1980; Sebba et al., 2012) and serves a variety of functions (Auer, 1988; Gumperz, 1982; Heller, 2007). Bilingual speakers are known to code-switch individual words, entire sentences, and even large portions of their conversations (Grosjean, 2001; Heredia & Altarriba, 2001; Milroy & Gordon, 2008). This process of integrating multiple languages on-the-fly can appear seamless, and often results in very few errors or overt breakdowns in communication (Poplack, 1980). But these switches may not be truly effortless: many behavioral and neurocognitive studies find that comprehenders incur costs associated with switching languages such as longer reaction times in

lexical decision tasks and increases in the neural response to code-switched words relative to non-switched words (see van Hell et al., 2015).

It is tempting to interpret these effects as evidence that comprehenders must actively switch from one language to another and that this process takes additional time and effort. But similar data patterns emerge in studies that do not involve code-switching. For example, hearing low frequency words or improbable sentence continuations will elicit similar effects in monolingual contexts (e.g. Forster & Chambers, 1973; Kutas & Federmeier, 2011). This raises an intriguing alternative hypothesis: perhaps the costs found in code-switching studies are caused by encountering unexpected input rather than a discrete process triggered by switching languages. The present study uses a novel paradigm combining electroencephalography (EEG) and naturalistic listening to explore this question. In the remainder of this Introduction, we evaluate the evidence showing the costs of code-switching and their variability (Section 2.1.1). Then, we describe two alternative theories about those costs (Section 2.1.2). We then outline the predictions that these theories make regarding two ERP responses: the N400 and the *Late Positive Complex* or LPC (Section 2.1.3). Finally, we end by describing the goals of the present study and the benefits of our novel paradigm, the *Storytime* task (Section 2.1.4).

2.1.1. The cost of code-switching in bilingual comprehension: Evidence from ERPs

Over the past few decades, many researchers have studied the effects of code-switching on comprehension using EEG and *event-related potentials* (ERPs). ERPs are averaged electrical responses collected at the scalp and time-locked to the onset of a stimulus. ERPs vary systematically in their amplitudes, latencies, and/or scalp distributions, making them useful for characterizing when and to what degree different variables affect cognitive processes like language

comprehension (Kappenman & Luck, 2011; Luck, 2004, 2014). The ERP literature on code-switching largely reports a biphasic response to code-switched words in sentence contexts: an early negativity (e.g. an N400) followed by a *Late Positive Complex* or LPC (see Fernandez et al., 2019; van Hell et al., 2015, 2018 for discussion). There are a few studies that do not find this biphasic pattern—for example, studies in which participants have lower levels of proficiency or experience with the matrix language often find no early negativities (e.g. Ruigendijk et al., 2016; Zeller, 2020 for discussion). We return to these studies in the General Discussion. But for now, we will focus on the studies that show the biphasic pattern during sentence comprehension and briefly describe the variability of these ERP effects in the code-switching literature.

2.1.1.1. Switch-related negativities and their variability in the literature

Most ERP studies on code-switching find some kind of early negativity followed by a late positivity in response to code-switched words. We will refer to these effects as being *switch-related* merely because they occur in response to code-switched words. In using this term, we do not mean to imply that these switch-related effects reflect a process of language selection or language switching—in fact, we will be arguing that one of these effects (i.e. the earlier negativity) reflects the effects of lexical prediction rather than language switching per se.

In the code-switching literature, the most common switch-related negativity is the N400. The N400 response is a negative-going deflection in an ERP waveform that typically peaks around 400 ms post-stimulus onset. This response is argued to reflect how easily a word is accessed and/or integrated into its context—the larger the N400 amplitude, the greater the processing difficulty (Kutas & Federmeier, 2009, 2011; Van Petten, 1993). Switch-related N400 effects have been taken

as evidence that switching between languages is costly and disrupts lexical processing (Alvarez et al., 2003; Grainger & Holcomb, 2009; van Hell et al., 2015, 2018).

Some of the earliest studies to find N400 effects used the aptly named *switch-task* paradigm in which a series of individual words are presented back-to-back and categorized by bilingual participants. In these tasks, “code-switching” occurs when a participant sees a word in one language and then a word in another language on the next trial. These studies find that switching languages between trials elicits increased N400 responses relative to non-switched trials (Alvarez et al., 2003; Midgley et al., 2009). In recent EEG studies, researchers have relied on more naturalistic paradigms to study code-switching. For example, some studies use single-sentence contexts or written discourses with intra-sentential code-switching, which is when one or more words are switched within a single utterance. The findings from these sentence comprehension studies largely support those from the earlier switch-tasks: there is an increased N400 response to code-switched words relative to the non-switched words (see Fernandez et al., 2019; van Hell et al., 2018).

One tricky aspect of the ERP literature on code-switching is that not *every* study finds an early negativity with the latency and scalp distribution of a canonical N400. For example, in a foundational study by Moreno et al. (2002), English-Spanish bilinguals read a mix of regular and idiomatic sentences (i.e. well-known proverbs). Their critical manipulation always occurred on the sentence-final word, which was either the expected, within-language word, its translation equivalent, or an unexpected, within-language word (see examples below).

(1) a. **Idiomatic sentences:** “Out of sight, out of...mind / brain / *mente* [mind].”

b. **Regular sentences:** “Each night the campers build a...fire / blaze / *fuego* [fire].”

In idiomatic sentences (1a), they found an LPC but no switch-related negativity—possibly because of the predictability of their idioms. We will return to this finding in the General Discussion, but for now, we focus on the results for the regular sentences (1b). In these more standard sentences, the authors found a left-lateralized negativity between 250 and 450 ms and an LPC between 450 and 850 ms in response to code-switched words (e.g. *fuego*). This early negativity was equivalent in magnitude to the canonical N400 elicited by their unexpected, within-language words (e.g. *blaze*). However, the left-frontal skew of this effect led the authors to interpret it as a *Left Anterior Negativity* (LAN) instead. Historically, LANs have been associated with increased demands on working memory (King & Kutas, 1995; Kluender & Kutas, 1993) and/or difficulties with morphosyntactic processing (e.g. Friederici, 2002; Gunter et al., 2000; Neville et al., 1991). More recently, Ng et al. (2014) found a similar biphasic LAN-LPC pattern in response to code-switched words. In this study, Spanish-English bilinguals read short stories in English. Throughout the stories, some nouns and verbs were occasionally code-switched into Spanish (e.g. “The wind and the **sol** (*sun*) were disputing which was the stronger. Suddenly they **miraron** (*saw*) a traveler coming down the street....”). They report a LAN (350–450 ms) and an LPC (500–900 ms) to both code-switched nouns and verbs.

Taken together, one interpretation of these two studies is that the switch-related LAN and the switch-related N400 are functionally distinct, and thus the cognitive processes invoked in the studies that find LANs and in the studies that find N400s are systematically different (see van Hell et al., 2018 for discussion). The burden of such an account would be to explain why the processes invoked by code-switches vary across studies and to identify replicable means of producing each distinct data pattern. In the code-switching ERP literature, there is little systematicity in the types

of stimuli that elicit LAN vs. N400 effects. For example, studies using sentence-final code-switches have found LANs, N400s, and sometimes both effects overlapping with one another (for LANs, see Moreno et al., 2002; for N400 effects, see Proverbio et al., 2004; Van Der Meij et al., 2011 with low proficiency bilinguals; FitzPatrick & Indefrey, 2014; Ruigendijk et al., 2016; Zeller et al., 2016; for both, see Van Der Meij et al., 2011 with high proficiency bilinguals). In fact, when we look beyond code-switching to the broader psycholinguistic ERP literature, we see similar variation in the LAN and N400 effects elicited by unexpected and/or ungrammatical lexical items (see Bornkessel-Schlesewsky & Schlewsky, 2019; Caffarra et al., 2019; Fromont et al., 2020; Molinaro et al., 2011, 2015; Royle et al., 2013; Steinhauer & Drury, 2012; Tanner, 2015 for discussions of this debate). However, we will postpone discussion of this debate to the General Discussion.

We have observed three treatments of switch-related negativities in the ERP literature. Some authors treat LAN and N400 effects (and other negativities) as categorically distinct, invoking a difference in their function when interpreting their findings (e.g. Moreno et al., 2002; Ng et al., 2014; see van Hell et al., 2018 for discussion). Some authors do not make strong predictions about which negativity they will find, conducting analyses consistent with both the N400 and the LAN (e.g. Kaan et al., 2020). Finally, some authors note when their effects do not have the canonical distribution of N400s but nonetheless interpret these effects as reflecting the processes underlying the N400 (e.g. van Hell et al., 2015; van Hell & Witteman, 2009).

Adding to the complexity of the prior literature, we note that the LAN and the N400 are not the only early negativities observed in code-switching experiments. Other studies have interpreted their switch-related negativities as being *Phonological Mismatch Negativities* (PMNs, see Liao & Chan, 2016), N1 effects (e.g. Proverbio et al., 2002, 2004), *N200 effects* (Khamis-Dakwar &

Froud, 2007), *left-occipital N250 effects* (e.g. Van Der Meij et al., 2011), *fronto-central negativities* (Hut & Leminen, 2017), “*broad*” *negativities* (Zeller, 2020), and finally *anterior negativities* (ANs, Litcofsky & Van Hell, 2017 for second code-switched word; Zeller, 2020). All of these effects appear in addition to or in place of canonical N400 effects, varying in their precise timings (starting as early as 130 ms and lasting as late as 900 ms post-stimulus onset) and in their scalp distributions (ranging from left anterior, bilateral, fronto-central, to widespread). However, these effects also show clear commonalities: Most of them take place, at least in part, during the typical N400 time window, and most show a scalp distribution that at least overlaps with the canonical N400 distribution. Furthermore, some of the variation in these effects could potentially be explained by differences in the presentation modality and the speed of language processing in a given population or during a particular task.

In the present paper, we have adopted the working hypothesis that the various switch-related negativities reflect a common underlying process (or set of processes)—and that this process is the same one that underlies the classic N400 effects. We do this both for ease of explanation and because we believe that it is the most parsimonious explanation given the existing data. On this hypothesis, the variation in latency and distribution would be attributed to the following: differences in processing speed due to features of the stimuli or the participants; differences in modality; differences in predictability; and differences in the other processes that are occurring within the same time window (see Moreno et al., 2002, 2008; Ng et al., 2014; Van Der Meij et al., 2011; Zeller, 2020 for similar interpretations). The challenge for such a hypothesis is to account for this variability and to make testable predictions. We return to this question in the General Discussion. Critically, our findings (and the validity of the experiment) do not depend on whether this working hypothesis is true. While our primary analysis will focus on the canonical N400 time

window and electrode sites, we will also conduct exploratory analyses that investigate the precise distribution and timing of our effects.

2.1.1.2. Switch-related LPCs and their variability in the literature

The second type of switch-related ERP components is the LPC, which peaks around 600 ms post-stimulus onset (over posterior electrode sites) and is argued to occur after initial lexical processing, i.e., after the N400/LAN (e.g. Fernandez et al., 2019; Moreno et al., 2002, 2008; Ng et al., 2014; Proverbio et al., 2004; van Hell et al., 2015, 2018; van Hell & Witteman, 2009). These late-emerging, long-lasting positivities often occur in response to code-switched words in sentence contexts, but they can be found in a variety of other linguistic (and non-linguistic) tasks (see Kuperberg et al., 2020; Van Petten & Luka, 2012). The precise interpretation of LPCs is still debated, but there seems to be agreement that they reflect the recognition of a high-level discrepancy (e.g. a language shift, a syntactic error, an unexpected event) and the reevaluation of the input to make sense of this unexpected event (Coulson et al., 1998; Friederici, 2005; Hagoort, 1993; Hahne & Friederici, 1999; Kaan et al., 2000; Kolk & Chwilla, 2007; Kuperberg, 2007; Kuperberg et al., 2020; Litcofsky & van Hell, 2017; Osterhout & Holcomb, 1992; Tanner et al., 2017). On this interpretation, switch-related LPC effects would reflect the recognition of the language switch (Moreno et al., 2002) and the costs associated with integrating the new language into the discourse (van Hell et al., 2018).

The switch-related LPC has been observed many times alongside switch-related negativities. There is, however, variation in the size of these effects (and when they occur) that seems to be related to factors like the predictability of the switch (FitzPatrick & Indefrey, 2014; Moreno et al., 2008; van Hell et al., 2018), the switching direction (Fernandez et al., 2019; Liao & Chan, 2016;

Litcofsky & van Hell, 2017), the participants' language proficiency (Alvarez et al., 2003; Moreno et al., 2002; Ruigendijk et al., 2016; Van Der Meij et al., 2011), and their experience with code-switching (e.g. Proverbio et al., 2004). We return to this variability in the General Discussion.

In sum, the ERP literature on code-switching provides strong, converging evidence that comprehenders are sensitive to an unforeseen shift in the language being used and experience some processing difficulties. The question addressed in the present study is whether these difficulties are specific to switching languages or whether they are simply an indirect consequence of processing an unexpected word.

2.1.2. Two theories about the costs of code-switching

Broadly speaking, there are two ways in which we could imagine bilinguals tackling the task of understanding multilingual utterances. First, language identification could precede lexical processing: a bilingual could initially determine which language is being spoken (perhaps on the basis of phonetic features, see Caramazza & Brones, 1979; Dijkstra, 2005; Dijkstra & Van Heuven, 2002; Grainger & Beauvillain, 1987; Grainger & Dijkstra, 1992; Macnamara & Kushnir, 1971; Scarborough et al., 1984; Soares & Grosjean, 1984; van Hell & Tanner, 2012). Then, they could switch into that language, and find the relevant word. On this account, code-switching costs would arise from the need to switch languages prior to accessing the code-switched word (Alvarez et al., 2003; Bultena et al., 2015; Grainger & Holcomb, 2009; Green, 1998). Second, lexical access could occur prior to (or independent of) recognizing the language of the word being processed: a bilingual could simultaneously map the sounds they hear (or the letters/signs they see) onto lexical forms in both of the languages that they know (Dijkstra et al., 1999; Duyck et al., 2007; van Hell & de Groot, 1998). On this second account, language identification might only be achieved after

the word is accessed and recognized as belonging to a particular lexicon (Dijkstra & Van Heuven, 2002; Moreno et al., 2002)—in fact, one could imagine a comprehension system in which the listener never *actively* recognized which language the word was in.

The evidence to date favors this second theory of bilingual comprehension. Many studies show that bilinguals simultaneously activate words from two languages as a spoken word unfolds. For example, Russian-English bilinguals hearing the sounds “*shar...*” will activate both the English word *shark* and the Russian word *sharik* (balloon), as both words match the initial phonemes that they heard (Marian & Spivey, 2003). The fact that bilinguals initially entertain both words and then arrive at the correct word after phonological disambiguation suggests that it is not necessary to distinguish between different lexicons during comprehension (e.g. Grainger & Beauvillain, 1987; Hartsuiker et al., 2004; Kroll et al., 2012; Li, 1996; Loebell & Bock, 2003; Marian & Spivey, 2003; Spivey & Marian, 1999). At first glance, this model of bilingualism is hard to reconcile with the studies that find costs to code-switching. If you do not need to switch from one lexicon to another before accessing a word, why would code-switched words be processed more slowly or effortfully? We see two alternative explanations for these code-switching effects, both of which come from an expectation-based framework for language comprehension (see Hale, 2001; Levy, 2008; Pickering & Gambi, 2018; Pickering & Garrod, 2013; Tanenhaus et al., 1995; Venhuizen et al., 2019).

One type of code-switching effect could result from later, post-lexical processes that occur after the listener realizes that the speaker has switched from one language to another. In EEG, we might expect these post-lexical effects to be indexed by the LPCs because they arise later than components linked to the processing of lexical forms and meanings (e.g. Brothers et al., 2015;

Grainger et al., 2006; Holcomb & Grainger, 2006; Lau et al., 2013; see Nieuwland, 2019 for review of form-based components).

A second type of code-switching effect might occur—not because the listener switches from one lexicon to another—but rather because they have made a specific prediction about the form (i.e. the language) of the word they are about to hear, and that prediction is violated when they hear a code-switched word instead. On this account, code-switched words are no different than any other unexpected word (cf. Moreno et al., 2002). This hypothesis is consistent with our current understanding of the N400. The N400 was first discovered in contexts where there is a highly predictable word that is replaced with an unexpected word, e.g., “He spread the warm bread with...*socks*” (Kutas & Hillyard, 1980). Subsequent studies have demonstrated that the magnitude of the N400 varies continuously with the predictability of a word (Borovsky, Elman, & Kutas, 2012; Brown & Hagoort, 1993; Federmeier & Kutas, 1999; Fernandez et al., 2019; Kutas & Federmeier, 2000; Kutas & Hillyard, 1984; Lau et al., 2009; Lau et al., 2013; Van Berkum et al., 1999). N400s also decrease for a given word as the cumulative contextual constraints make that word increasingly likely (Van Petten, 1993). This pattern is compatible with a framework in which top-down processes generate predictions about upcoming words (Altmann & Kamide, 1999; Kamide et al., 2003), making it easier to access the meanings of words that are consistent with those predictions (Federmeier, 2007; Hale, 2001; Kuperberg, 2016; Kuperberg et al., 2020; Levy, 2008; Schwanenflugel & LaCount, 1988). Some studies reveal that the N400 is sensitive to expectations that are linked to meaning (or semantic features) of the word (Federmeier et al., 2002; Federmeier & Kutas, 1999; Kuperberg, 2007; Kuperberg et al., 2020). Other studies have also found evidence that these expectations can lead to the pre-activation of syntactic or phonological features of the word (DeLong et al., 2005, 2017; Van Berkum et al., 2005; Wicha, Bates, et al.,

2003; Wicha et al., 2004; Wicha, Moreno, et al., 2003; but see, Ito et al., 2017a; Nieuwland, 2019). If top-down processing generates an expectation for a particular lexical item within a particular language (rather than just the concept encoded in that word), then we would expect to get N400 effects to code-switched words purely as a side effect of these lexical predictions. In the next section, we consider both hypotheses (i.e. discrete switch costs vs. general expectation-based costs), and then discuss an experimental manipulation designed to tease them apart.

2.1.3. Testing the one-cost and two-cost hypotheses

The present study tests two alternative theories for the N400 effects in code-switching studies. The first theory claims that language identification or recognition (e.g. English or Spanish) precedes lexical processing. On this hypothesis, the N400 to code-switched words reflects the cost of switching from one lexicon to another, while the N400 to words that do not fit the context reflects a slowdown in lexical access in the absence of contextual clues. We will call this the *two-cost* hypothesis. The second theory claims that lexical processing occurs prior to and independent of recognizing the language of the word being processed. On this hypothesis, the N400 to code-switched words and the N400 to words that do not fit the context have the same root cause. In both cases, the listener hears a word that is inconsistent with their expectations, and thus that word is more difficult to access. We call this the *one-cost* hypothesis.

These accounts make different predictions about what would happen if a bilingual encountered a code-switched word *that was also a poor fit for the context*. The two-cost account predicts additivity in the N400 response, such that the size of the N400 would be roughly equivalent to the sum of the N400 effects for the two separate violations. Critically, the one-cost account predicts that the N400 response to the double violation should be roughly the same as the

cost for either the (correct) code-switched word or the poorly fitting word in the same language—as in all three cases, the expected word did not appear. The present study tests whether these two N400 effects are additive or whether all three violations have the same N400 effect.

There are two studies with data that bear on these hypotheses: Both Liao and Chan (2016) and FitzPatrick and Indefrey (2014) conducted experiments with 2×2 manipulations of the language of the target word (switched vs. non-switched) and how well the word fits into the preceding sentence context. These factorial designs provide critical data that could test the question of additivity—however, neither group considered their data in light of the two hypotheses above. Instead, they arrived at very different conclusions, perhaps because they set out to test different questions or used different interpretive frameworks to understand their data.

In the first relevant study, Liao and Chan (2016) had Mandarin-Taiwanese bilinguals listen to sentences that were played word-by-word with 200 ms pauses between each word. In addition to manipulating the presence/absence of a code-switch and the contextual fit of their target words, they also manipulated the direction of the language switching, i.e., switching from the participants' dominant language into their weaker one or vice versa. The authors concluded that the costs associated with code-switching are greater in cases of dominant-to-weaker language switching (as in the present study). When collapsing across switching direction, the authors find an interaction between contextual fit and code-switching in an early negative component—namely, the Phonological Mismatch Negativity, which emerged between 250–350 ms. This interaction is consistent with the one-cost hypothesis, as the negativities are similar across all three violation conditions and significantly different from the baseline condition (i.e. the expected word in the expected language).

Three features of this study, however, prevent us from drawing strong conclusions with respect to our current hypotheses: 1) The use of word-by-word auditory presentation may have resulted in processing strategies that are different from those in a more naturalistic listening task, perhaps because it leaves more time for prediction and sub-articulation; 2) The interaction was found on a component that is not typically observed in code-switching studies, possibly due to the presentation method. Thus, it is unclear whether or not this finding would generalize to the N400 and LAN effects, which are the most prevalent initial effects of code-switching in the ERP literature; 3) Because the pattern of effects is radically different across the two switching directions, it is difficult to know how to interpret findings that appear when you collapse across them. Taken together, we cannot tell (from the data presented) whether the one-cost data pattern is reliably present in either switching direction.

In the second relevant study, FitzPatrick and Indefrey (2014) had Dutch-English bilinguals listen to sentences in either their native language (Dutch) or their non-native language (English). The authors' central goal was to explore bilinguals' comprehension of interlingual homophones (e.g. *pet* can mean *an animal companion* in English or *a cap-like hat* in Dutch). However, they also conducted two experiments (one in English, one in Dutch) with a 2×2 manipulation of code-switching and semantic congruence. In this study, the authors pursued a different analytic approach. Rather than directly comparing the three violation conditions to one another, they focused on the presence and the timing of the semantic congruity effects within each language separately. Their conclusions are broadly consistent with the two-cost hypothesis. Specifically, they propose that there are semantic incongruity effects for all incongruous words (regardless of language) and that these effects emerge earlier for non-switched words because they can be accessed more easily. They also propose that there is an early transient negativity for congruous

code-switched words due to the priority given to the matrix language during lexical access. But curiously, their data patterns also seem consistent with the one-cost hypothesis: there is an interaction in the early N400 time window due to the fact that the effects in the three violation conditions appear to be similar in magnitude, timing, and scalp distribution. However, comparisons across these conditions are hindered by the fact that the sentence frames are different for each condition.

In sum, the studies to date do not provide conclusive evidence for either the one-cost or the two-cost hypothesis, largely because these studies did not originally set out to address these particular hypotheses. The two studies above reached divergent conclusions, despite having broadly similar data patterns—an issue that we will return to in the General Discussion.

2.1.4. The present study

The present study asks whether the neural markers associated with code-switching costs are best understood as difficulties specific to switching languages or as indirect consequences of processing unexpected words or encountering unexpected events. To test this question, we systematically compared bilinguals' ERP responses to three types of words: code-switched words, unexpected (within-language) words, and double violations (code-switched words that weakly fit the context). If there are unique costs to code-switching, we should expect an additive N400 response to the double violation condition. This study differs from most of the prior ERP studies on code-switching in one critical way. Many of the prior studies have used artificial tasks (e.g. lexical decisions, naming tasks) with unnatural code-switches (i.e. switches in formal, written texts or predictable sentences with the last word switched). Recently, many observers have noted that these artificial contexts are not the most accurate way to study how bilinguals understand code-

switching in the wild (Blanco-Elorrieta & Pykkänen, 2016, 2017, 2018; Fernandez et al., 2019; van Hell et al., 2015, 2018). The present study takes a critical step toward a more naturalistic approach with our *Storytime* task. In our paradigm, participants listen to real, unscripted, spoken narratives. We took these narratives and spliced in a carefully counterbalanced experimental manipulation. This allowed us to precisely study intra-sentential code-switching in a rich, variable, naturalistic context. To preview our findings, we were able to successfully replicate the N400 and LPC effects found in the prior literature using our design. And critically, this design allowed us to test whether the processing costs of language and contextual fit were additive.

2.2. Method

2.2.1. Participants

Thirty-four Spanish-English bilinguals from Harvard University participated in this experiment. We excluded two participants due to experimenter error, resulting in 32 participants in the final sample. We did not perform an a priori power analysis—rather we based our final sample size on the prior literature. During data collection, however, we implemented a stopping rule: participants' EEG data were cleaned incrementally and replaced if more than 25% of all trials were rejected (see rejection criteria in Section 2.2.4.1). This procedure continued until we had usable data from 32 participants.

We recruited participants from student-run organizations and from Harvard's study pool. Participants were compensated \$10/hour or received two study credits for participating. We screened participants for eligibility by asking them six questions about their proficiency in Spanish and their overall exposure to code-switching in their community. This language screener is accessible on the Open Science Framework (OSF; see <https://osf.io/jwqpr/>). Participants self-

reported that they were highly proficient in Spanish (*intermediate-level* = 1, *advanced-level* = 3, *native-level* = 28) and had considerable exposure to code-switching (*never heard code-switching* = 1, *sometimes* = 9, *often* = 22). All participants reported learning Spanish before age eight with the average age of acquisition being 0;11 ($SD = 1;11$).

We did not intend for this initial screener to be a robust language history survey—thus, after the experiment, we recontacted all of our original participants to have them complete the Language Experience and Proficiency (LEAP) Questionnaire (Kaushanskaya et al., 2020; Marian et al., 2007). Roughly two-thirds of our study population agreed to participate and completed the LEAP Questionnaire (22 out of 32 participants). After receiving the additional information from the LEAP Questionnaire, we determined that our population was dominant in English and considered Spanish to be their first language. Although participants considered themselves to be largely dominant in English, the levels of proficiency across both languages were comparable with slightly more variability in the proficiency for Spanish (see Table 2.1). The results from the LEAP Questionnaire are largely consistent with our findings from the initial screener—more information can be found on OSF (<https://osf.io/jwqpr/>).

Table 2.1: Participants' responses from the LEAP Questionnaire.

Language History Measures from the LEAP Questionnaire		
<i>Language Proficiency Assessment</i>	Spanish	English
Average age of acquisition	0;8 (0;4)	3;9 (0;8)
Average age of reaching fluency (speaking)	6;0 (1;0)	6;10 (0;11)
Average age of reaching fluency (reading)	8;5 (1;1)	7;7 (0;10)
Percentage of time exposed to the language (out of 100%)	30.5 (3.4)	69.3 (3.3)
Average composite score of language exposure ¹ (highest score = 10)	4.3 (0.3)	7.1 (0.3)
Percentage of time choosing to speak the language (out of 100%)	32.6 (4.5)	67.3 (4.5)
Self-rated proficiency in speaking the language (out of 10)	8.6 (0.3)	9.8 (0.1)
Self-rated proficiency in understanding the spoken language (out of 10)	9.1 (0.2)	9.8 (0.1)
Self-rated proficiency in reading the language (out of 10)	8.0 (0.1)	9.8 (0.1)
Average composite of language proficiency ² (highest score = 10)	8.6 (0.2)	9.8 (0.1)
<i>Language Dominance Assessment</i>	Spanish	English
Number of participants listing this language as dominant (out of 22) ³	2	20
Number of participants listing this language as first acquired (out of 22)	19	3
<p>Notes. Means and standard errors are reported for both Spanish and English separately. All ages are reported as years followed by months.¹ The composite scores for <i>language exposure</i> were created by averaging self-reported ratings (out of 10) of how often participants are currently exposed to Spanish/English when 1) interacting with friends, 2) interacting with family, 3) reading, 4) language apps or websites, 5) watching TV, and 6) listening to radio/music.² The composite scores for <i>language proficiency</i> were created by averaging the self-reported scores (out of 10) for speaking, understanding, and reading in each language.³ We were only able to get LEAP responses from 22 of the original 32 participants.</p>		

2.2.2. Stimuli

To preview our stimuli, we manipulated 120 target words in sentences within two oral stories, which were largely in English. There were two factors in our design: 1) how well does a word fit

into its preceding context putting aside its language (strong-fit, weak-fit) and 2) what language is the word in (English, Spanish). Below is an example of a target sentence in all four conditions:

- (2) a. And the wig itself is so hot and heavy on my **head**. (Strong-fit English)
- b. And the wig itself is so hot and heavy on my **cabeza**. (Strong-fit Spanish)
- c. And the wig itself is so hot and heavy on my *cranium*. (Weak-fit English)
- d. And the wig itself is so hot and heavy on my *cráneo*. (Weak-fit Spanish)

Thus, the Spanish conditions involved code-switching while the English conditions did not. Note, all Spanish target words were translations of either the English strong or weak-fit conditions (i.e. **head-cabeza**, *cranium-cráneo*). In the sections below, we explain how these stimuli were created.

2.2.2.1. Oral story selection

We selected two stories from a collection of unscripted, oral performances known as Moth stories. The Moth is a non-profit organization dedicated to “the art and craft of storytelling” (see <https://themoth.org/>). These ‘Moth’ stories contain properties of naturalistic speech (e.g. disfluencies, redundancies, colloquialisms) that are typically absent in more formally scripted performances. The stories that we selected were originally performed in English (roughly 20 minutes each) and had a combined total of 343 sentences. From these stories, we selected 120 target sentences (described below), resulting in an approximate 2:1 filler-to-target sentence ratio.

2.2.2.2. Target sentence and English noun selection

Words that are preceded by a supportive context are processed more easily, as indexed by faster behavioral responses and reduced N400 responses (Ehrlich & Rayner, 1981; Fischler & Bloom, 1979; Jordan & Thomas, 2002; Kutas & Federmeier, 2011). Thus, for our experiment, we wanted the original target words to be as predictable (and as easy to process) as possible. One limitation to using naturally produced narratives is that we were not able to create highly predictable sentence contexts—so, we settled on selecting the most predictable nouns available in the Moth stories. To characterize the predictability of the nouns in these stories, we conducted two cloze tasks: one using written versions of our stories and the other using our final auditory stimuli (see Taylor, 1953 for information on cloze tasks). Both of these cloze tasks were created on the IbexFarm experimental software (<http://spellout.net/ibexfarm/>) and made available to participants on Amazon’s Mechanical Turk (<https://www.mturk.com>).

The first cloze task was designed to collect the cloze probabilities for *all* of the nouns in both stories. To do this, 72 participants read one of the two stories from beginning to end. Participants saw short, fragmented sentences that ended right before a noun. They were then asked to guess the next word. After guessing, they were shown the actual noun, and this procedure continued until participants had guessed every noun in the story. We then calculated each noun’s cloze probability or the proportion of times that participants provided the target noun given its context. Cloze probability is argued to be a good measure of how easily a word can be predicted during language comprehension (Federmeier & Kutas, 1999; Staub et al., 2015), and it is inversely correlated with a word’s N400 amplitude (Kutas & Hillyard, 1984). Based on these data, we identified the most predictable nouns by sorting the cloze probabilities and taking the top 120 targets. Occasionally, one noun (e.g. head) had high cloze probabilities in multiple sentences (i.e. one noun type had

multiple high-cloze tokens); however, we never used the same target noun more than three times (and never more than twice in a single story). These 120 nouns became the strong-fit English target words, and they had an average cloze probability of 61% with a range of 13–98%.

Given this range of cloze values, we designed a second cloze task to characterize how predictable our target words were in the final recordings that we used in our EEG study. Note, in Section 2.2.2.7 below, we describe how these final recordings were created. In this audio cloze task, 45 participants heard both stories in their entirety, and we counterbalanced which story was played first. The recordings would pause right before each target word, and participants would then guess the next word. After guessing, the recording would rewind to the start of that target sentence, so that the participants could hear the actual story continuation. At the end of the task, participants were asked a series of questions to determine their level of engagement and overall comprehension during the task. Results indicated that participants understood the stories, as their comprehension accuracy was 91.7% ($SE = 2.1\%$). Similar to the written cloze task, the target words had an average cloze probability of 61.3% ($SE = 2.3\%$); however, this time the range was slightly wider with values ranging from 2.8–94.3%. In the General Discussion, we will address the implications of having a wide range of cloze values for our targets. The full list of target nouns and their cloze values can be found on OSF (see <https://osf.io/jwqpr/>).

2.2.2.3. *Weak-fit English noun selection*

Next, we selected the weak-fit English target nouns. We wanted the strong and weak-fitting pairs to be semantically related to one another. To do this, we took the same sentences that we identified above and replaced the high cloze noun with a noun that still made sense in that context—but had never been used by our MTurk norming sample in the written cloze task (e.g.

And the wig itself is so hot and heavy on my *cranium*). Thus, these weaker alternatives had a cloze probability value of roughly zero given our target contexts—although, this does not imply that these words could *never* be used in our sentences. Critically, the weak-fit nouns expressed events that were plausible and did not disrupt the overall storyline (e.g. It had a *mind/brain* of its own; I put it on one of my dresser *drawers/shelves*; My hair does fall out, first in these strands in my brush, and then in clumps in my shower *drain/hole*).

After creating the strong and weak-fit pairs, we quantified the semantic relatedness between them by calculating their cosine similarities. We used the *LSAfun* package (Günther et al., 2015) in the R statistical computing environment (R Core Team, 2020). To do this, we first selected a semantic space in which each target word is represented as a single vector—for our analyses, we used the semantic space from Baroni et al. (2014). Then, we measured the cosine of the angle between the vectors for each strong and weak-fit word pair. Cosine similarity values can range between -1 (highly dissimilar) and 1 (highly similar). A cosine value of zero indicates that the two words are orthogonal to one another (for more information, see Günther et al., 2015). Across all pairs, the average cosine similarity was .26, which is equivalent to the similarity between the words *dog* and *mouse* in this semantic space. The range of similarities was .02–.69, which is similar to the comparison of *dog* and *osprey* and then *dog* and *puppy* respectively.

Next, we decided to assess the fit of our words within our target sentences. To do this, we conducted a naturalness rating task on IbexFarm and Amazon’s Mechanical Turk. In this task, participants read all of the target sentences from one of the two stories. The sentences were presented in their entirety with either the strong-fit or weak-fit target. All target words were marked with asterisks (e.g. *head*). Participants were then instructed to rate the naturalness of the target word in the sentence using a 7-point Likert scale ($7 = \textit{Very Natural}$, $1 = \textit{Very Unnatural}$). An

unnatural word was described as a word that a person might have a hard time imagining someone saying in this context (and vice versa for a natural word). To determine any differences across conditions, we used a two-tailed paired-samples *t*-test. We found that our strong-fit targets were rated significantly higher ($M = 6.36$, $SE = 0.06$) than our weak-fit targets ($M = 3.28$, $SE = 0.08$), $t(129) = 32.19$, $p < .001$.

2.2.2.4. Spanish noun selection

In our design, English was the matrix language, which meant that the Spanish words served as the code-switched items, and the English words served as controls. We assumed that English would be the dominant language for the majority of our study population, as they all attended a university where the language of instruction is English. Results from the LEAP Questionnaire confirmed that this assumption was correct, as most participants said English was their dominant language. Prior studies have shown that bilinguals are better able to predict upcoming words when comprehending sentences in their dominant language (see Ito, 2016; Ito et al., 2017b, 2018; Ito & Pickering, 2021; Kotz & Elston-Güttler, 2004; Liao & Chan, 2016).

In the present study, the critical words in the Spanish conditions were translation equivalents of the strong-fit and weak-fit English targets. The translations were provided by the second author (EM), who is a native Spanish speaker. In some cases, the direct translation of an English target was a Spanish cognate (e.g. *ceremony* and *ceremonia*). The use of cognates was judged to be unavoidable, so we equated the number of cognates in the strong-fit and weak-fit conditions: 36 cognates per condition, 72 cognates in total out of 240 Spanish words. After testing, we realized that two of these cognates are considered to be variants (*subjeto* for *sujeto*) or are not formally accepted (*disturbia* for *disturbio*) according to the *Diccionario de la lengua española* (Dictionary

of the Spanish language). Given this finding and the number of cognates in our study, we conducted our primary analyses with and without these cognate items. However, removing the cognates did not change the overall pattern of findings. The results from the analyses without cognates are available in our annotated analyses on OSF (see <https://osf.io/jwqpr/>). Again, all target trials are listed on OSF.

2.2.2.5. Assessing our critical manipulations within our study population

As we mentioned above, we recontacted all of our original study participants and asked them to complete a language survey and a ratings task. In the ratings task, we asked participants to re-read the original stories and provide naturalness ratings for our strong and weak-fitting English words. To do this, participants read both stories in their entirety, chunk-by-chunk. Each chunk contained one English target word (either the strong or weak-fitting version). Participants then rated the naturalness of the word given the story context on a sliding scale from 0 (*Very Unnatural*) to 100 (*Very Natural*). After rating this target word, participants were presented with a potential Spanish translation of that word—half of these translations were the ones used in the ERP study while the other half were foils. The foils were simply the Spanish translations of other words from the trials that the participant did not see. Finally, participants rated these translations as being acceptable or unacceptable (or they indicated that they did not know the Spanish word, the English word, or both of the words presented). This procedure continued until participants had read both stories and rated all of the target words that they had encountered in the original ERP study.

We had 20 out of the original 32 participants complete this ratings task. Results indicated that participants consistently rated the strong-fit English words as being more natural ($M = 94.3$,

$SE = .9$) than their weak-fit English alternatives ($M = 35.9$, $SE = 2.0$). Participants also strongly accepted the translations that we used in the original study ($M = 90\%$ acceptable, $SE = 1.0\%$) and strongly rejected our foil translations ($M = 2.8\%$ acceptable, $SE = 1.0\%$). Looking at the code-switched conditions individually, we found that strong-fit and weak-fit Spanish translations were accepted 95.8% ($SE = 1.2\%$) and 83.5% ($SE = 2.2\%$) of the time respectively. Finally, participants knew nearly all of our target words: strong-fit English words ($M = 99.8\%$ known, $SE = 0.1\%$); strong-fit Spanish words ($M = 97.5\%$, $SE = 0.7\%$); weak-fit English words ($M = 98.4\%$, $SE = 0.4\%$); and weak-fit Spanish words ($M = 93.5\%$, $SE = 1.2\%$).

2.2.2.6. Other properties of the stimuli: word frequency, length, and sentence position

There are a few other stimulus properties that we did not consider when initially constructing our target words; however, they are still important to characterize. These properties are word frequency, word length, and the word's position in our target sentences—and we describe them in more detail below.

Word frequency. For our target words, we used the standardized word frequencies (per million words) from the SUBTLEX_{US} (Brysbaert & New, 2009) and the SUBTLEX_{ESP} (Cuetos et al., 2011) subtitle corpora. The SUBTLEX_{US} corpus has roughly 51 million words from American English subtitles (1990–2007). The SUBTLEX_{ESP} corpus has roughly 41 million words from Spanish subtitles (1990–2009) that contain both Iberic and Latin American language variants. These Spanish subtitles came from a range of Spanish-speaking countries such as Argentina, Chile, Colombia, Mexico, Peru, and Spain, as well as from the United States. To evaluate differences in word frequencies across conditions, we used a series of two-tailed, paired-samples *t*-tests (Bonferroni-corrected $\alpha = .01$). The two irregular cognates (mentioned above) did not have any

frequency values, so we included the frequency values for their accepted forms: *sujeto* and *disturbio*. When comparing the frequency of all English words ($M = 176.54$, $SE = 22.5$) to all Spanish words ($M = 166.98$, $SE = 18.7$), there was no significant difference in frequency, $t(239) = .33$, $p = .74$. However, there were pairwise differences between conditions, such that the strong-fit English words ($M = 328.72$, $SE = 40.30$) were more frequent than the weak-fit English words ($M = 23.11$, $SE = 4.32$), $t(119) = -7.70$, $p < .001$, and the strong-fit Spanish words ($M = 287.81$, $SE = 32.65$) were more frequent than the weak-fit Spanish words ($M = 43.67$, $SE = 9.48$), $t(119) = -7.28$, $p < .001$. There were no significant differences between the two strong-fit conditions, $t(119) = -1.47$, $p = .14$, nor the two weak-fit conditions, after correcting for multiple comparisons, $t(119) = 2.48$, $p = .015$.

Word length. Next, we compared the length of our target words. To do this, we calculated two measures of word length: the raw number of syllables and the duration (ms) of the words from our recordings. To evaluate any differences, we again used a series of two-tailed, paired-samples *t*-tests (Bonferroni-corrected $\alpha = .0125$). Unsurprisingly, there were significant differences in the number of syllables between all four conditions, reflecting the tendencies for both Spanish words and less frequent words to have more syllables in general: the strong-fit English words ($M = 1.40$, $SE = .06$) were shorter than the strong-fit Spanish words ($M = 2.63$, $SE = .09$), $t(119) = 14.49$, $p < .001$; the weak-fit English words ($M = 1.79$, $SE = .07$) were shorter than the weak-fit Spanish words ($M = 2.99$, $SE = .09$), $t(119) = 13.60$, $p < .001$; the strong-fit English words were shorter than the weak-fit English words, $t(119) = 4.97$, $p < .001$; and the strong-fit Spanish words were shorter than the weak-fit Spanish words, $t(119) = 2.94$, $p < .01$. For target word duration in milliseconds, the strong-fit English words ($M = 538.75$, $SE = 23.48$) were significantly shorter than both the weak-fit English words ($M = 617.24$, $SE = 24.15$), $t(119) = -4.67$, $p < .001$ and the strong-fit Spanish

words ($M = 647.17$, $SE = 25.51$), $t(119) = -6.25$, $p < .001$. However, there were no significant differences between the durations of the weak-fit Spanish words ($M = 700.99$, $SE = 27.10$) and the strong-fit Spanish words, $t(119) = -1.77$, $p = .078$, nor the two weak-fit conditions, $t(119) = -2.23$, $p = .027$ (after correcting for multiple comparisons).

Sentence position. Finally, we evaluated where our critical words appeared in each target sentence. Roughly 30% of all target words appeared at the end of the sentence—the rest of the targets appeared somewhere in the middle. To quantify the position of our target words, we calculated how many words preceded the targets and how much of the total sentence had been heard prior to the target. On average, there were 13 words prior to our critical words (range: 3–35 words) and roughly 70% of the entire sentence had been heard by the onset of our targets (range: 12.5–100%). Below, we have summarized all of the relevant properties of our naturalistic stimuli (see Table 2.2).

Table 2.2: *Critical properties of our experimental stimuli.*

Stimulus Properties across Experimental Conditions						
<i>Average cloze probabilities</i>		Written Cloze (N = 36) 61.0% Range: 13–98%			Auditory Cloze (N = 45) 61.3% Range: 3–94%	
<i>Average sentence positions</i>		Amount of prior context 70.6% Range: 13–100%			Number of prior words 13 words Range: 3–35 words	
Condition	SUBTLEX Frequency	Number of Syllables	Duration (ms)	Word Known ¹	Translation Acceptability	Naturalness (out of 100)
<i>Strong-fit English</i>	328.72 (40.3)	1.40 (0.1)	538.7 (11.7)	99.8% (0.1)	--	94.3 (0.9)
<i>Weak-fit English</i>	23.11 (4.3)	1.79 (0.1)	647.2 (12.7)	98.4% (0.4)	--	35.9 (0.2)
<i>Strong-fit Spanish</i>	287.81 (32.6)	2.63 (0.1)	617.2 (12.1)	97.5% (0.7)	95.8% (1.2)	--
<i>Weak-fit Spanish</i>	43.67 (9.4)	2.99 (0.1)	701.0 (13.5)	93.5% (1.2)	83.5% (2.2)	--
<p>Notes. *Averages are presented with standard errors in parentheses. ¹Word Known variable represents the percentage of words recognized by bilinguals in each condition.</p>						

2.2.2.7. Audio stimulus creation

After selecting our target nouns and sentences, we created the materials for our *Storytime* paradigm. First, we recorded the two stories in their entirety, making an effort to preserve the disfluencies and redundancies from the original Moth performances. We then recorded each of the target sentences individually in each of the four conditions. Next, we spliced all target words from these individual recordings into the larger story recordings, which allowed us to keep the audio before and after the targets identical across conditions. To avoid splicing artifacts, we respected co-articulation by splicing in, at most, the word before and after the target. Finally, we found all

of the target onset times manually using the phonetic software, PRAAT (Boersma & Weenink, 2001), which we later used to time-lock our ERP responses.

Since we used these onset times in our final analyses, we wanted to ensure their accuracy. First, all onset times were determined by the second author (EM), who is a native speaker of Spanish. Second, these onset times were confirmed by the first author (AY), who is a native speaker of English. Third, we transcribed each target word into IPA and considered how the phonetic differences across the two languages could cause differences in onset times (e.g. the lack of aspiration for word-initial plosives in Spanish). Finally, we relied on visual cues like formant transitions, changes in pitch, frequency, and/or intensity, as well as our own intuitions when marking word boundaries. An example of a specific time-locked stimulus, and all of our phonetic transcriptions can be found on OSF (see <https://osf.io/jwqpr/>).

We used a Latin Square design with four lists. Targets were assigned to item groups in such a way that in any given list no adjacent items were ever in the same condition. In our *Storytime* paradigm, there is no traditional trial structure, meaning that all intervening sentences serve as fillers. Some of our target sentences occurred back-to-back, whereas others occurred with 16 sentences in between. On average, the number of intervening sentences was 1.73 ($SE = .23$). This number may make it seem like our target words were presented in rapid succession; however, our sentences varied widely in their total durations. Thus, a more accurate characterization of the timing between target trials is the *interstimulus interval* (ISI), which represents the time between the offset of one target word and the onset of the next. The average ISI in our recordings was 17.8s ($SE = 1.35$) with a range of 1.8 to 74.3 seconds.

Finally, we wanted to characterize the speech rate and the average *stimulus onset asynchrony* (SOA) of our recordings. These two properties are often tightly controlled in traditional

psycholinguistic experiments—thus, we calculated these metrics for ease of comparison. To calculate the speech rate, we used a PRAAT script written by de Jong and Wempe (2009), which finds the nucleus of each syllable in a recording and uses that information to determine metrics like speech rate, phonation time, and average syllable duration automatically. The average speech rate across recordings was 2.84 ($SE = .02$) syllables per second. Next, we calculated the average SOA (i.e. the time between the onset of a target word and the onset of the next target word). To do this, we obtained the onset times for each word in our recordings using the Gentle forced aligner from Ochshorn and Hawkins (2017). Then, we calculated the SOA values for each word, removing the values for all of the target words (as they differed across story versions) and all of the function words (as they are extremely short and would skew the overall average). Results indicated that, on average, the SOA in our stories was 521.9 ms ($SE = 8.0$).

2.2.3. Procedure

In the present study, participants passively listened to two short stories during a single EEG recording session. This naturalistic listening technique circumvents typical difficulties associated with traditional experimental designs, i.e., the need for many disjointed, out-of-the-blue sentences, and an extensive use of filler sentences. This technique also allows participants to hear rich discourses, promotes attention, and is arguably more engaging. Each story lasted about 20 minutes, and there was a short break in between them. The order of story presentations was counterbalanced across participants. We intended to test the same number of participants in each list, but one participant was run in the wrong list, resulting in a slightly uneven number (i.e. 7, 9, 8, and 8 per list). All participants sat approximately 40 inches in front of a TV monitor, which displayed an unrelated video of a beach sunset (available on OSF, <https://osf.io/jwqpr/>). In this video, the sun

was slowly moving along a vertical axis in the center of the display. This video served as a focal point for participants and helped minimize sharp horizontal and vertical eye movements throughout the study. We encouraged participants to blink as little as possible and to reduce facial tension (i.e. keep their forehead and jaw relaxed). At the end of the study, participants were fully debriefed and given the opportunity to ask any questions. All of our experimental procedures were approved by the Harvard Committee on the Use of Human Subjects (CUHS).

2.2.4. EEG Recording

We recorded the electroencephalogram using Brainvision's actiChamp System. Online signals were recorded from 31 active Ag/AgCl electrodes embedded in an elastic cap (EASYCAP GmbH). The ground and reference electrodes were the pre-frontal electrodes FPz and FP1 respectively. A pair of passive EOG electrodes connected to the BIP2AUX adapter was attached above and below the left eye to monitor for vertical eye movements. We continuously recorded at a sampling rate of 500Hz and kept electrode impedances below 20 k Ω .

2.2.4.1. EEG Pre-processing

We pre-processed and analyzed our EEG data using both EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon, & Luck, 2014) toolboxes. First, we downsampled the data to 200 Hz and re-referenced offline to the average of the left and right mastoids. The EEG signals were then filtered using an IIR filter with a bandwidth of 0.01–30 Hz. We then identified and corrected eye blink artifacts using an Independent Component Analysis (ICA). Next, we created epochs that extended from 200 ms before stimulus onset to 2000 ms post-stimulus onset. We then performed a two-step artifact rejection process: First, we subjected all epochs to an automatic

rejection procedure that removed trials with voltages exceeding -90 or $90 \mu\text{V}$. This procedure rejected 15.2% of all 3,840 trials (30 trials per condition \times 32 electrodes = 3,840 observations). No condition had more than 17% of their trials rejected: strong-fit English = 16.25% ($SE = .02\%$); strong-fit Spanish = 16.67% ($SE = .02\%$); weak-fit English = 14.06% ($SE = .01\%$); weak-fit Spanish = 13.75% ($SE = .02\%$). Using a generalized logistics mixed model, we confirmed that there were no statistical differences in rejection rates between the two strong-fit conditions ($b = .03$, $SE = .13$, $z = .27$, $p = .79$), the two English conditions ($b = -.17$, $SE = .13$, $z = -1.33$, $p = .18$), nor the strong-fit English and the weak-fit Spanish conditions ($b = -.21$, $SE = .13$, $z = -1.59$, $p = .11$). Second, we visually inspected each electrode to check the quality of the EEG recording. Specifically, we looked for eye motion artifacts (e.g. horizontal movements), electrocardiographic (ECG or EKG) and other muscular artifacts, and instances of power line noise, channel noise, and channel pop-off effects. If a single electrode had multiple artifacts that were not removed or corrected using the prior methods, we interpolated the entire electrode channel. However, we never interpolated the three main electrodes along the midline (Fz, Cz, and Pz). On average, we interpolated roughly 3 out of 32 electrodes across all participants.

2.2.5. Statistical Analyses

2.2.5.1. Pre-registered mean amplitude analyses (300–500 ms post-stimulus onset)

We first averaged the ERP amplitudes from our pre-registered N400 time window of 300–500 ms post-stimulus onset for each trial (and channel location). This process created 3,840 mean amplitude values (120 trials \times 32 electrodes) for each participant prior to exclusions. For our mean amplitude analyses, however, we only used the averages from three midline electrodes (Fz, Cz, and Pz). We then modeled these averages with a linear mixed effects model using the *lme4* package

in the R statistical computing environment (Bates et al., 2014; R Core Team, 2020). Our model had dummy-coded, fixed effects of *contextual fit* (strong-fit = 0, weak-fit = 1) and *language* (English = 0, Spanish = 1) as well as their interaction.² The model had a maximal random effects structure: there were random intercepts and random slopes for *language*, *contextual fit*, and their interaction for both participant and item grouping factors.

To evaluate significance, we adopt the convention of having an absolute value of t greater than 2 (Gelman & Hill, 2006). This is due to the on-going debate about how to best calculate the appropriate degrees of freedom for the test statistics in linear mixed effects models (see Baayen et al., 2008). However, we also report the p -values as calculated by the *lmerTest* package, as both methods of evaluation arrived at the same conclusions. The code for our statistical analyses and model comparisons can be found on OSF (see <https://osf.io/jwqpr/>).

2.2.5.2. Exploratory permutation-based cluster mass analyses (0–2000 ms post-stimulus onset)

In an exploratory analysis, we used a permutation-based cluster mass technique (Fields & Kuperberg, 2020; Groppe et al., 2011a; Maris & Oostenveld, 2007) to investigate the full range of effects during comprehension. This approach should both confirm any effects observed between 300–500 ms and detect other effects not captured by our pre-registered mean amplitude analyses. Mean amplitude analyses have been shown to have limited power for detecting small, long-lasting effects, as they often involve averaging across many electrodes and time points with small or absent effects. Permutation-based cluster mass analyses do not have this limitation, instead they

² We ran additional models that included *midline electrode site* (Fz, Cz, or Pz) as either a random effect or a control variable. The models with *midline electrode site* as a random effect resulted in singular fits. The model with *midline electrode site* as a control variable did not improve model fit, $\chi^2(27, N = 32) = .65, p = .72$; thus, we collapsed across these midline electrode sites in our final analysis.

preserve power for effects that emerge slowly over time and broadly across the scalp (Fields, 2019; Groppe et al., 2011b; for simulations, see Fields & Kuperberg, 2019).

Permutation-based cluster mass analyses employ the following procedure: First, an ANOVA is performed at each electrode site for each time-point in the target window. The results from each of these spatially and temporally distinct ANOVAs are compared to a threshold for cluster inclusion. We used a p -value of 0.01, as recommended for exploratory analyses looking at long time-windows (see Fields, 2019). All spatially and temporally adjacent points (i.e. neighboring electrodes at similar times) with p -values exceeding this threshold are grouped into a single cluster. Then, for each cluster, we calculate a cluster mass statistic by summing all of the cluster's F -values. Finally, we evaluate a cluster's significance using permutation-based corrections for multiple comparisons. To do these corrections, we first create a distribution of possible cluster statistics computed from randomly permuted data (with null effects). We then compare our observed cluster statistics to the null distribution to determine significance at a predetermined alpha level. For example, if we set $\alpha = 0.05$, a significant cluster statistic would need to fall outside of the 95 percentile (i.e. $1 - \alpha$) of the null distribution.

Prior to our cluster analyses, we downsampled the data to 100 Hz using the boxcar filter, which averages adjacent time-points together, reducing the data to the desired sampling rate. This procedure left us with 200 samples between -5 to 1985 ms post-stimulus onset. Note, this unusual time-window is a product of downsampling and re-baselining from -200 to 0 ms. We implemented our analyses using the Factorial Mass Univariate Toolbox extension (FMUT; Fields, 2017; Fields & Kuperberg, 2020) for the Mass Univariate Toolbox (MUT; Groppe et al., 2011a). We used the recommended number of 100,000 permutations and $\alpha = 0.05$ (Fields, 2019) for our main ANOVA and our four pairwise comparisons, which further addressed effects of contextual fit and language.

The pairwise tests were corrected for multiple comparisons by applying a new Bonferroni-corrected alpha level ($\alpha = 0.0125$). In the sections below, we first present the results from the mean amplitude analyses, and then those from the cluster mass analyses.

2.3. Results and Discussion

2.3.1. Averaged waveforms and topographic voltage maps

Figure 2.1 shows the averaged waveforms for the three midline electrodes (Fz, Cz, and Pz), as well as the combined averages for left anterior, right anterior, left posterior, and right posterior electrodes. In interpreting these waveforms, it is important to take into account the differences that are often found between auditory and visual ERPs. Typically, auditory ERP components have earlier onsets, later offsets, and wider/broader distributions across the scalp relative to visual ERP components (e.g. Fernandez et al., 2019; Grey et al., 2019; Grey & van Hell, 2017; Holcomb & Neville, 1991; Kutas & Federmeier, 2011; Liao & Chan, 2016; Ruigendijk et al., 2016). These differences are presumably due to the fact that visual words are presented all at once while auditory words unfold over hundreds of milliseconds (see Connolly et al., 1995; Van Petten et al., 1999).

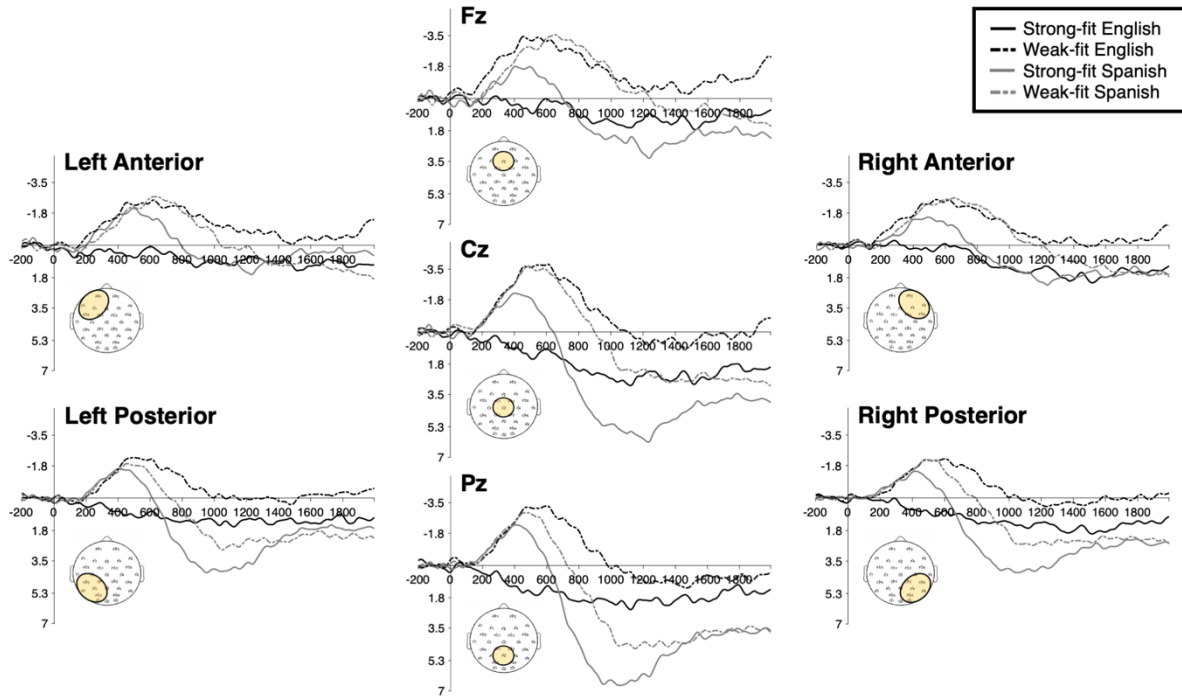


Figure 2.1: *Grand waveforms for all conditions.* The averages (μV) for the midline electrodes Fz, Cz, and Pz are presented in the center panel. The combined averages for left anterior, right anterior, left posterior, and right posterior electrodes are positioned in their respective quadrants. The two dark lines at each site indicate the English conditions, while the lighter lines indicate the Spanish code-switched conditions. The solid lines indicate strong-fitting conditions, while the dotted lines indicate weak-fitting conditions. The canonical N400 effect is seen for all violation conditions (300–500 ms) and the LPC effect is seen for all code-switched words (700–1200 ms). Both effects are most prominent over parietal site (Pz). The sustained negativity for weak-fitting words (700–1200 ms) is most prominent over frontal site (Fz). All waveforms were subjected to an additional low-pass filter at 10Hz for plotting purposes.

2.3.2. Mean Amplitude Analysis at Fz, Cz, and Pz (300–500 ms): The N400

In our pre-registered time window of 300–500 ms, there were significant effects of contextual fit ($b = -3.51$, $SE = .72$, $t = -4.86$, $p < .001$) and language ($b = -2.62$, $SE = .66$, $t = -3.99$, $p < .001$) at midline electrodes Fz, Cz, and Pz. The weak-fitting words and Spanish code-switches elicited greater N400 amplitudes relative to strong-fitting words and English non-switches respectively.

These main effects, however, were superseded by an interaction between contextual fit and language ($b = 2.87, SE = 1.01, t = 2.86, p < .01$).³

To unpack this interaction, we performed planned pairwise comparisons using the *emmeans* package in R (Lenth et al., 2021). The reported p -values were first obtained by comparing the pairwise estimates against a standard normal distribution (rather than the t distribution) and then adjusted for multiplicity using the Tukey method. The pairwise comparisons revealed a main effect of contextual fit between English conditions ($b = 3.52, SE = .72, z = 4.86, p < .0001$) such that weak-fit English words were more difficult to process, eliciting greater N400 responses relative to strong-fit English words. There were no differences between the N400 amplitudes elicited by the two Spanish code-switch conditions, suggesting that they were similarly difficult to process for listeners ($b = .64, SE = .68, z = .95, p = .78$). There was also a main effect of language between strong-fit words such that the strong-fit Spanish code-switches were more difficult to process and elicited greater N400 responses relative to strong-fit English words ($b = 2.62, SE = .66, z = 3.99, p < .001$). Finally, there were no differences between the two weak-fit conditions ($b = -.25, SE = .66, z = -.38, p = .98$) nor the strong-fit Spanish and the weak-fit *English* conditions ($b = .89, SE = .58, z = 1.53, p = .42$). Taken together, these comparisons show that all unexpected conditions (i.e. the weak-fit English and both Spanish conditions) were more challenging for bilingual listeners than the expected strong-fit English condition—and moreover, that all forms of unexpected words elicited the same magnitude of comprehension difficulty (as indexed by their equally-sized N400 responses between 300–500 ms).

³ We also implemented these models with covariates for participants' proficiency levels in English and in Spanish. The pattern of significance did not change when controlling for proficiency differences in the subset of 22 participants that completed the LEAP Questionnaire. More information about these analyses can be found in our annotated analysis script on OSF (<https://osf.io/jwqpr/>).

For ease of comparison to prior work, we also conducted a set of exploratory analyses to see if there were any distributional differences across our N400 effects. To do this, we ran three separate mixed effects models: The first model compared strong-fit English words to both Spanish conditions (collapsing across contextual fit). The second model compared strong-fit English words to the weak-fit English words. The last model compared the three violation conditions to each other. All three models included distributional factors of *hemisphere* (left vs. right), *laterality* (lateral vs. medial), and *anteriority* (pre-frontal, frontal, centro-temporal, and occipital).⁴ Results indicated that, when comparing strong-fit English conditions to both Spanish conditions (collapsing across contextual fit), there was an effect of language ($b = -1.46$, $SE = .34$, $t = -4.24$, $p < .001$), confirming our prior findings. The only distributional factor that interacted with language in this model was laterality such that the N400 effect for Spanish words was more negative at medial electrode sites than at lateral ones ($b = -1.55$, $SE = .49$, $t = -3.18$, $p < .01$). There were no other significant two-way, three-way, or four-way interactions. In the second model, when comparing strong-fit English to weak-fit English conditions, there was an effect of contextual fit ($b = -1.41$, $SE = .39$, $t = -3.66$, $p < .001$), again confirming our prior results. Similar to the Spanish conditions, the only distributional factor that interacted with contextual fit was laterality, as the N400 effect was more negative at medial electrode sites than at lateral ones ($b = -1.92$, $SE = .54$, $t = -3.53$, $p < .001$). Again, there were no other significant two-way, three-way, or four-way interactions. Finally, the direct comparison of all three unexpected conditions did not yield any significant differences, reaffirming that all three N400 effects had similar magnitudes and scalp distributions.

⁴ These distributional factors are originally from Moreno et al. (2002) and Ng et al. (2014). Both studies used these factors to argue for left-lateralization of their switch-related negativities (i.e. LAN effects). Specific details about which electrodes were used in each group can be found in our annotated analyses on OSF (<https://osf.io/jwqpr/>).

2.3.3. *Permutation-based cluster mass analysis across all electrodes*

We conducted an exploratory permutation-based cluster mass analysis using all electrodes and all 200 time points between -5 to 1985 ms post-stimulus onset. We first report the results for the interaction in the main ANOVA and then the results from four pairwise comparisons. Extensive information about all of the results, the statistical procedure, and the raw output can be found on OSF (see <https://osf.io/jwqpr/>). First, our analysis revealed a significant cluster for the interaction that lasted between 355–535 ms (Summed F -statistic = 2125.001, $p < 0.05$). This cluster was broadly distributed across centro-parietal electrode sites, and the effect was greatest at 385 ms over electrode FC1, which neighbors electrode Cz (see Figure 2.2 for raster plots, waveforms, and scalp topographies). This significant interaction confirms the findings from our mean amplitude analyses, which revealed an interaction in the pre-registered 300–500 ms window such that the Spanish code-switched words and weak-fitting English words elicited similar N400 responses.

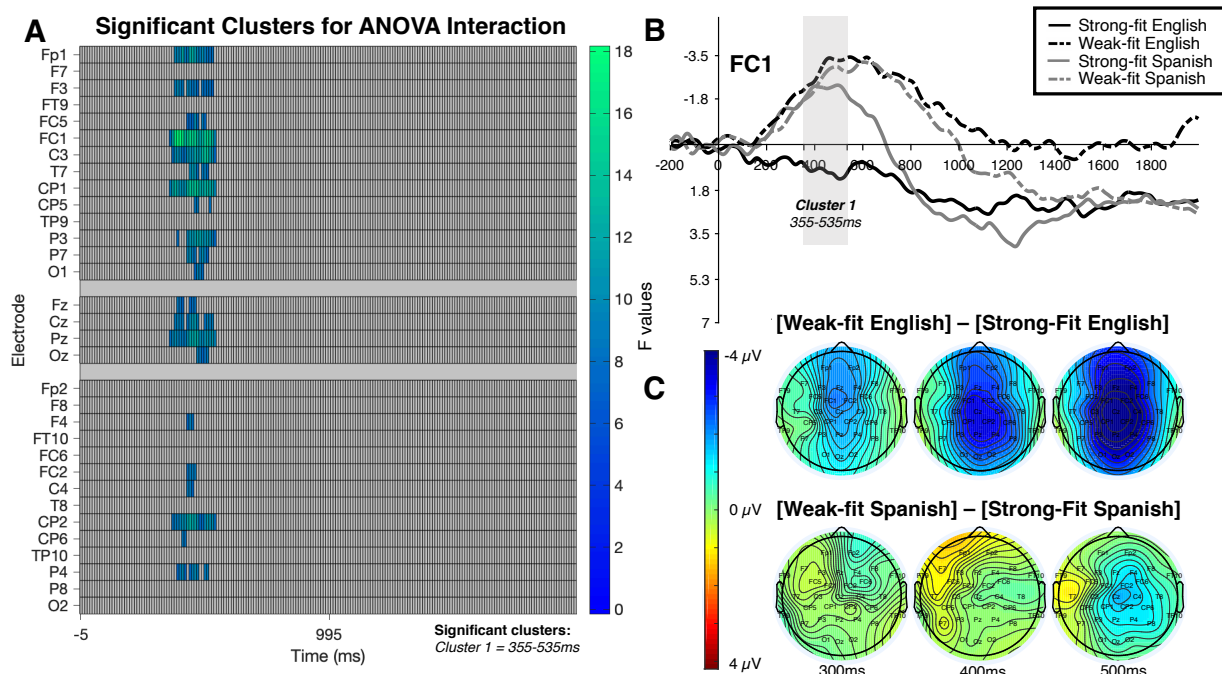


Figure 2.2: Cluster mass results from main ANOVA interaction. This graphic indicates A) when and where the main interaction was significant in the trial (i.e. broadly distributed effect between 355–535 ms); B) the waveform at electrode FC1, where the interaction effect was maximal; and C) the topographic maps of the difference waves for contextual fit between language conditions. All values in (B) and (C) are μV s. These plots show a significant interaction represented as typical N400 effects for all three violation types.

Given this pattern of effects, one might wonder whether there are any effects of code-switching that are independent of predicting a specific word. To explore this, we conducted a cluster analysis comparing the weak-fit English and the weak-fit Spanish conditions. Both are unpredicted words, but the latter involves a language shift. The analysis revealed a late positivity restricted to parietal electrodes between 885–1985 ms (Summed F -statistic = 9310.81, $p < 0.01$). This LPC effect peaked at 1025 ms over parietal electrode, P4 (see Figure 2.3). We take this LPC effect as evidence that bilinguals recognized that the speaker switched languages, i.e., a high-level (unexpected) discrepancy that needed to be re-evaluated (Friederici, 2005; Kaan et al., 2000; Kolk & Chwilla, 2007; Kuperberg et al., 2020; Litcofsky & van Hell, 2017; Van Petten & Luka, 2012).

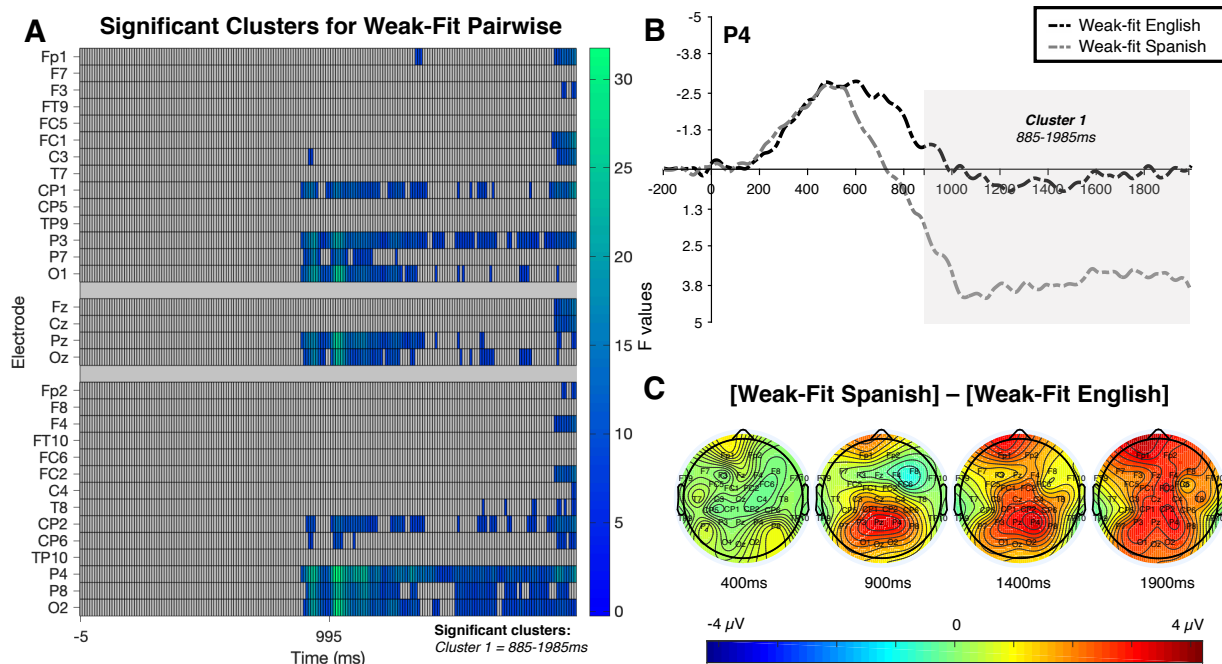


Figure 2.3: Cluster mass results from weak-fit comparison. This graphic indicates A) when and where the weak-fit conditions differed significantly (i.e. posterior LPC effect between 885–1985 ms); B) the waveform at electrode P4, where the effect was maximal; and C) the topographic maps of the difference wave for language within weak-fitting conditions. All values in (B) and (C) are μVs . These plots show a late-emerging posterior positivity in response to code-switching for the weak-fitting conditions.

Similarly, one might ask whether there are any effects of contextual fit that are independent of predicting a specific word. We explored this by comparing the strong-fit Spanish and the weak-fit Spanish conditions. Both are unpredicted word forms, but the latter condition also involves a concept that is a poor fit for the context. This analysis revealed a broadly-distributed negativity lasting between 545–1265 ms (Summed F -statistic = 11242.94, $p < 0.0125$). This sustained negativity peaked at 955 ms over the midline electrode, Cz (see Figure 2.4); however, the effect became more frontally distributed towards the end of the cluster. Sustained negativities, especially those that are frontally distributed, have been associated with increased working memory demands (e.g. Coulson & Kutas, 2001; King & Kutas, 1995), continued activity associated with word identification (Liao & Chan, 2016), and/or cognitive control processes (Lee & Federmeier, 2006,

2009, 2012; Nieuwland et al., 2007; Nieuwland & Van Berkum, 2006). We interpret this sustained negativity as reflecting increased or persisting difficulties with integrating a weak-fitting word into the unfolding context.

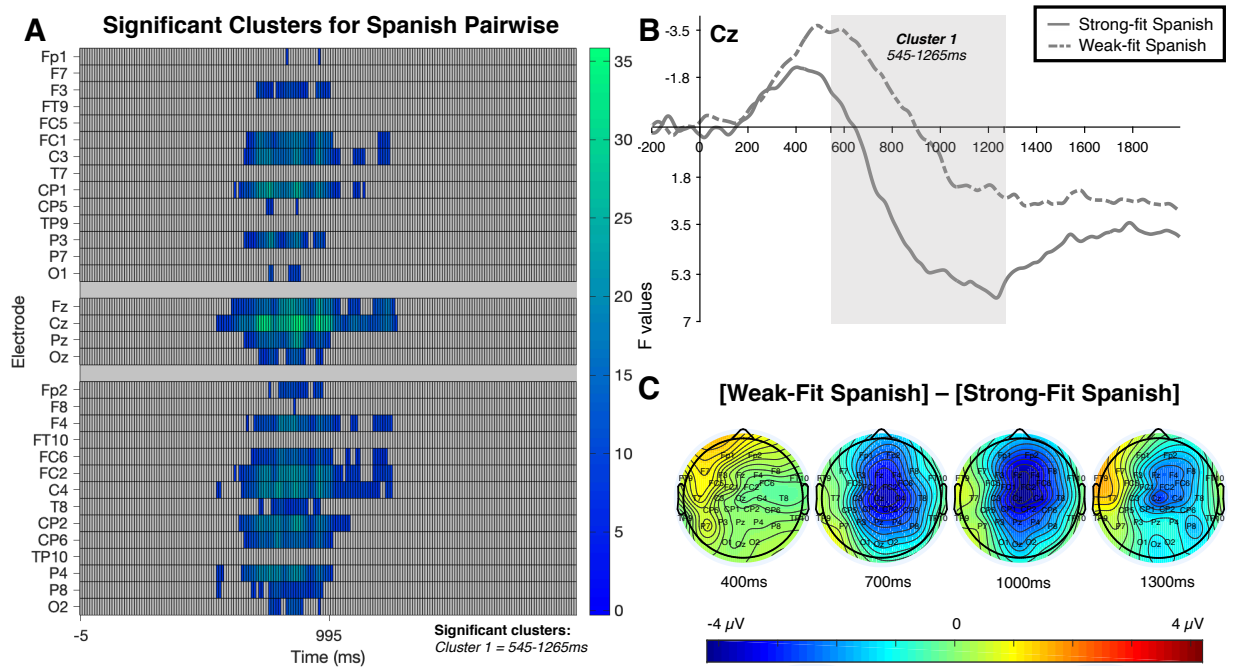


Figure 2.4: Cluster mass results from Spanish comparison. This graphic indicates A) when and where the Spanish code-switched conditions differed significantly (i.e. sustained negativity between 545–1265 ms); B) the waveform at electrode Cz, where the effect was maximal; and C) the topographic maps of the difference wave for contextual fit within Spanish conditions. All values in (B) and (C) are μVs . These plots show a sustained negativity in response to weak-fitting words within Spanish conditions.

The remaining two pairwise comparisons involve conditions that differ along one dimension in our original experimental design but, by hypothesis, differ in terms of two cognitive processes each. In comparing the strong-fit English and the strong-fit Spanish conditions, we are comparing a word that is both predictable and in the matrix language to a word whose word form is unpredictable and involves code-switching. Thus, we might expect to see two effects: an early effect (i.e. the N400) reflecting the unexpected word (as in Figure 2.2) and a late positive component

reflecting the language shift (as in Figure 2.3). The analysis revealed two significant clusters. The first cluster was a negativity distributed along the midline, which lasted between 235–595 ms (Summed F -statistic = 9431.797, $p < 0.0125$) and peaked at 435 ms over parietal electrode, Pz (see Figure 2.5). This early negativity reflects the N400 response captured in our previous analyses. The second cluster was a late positivity distributed across the parietal electrodes between 755–1305 ms (Summed F -statistic = 7472.399, $p < 0.0125$), which peaked at 965 ms over parietal electrode, P3 (see Figure 2.5). Again, the presence of the LPC seems to index recognition of the language switch, as these effects appeared in both Spanish code-switched conditions regardless of contextual fit.

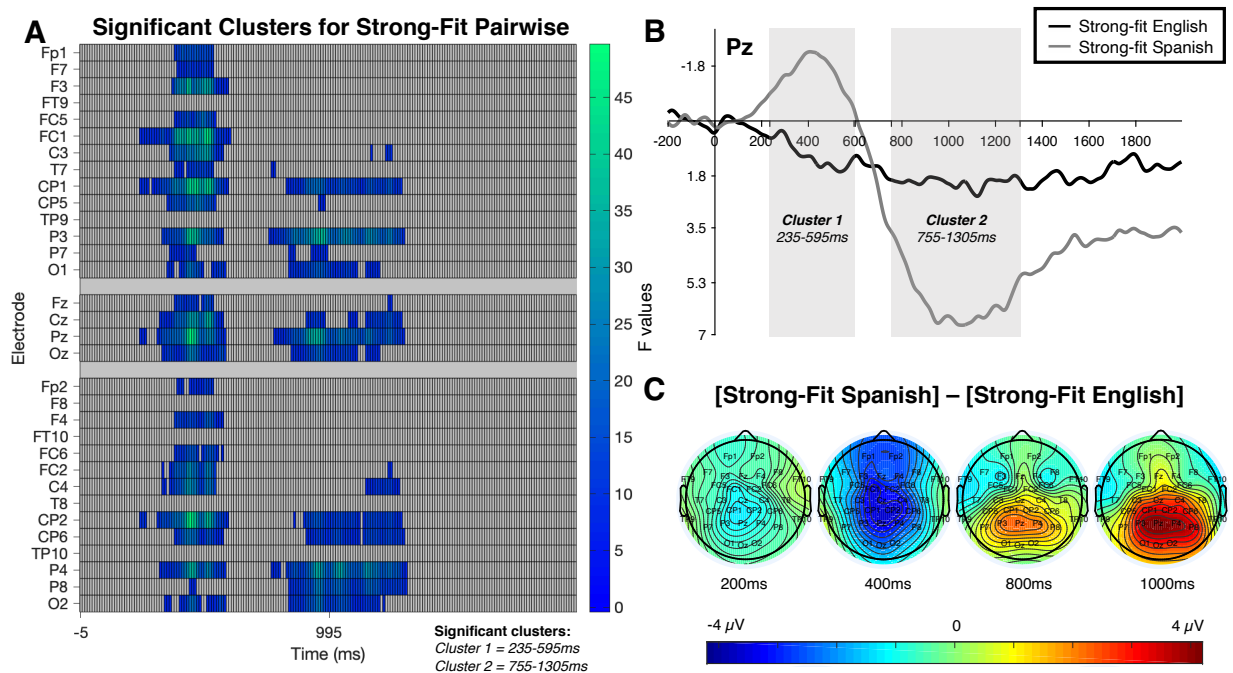


Figure 2.5: Cluster mass results from strong-fit comparison. This graphic indicates A) when and where the strong-fit conditions differed significantly (i.e. early N400 effect between 235–595 ms and posterior LPC effect between 755–1305 ms); B) the waveform at electrode Pz, where the effects were maximal; and C) the topographic maps of the difference wave for language within strong-fit conditions. All values in (B) and (C) are μ Vs. These plots show an early N400 effect for the unexpected strong-fit Spanish word and a late-emerging posterior positivity reflecting the code-switch.

In the last comparison, between the strong-fit English and the weak-fit English conditions, we are comparing a word that is predictable and easily integrated into the context with one that is unpredicted and hard to integrate. Thus, we might expect to first see an initial N400 effect reflecting the processing of the unexpected word (as in Figures 2.2 and 2.5) followed by a later long-lasting negativity that begins toward the end of this time window (as in Figure 2.4). This long-lasting negativity may reflect the continued difficulty of integrating weak-fitting words (regardless of language) into a broader discourse (see Liao & Chan, 2016 for similar effects). Since these two effects are adjacent in time, overlapping in space, and in the same voltage direction, they should be continuous in the cluster analysis, resulting in a single long-lasting cluster. Indeed, the analysis revealed a long-lasting negativity between 265–1195 ms (Summed F -statistic = 35173.77, $p < 0.001$) that peaked at 515 ms over electrode CP1, which neighbors electrode Cz (see Figure 2.6); however, the effect became more frontally distributed toward the end of the cluster (similar to the sustained negativity in Figure 2.4). We believe this long-lasting cluster reflects the summation of the (centro-parietal) N400 prediction effect and the sustained negativity observed in the other weak-fit condition.

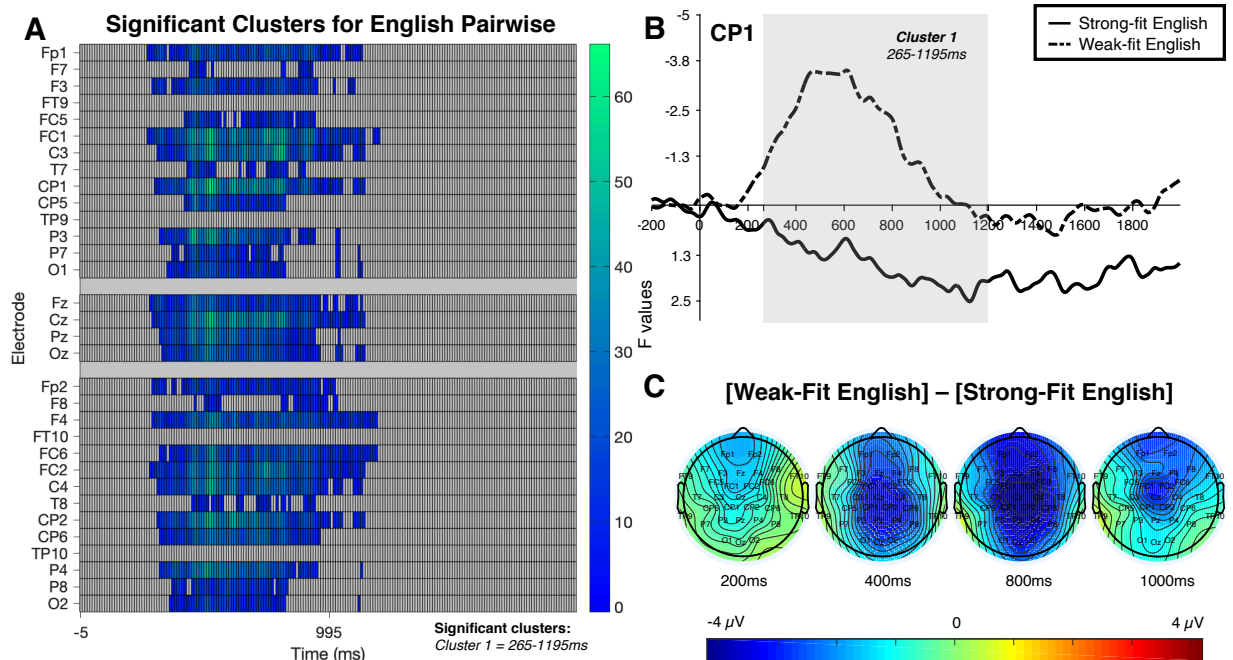


Figure 2.6: Cluster mass results from the English comparison. This graphic indicates A) when and where the English conditions differed significantly (i.e. one long-lasting negativity between 265–1195 ms); B) the waveform at electrode CP1, where the effect was maximal; and C) the topographic maps of the difference wave for contextual fit within English conditions. All values in (B) and (C) are μVs . These plots show a long-lasting negativity, which we interpret as an overlapping N400 effect and a sustained negativity for the weak-fitting, unexpected English word.

2.4. General Discussion

In this study, we tested whether switch-related ERP effects are better understood as direct costs associated with switching languages or as indirect consequences of processing unexpected words. To explore this, we factorially manipulated contextual fit and the presence of code-switching to explore bilinguals' responses to weak-fitting, within-language words and to strong and weak-fitting code-switched words. Given prior findings, we expected to find increased switch-related negativities for all three violation types and LPC effects for code-switches regardless of contextual fit (e.g. FitzPatrick & Indefrey, 2014; Liao & Chan, 2016; van Hell et al., 2015, 2018). Using our novel *Storytime* paradigm, we successfully replicated these prior findings, as well as a lesser-known sustained negativity effect for weak-fitting words (e.g. Liao & Chan, 2016).

Critically, we found that the N400 effect for double violations (i.e. weak-fitting, code-switched words) was equivalent to the effects for words that were either weak-fitting or code-switched. This pattern suggests that the N400 effects for code-switching are simply a specific case of lexico-semantic predictions being violated. In these rich contexts, listeners predict a particular word in a particular language. When that prediction is violated, lexical access and/or integration is more difficult, resulting in an increased N400 response. This cost is the same regardless of whether the prediction is violated due to the language, the meaning, or both. This pattern contrasted with the two later effects that we found: the LPC and a sustained negativity. The LPC was unique to the code-switching conditions and occurred regardless of contextual fit, and the sustained negativity was unique to words whose meaning did not quite match the context (regardless of whether they were in English or Spanish).

We interpret the present findings as supporting the one-cost hypothesis presented in the Introduction. These findings demonstrate that early lexical processing can often occur prior to or independent of recognizing the language of the word being processed. Our results are consistent with a body of evidence showing that bilinguals can simultaneously map the sounds that they hear (or the letter/signs they see) onto lexical forms in both (or all) of the languages that they know (e.g. Dijkstra et al., 1999; Duyck et al., 2007; van Hell & de Groot, 1998). Given this simultaneous mapping *and* a system that predicts lexical forms, we should expect that any unique cost of switching languages should occur after the initial costs of processing an unexpected word.

This data pattern is quite robust. As we mentioned in the Introduction, there are two other code-switching studies that used similar designs to our own and found similar data patterns (FitzPatrick & Indefrey, 2014; Liao & Chan, 2016). However, because these studies were designed

with different questions in mind, they did not directly assess the predictions of the one-cost and two-cost hypotheses.

In the remainder of this General Discussion, we will address five issues: First, we will discuss the Liao and Chan (2016) and FitzPatrick and Indefrey (2014) studies in more detail, evaluating their interpretations of their findings in light of the present data (Section 2.4.1). Next, we discuss the prior code-switching studies that find LAN effects rather than canonical N400 effects, situating these data in the larger debate about the functional difference between them (Section 2.4.2). We then turn to prior studies that do not find *any* early negativities related to code-switching (Section 2.4.3). Given these findings, we then consider the implication of our results for theories about the functional significance of the N400 (Section 2.4.4). Finally, we integrate our findings into the prior literature on the LPC (Section 2.4.5), and end by describing the methodological contribution of this paper by examining the advantages and limitations of the *Storytime* paradigm (Section 2.4.6).

2.4.1. Previous factorial manipulations of contextual fit and language switching

As we noted above, there are two other code-switching studies that used designs similar to ours and found very similar data patterns but interpreted them in different ways. Both Liao and Chan (2016) and FitzPatrick and Indefrey (2014) found the three basic ERP effects that were present in our study: an early interaction in a negative component (~250–450 ms), a longer-lasting negativity for all words that did not fit the context (collapsing across language), and an LPC effect for all code-switched words (collapsing across fit).⁵ Despite the similarities across studies, the authors arrived at divergent conclusions about the costs of switching languages during

⁵ Although, it is important to note that the LPC effects for the double violation conditions in Liao and Chan (2016) and FitzPatrick and Indefrey (2014) were heavily reduced (and sometimes non-significant) due to the overlapping sustained negativities that were also present in this particular condition.

comprehension. The main issues surrounding these alternative interpretations involve the conceptualization of the effects themselves and whether or not the authors posit a unique cost associated with code-switching during comprehension.

For example, Liao and Chan (2016) interpreted their early negativities as a variant of the PMN, which emerges after listeners hears a word with phonological features that mismatch the features of the word they expected to hear given the context (see Connolly et al., 1995; Connolly & Phillips, 1994). In their study, they argue that bilinguals were able to pre-activate information about the form of the upcoming words (and not just their semantic features). This interpretation is highly plausible given their design: participants listened to word-sized audio chunks presented one-by-one with 200 ms pauses in between them, and all of their target words were sentence-final. Thus, the slow and choppy presentation of their sentences, coupled with the fact that their targets always appeared in the same sentence position, could have allowed for robust prediction. Critically, under their interpretation, there is no unique cost to code-switching—rather, there is just one cost associated with perceiving an unexpected sound, and that cost is similar across all violation conditions. Thus, their theoretical conclusions are broadly compatible with ours; the outstanding issue is whether their PMN effects can be interpreted as reflecting an identical (or at least similar) set of underlying processes as our N400 effects. There are three reasons to think this might be the case: First, as the authors note, their PMN did not have the canonical frontal distribution of a typical PMN but instead had a distribution more similar to an N400. Second, their slower presentation method and the predictable position of their violations may have speeded up processing, shifting the N400 effect to a slightly earlier time window (see Brothers et al., 2015; Kutas et al., 2006 for related discussions). Lastly, as we will discuss below in Section 2.4.2.2, there

are good reasons to believe that all language-related negativities come from the same functional family and vary continuously rather than categorically.

In contrast, FitzPatrick and Indefrey (2014) conceptualize their results in a very different way: They argue that the N400 effect for the strong-fitting code-switches reflects the initial unavailability of the meaning of the code-switched word. They argue that the cost of code-switching does not manifest itself as a greater amplitude on the N400 response but rather as a delay in lexical access that results in a transient negativity while lexical meaning is unavailable. On this account, weak-fitting code-switched words do not show greater N400 amplitudes than strong-fitting code-switched words because, in both cases, the meaning is not initially accessed.

To address these competing hypotheses, we can look at the predictions that each account makes for code-switched words in various sentence contexts. According to FitzPatrick and Indefrey, the meaning of code-switched words should always be delayed. Thus, we should see the initial N400 effect (the transient negativity) for both code switches regardless of whether the target word is predictable or not. According to our proposal (the one-cost account), the initial N400 effect reflects the violation of a lexical prediction, not the evaluation of meaning. Thus, we should only see this pattern when the word is predictable. In contrast, the later sustained negativity reflects the degree to which the meaning of the word matches or mismatches the prior context (regardless of how predictable the word was or whether it was code-switched or not).

In order to test these predictions, we would want to look at sentences with unpredictable target words, as this condition provides the most robust contrast between the hypotheses: On the delayed access account, strong-fitting code-switches should continue to show an initial negativity (relative to same language continuations) because their meaning is always initially unavailable. Moreover, the magnitude of this negativity should not be influenced by the constraint of the

sentences. In contrast, our one-cost account predicts that strong-fitting code-switches should not show any early negativity, because the lexical form of the target word cannot be predicted in unpredictable contexts. This finding would indicate that, in the unpredictable contexts, the semantic features of the input (regardless of language) were interpreted in a bottom-up fashion with all strong-fitting words showing smaller N400 responses than all weak-fitting words. This account would also predict that the only index of recognizing the language switch would occur later as an LPC effect.

In our exploratory analyses, we compared the ERP effects from trials with very high cloze values and very low cloze values. Specifically, we analyzed the top 15% of trials ($M = 91\%$, Range = 88.5–94.3%) and the lowest 15% of trials ($M = 17\%$, Range = 2.9–34.3%). In the highest cloze group, we saw the same pattern that we found in the primary analysis: a very large early N400 effect that was similar in size across the three violation types, a later sustained negativity for weak-fitting words, and an LPC for code-switched words (see Figure 2.7 below). In the lowest cloze group, there was little to no early N400 effect, but there was still a sustained negativity and an LPC (for similar findings using eye-tracking methods in low constraint sentences, see (Altarriba et al., 1996; but see, Hoversten & Traxler, 2020)).

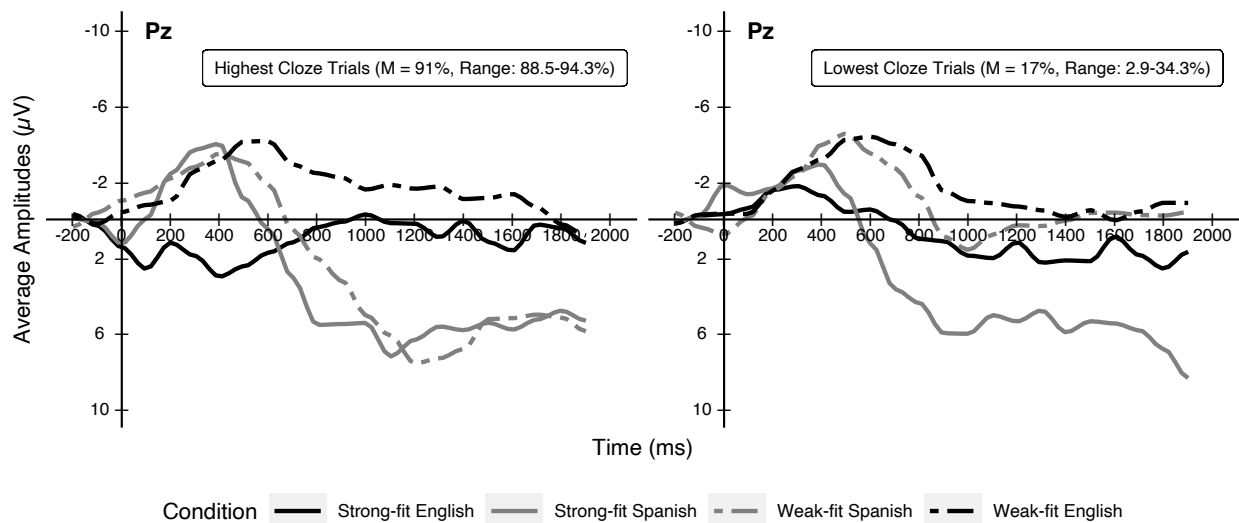


Figure 2.7: Average waveforms at Pz across highest and lowest cloze items. These waveforms were recreated for plotting purposes by taking the average amplitude values every 100 ms from -200 to 2000 ms (e.g. 100 – 200 ms, 500 – 600 ms, 1200 – 1300 ms). The lines were then fit to these averages using local regression (loess) smoothing techniques in R (see our annotated code on OSF, <https://osf.io/jwqpr/>). The two dark and lighter lines represent English words and Spanish code-switches respectively. The solid and the dotted lines represent strong and weak-fitting conditions respectively.

We confirmed these findings with a post-hoc mixed effects model that looked at average N400 amplitudes from centroparietal electrodes between 300 – 600 ms. We included random intercepts for participants and items, and fixed effects of *language* (English = 0, Spanish = 1), *contextual fit* (strong-fit = 0, weak-fit = 1), *cloze* (highest cloze = 0, lowest cloze = 1), and all of their interactions. We found a significant three-way interaction ($b = -1.04$, $SE = .51$, $t = -2.04$, $p < .05$), which suggested that the lower two-way interaction of contextual fit and language differed across cloze groups between 300 – 600 ms. Pairwise comparisons revealed that this effect was due to the fact that the two code-switched conditions significantly differed in the lowest cloze items ($b = -1.12$, $SE = .25$, $z = 4.51$, $p < .001$) but not in the highest cloze items ($b = .46$, $SE = .26$, $z = 1.75$, $p = .08$). All other pairwise comparisons (within each cloze group) were significant (see our analyses on OSF for a full model summary). These findings suggest that the meanings of code-

switched words are initially available, and that lexical access is not delayed, resulting in the early differentiation between the two code-switched conditions in the unpredictable contexts.

There are two other interesting findings to pull from these exploratory analyses: First, the LPC effects for the weak-fitting code-switches were heavily reduced in the lowest cloze trials—a similar pattern was reported in FitzPatrick and Indefrey (2014). Second, the N400 response for the baseline condition (i.e. the strong-fitting English words) increased in the lowest cloze group, confirming that the prediction of the target word was not as robust in these less predictable trials. In fact, we found a small correlation between the size of the N400 response for baseline conditions and the predictability of the target word (see Figure 2.8). As the cloze probability of the target word increases, the N400 response for that word becomes less negative, $r(118) = .18, p < .05$.

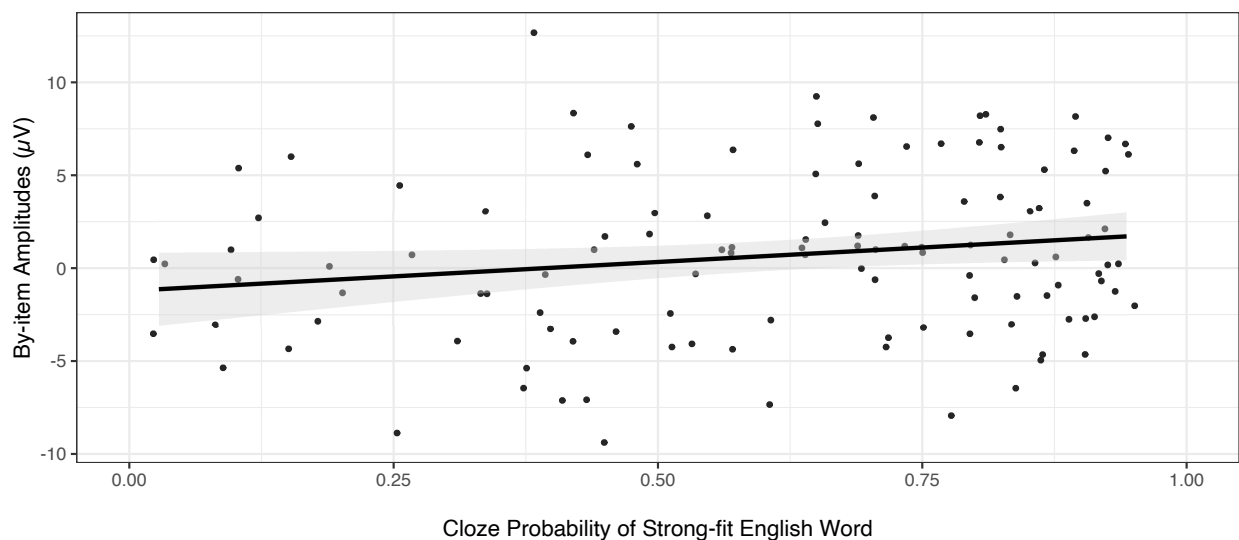


Figure 2.8. Average N400 response amplitudes across target predictability. Each observation represents the average amplitude between 300–600 ms from midline electrodes Fz, Cz, and Pz for strong-fitting English words. The x-axis is plotting the cloze probability values obtained from our audio cloze ratings task. As the predictability of the target word increases, the N400 response to that word is reduced (i.e. it becomes more positive). There is a small correlation between response size and predictability, $r(118) = .18, p < .05$.

Taken together, these exploratory findings suggest that the meanings of code-switched words are accessible during the initial stages of lexical processing—and moreover, that the variability in the magnitude of this early negativity heavily depends on how much the listener expected to hear a particular word in a particular language.

2.4.2. Understanding variability in switch-related negativities

In the Introduction, we noted that most ERP studies on code-switching find a biphasic pattern consisting of an early negativity followed by later positivity in response to code-switched words in sentence contexts. However, we also noted that there is considerable variation in the timing and scalp distribution of these switch-related effects. In this section, we focus on understanding this variability. Researchers have used many different labels for the switch-related negativities in their studies—for example, switch-related PMNs, N1s, N200s, N250s, N400s, and LANs have all been reported (see Kutas et al., 2009; Payne et al., 2020 for an overview). These effects vary in both their timing and their distribution, ranging from effects that are localized to left anterior or fronto-central electrode sites to effects that spread widely across the scalp (see Litcofsky & van Hell, 2017; Moreno et al., 2008 for reviews). Nevertheless, there is a family resemblance between them—namely, they are all negativities that take place, at least in part, during the typical N400 time window (~200–600 ms) and overlap in distribution with a typical N400 (i.e. widespread with a centroparietal focus).

The present study found a switch-related negativity with a classic N400 distribution and timing—thus, we interpreted the effect in line with prior treatments of the N400 in the broader psycholinguistic literature. Specifically, we treated it as an index of the difficulty of lexico-semantic prediction (e.g. Federmeier, 2007; Kuperberg et al., 2020; Kutas et al., 2006; Lau et al.,

2013; Otten & Van Berkum, 2008; Van Berkum et al., 2005; Wlotko & Federmeier, 2015; see Kutas & Federmeier, 2011 for a review). While many other studies have found switch-related components that look like classic N400s, switch-related effects that look like LANs are also common. In the broader psycholinguistic literature, LAN effects are often associated with an increased demand on working memory (King & Kutas, 1995; Kluender & Kutas, 1993) and/or difficulties with morphosyntactic processing (e.g. Friederici, 2002; Gunter et al., 2000; Neville et al., 1991).

In reading the code-switching literature, we found several approaches for dealing with this variability. One approach is to treat these two effects as functionally distinct, with each component reflecting a different process. On this approach, switch-related LANs are argued to index difficulties associated with integrating two language systems with different morphosyntactic features (e.g. Moreno et al., 2002; Ng et al., 2014) whereas switch-related N400s reflect difficulty in accessing the meaning of the code-switch and integrating it into the prior context (e.g. Proverbio et al., 2004; Ruigendijk et al., 2016). A second approach is to avoid making explicit commitments about the nature of these components. In practice, this often means not making a strong prediction about which effect will occur in a given study and conducting an analysis that should capture either effect if it is present (e.g. Kaan et al., 2020). A final approach is to simply note that switch-related effects vary in their distribution and often do not have the canonical N400 morphology but then treat the effects as being functionally equivalent to the N400 (see van Hell et al., 2015; van Hell & Witteman, 2009).

The present study was framed with the working hypothesis that LANs and N400s reflect a common underlying process (or set of processes). We chose this framing for the sake of simplicity and clarity, but also because we believe that there is compelling evidence that all language-related

negativities within this time window belong to the same functional family of mismatch negativities (see Bornkessel-Schlesewsky & Schlewsky, 2019 for parallel discussion). Nothing in the design of our study depended on this working hypothesis; as we noted above, many researchers have used these measures without making this theoretical commitment. However, a full interpretation of our findings and of the prior literature requires that we revisit this hypothesis. We begin by laying out the case for functionally distinct LANs and N400s and then evaluating the argument in light of the data (Section 2.4.2.1). Next, we make the case for the unitary nature of switch-related negativities, examining how this theory would account for the observed differences across studies (Section 2.4.2.2). Finally, we address an alternative theory that LAN components are an epiphenomenon resulting from instances of component overlap (Section 2.4.2.3).

2.4.2.1. The LAN and N400 as functionally distinct components

There is a long tradition in the psycholinguistic literature of treating LANs and N400s as functionally distinct components, with the LAN indexing morphosyntactic processes and the N400 indexing lexico-semantic processes (see Caffarra et al., 2019 for discussion). Thus, it is unsurprising that these two effects are often treated as distinct in the code-switching literature as well (see Litcofsky & van Hell, 2017). If these two effects are functionally distinct, then we should expect to find them in different populations or under different conditions. There are two sets of findings in the code-switching literature that provide *prima facie* support for this claim, but this evidence is limited and open to alternative interpretations.

First, there are studies that find the switch-related LANs and N400s in different populations. For example, Van Der Meij et al. (2011) recruited native Spanish speakers with either high or low proficiency in English and asked them to read English sentences. Half of these sentences had a

word code-switched into Spanish. As predicted, the authors found a biphasic pattern in response to code-switched words. However, there were distributional differences in the switch-related negativities based on speakers' proficiency in English: For low proficiency speakers, there was a canonical N400 effect, which the authors took as evidence that accessing words from the non-matrix language incurred additional processing costs. For high proficiency speakers, there was a widespread negativity that extended to left frontal electrode sites, which the authors suggested may be more akin to the LAN effects observed in balanced bilingual populations (e.g. Moreno et al., 2002; Ng et al., 2014). This distributional difference was taken as tentative evidence that highly proficient L2 speakers (and balanced bilinguals) are more influenced by the grammar of their second language than less proficient speakers, resulting in more difficulty integrating codeswitches with the matrix language.

This hypothesis, however, has not stood up well to further scrutiny. In a very similar study with Finnish-English bilinguals, Hut and Leminen (2017) found essentially the opposite pattern—a widespread negativity that extended to left anterior electrodes in their less proficient group and a canonical N400 effect in their more proficient group. Two other studies comparing bilinguals with varying levels of proficiency found canonical N400 effects with no topographic differences between the groups (e.g. Proverbio et al., 2004; Ruigendijk et al., 2016).

The second way to demonstrate a functional dissociation between switch-related LANs and N400s would be to identify stimulus factors that influence one component but not the other. Ng et al. (2014) report one analysis of this kind. They presented short stories (averaging four sentences in length) to Spanish-English bilinguals. These stories contained four instances in which target words were code-switched into Spanish. The authors report a left-lateralized negativity in response to these code-switched words, which they interpreted as a LAN effect. This left-lateralized effect

was unexpected; the authors intended to explore how the size of the switch-related *N400 effect* was influenced by the position of the word in the story. N400 effects are typically smaller for words that appear later in a sentence, consistent with standard accounts linking N400 effects to lexical access and integration (see Federmeier, 2007; Kambe et al., 2001; Van Petten & Kutas, 1990, 1991; but see, Van Petten, 1995 for connected discourse). They found that this switch-related negativity was not modulated by the position of the code-switched word in the story, as the amplitude of the effect was no different for the first two switches than the last two. However, when collapsing across switched and non-switched words, they found an N400 effect that was modulated by discourse position. The authors concluded that their switch-related negativity was a LAN and that it was functionally distinct from the N400 in their study.

Critically, Ng and colleagues' analysis rests on a null interaction in a study that may not have sufficient power to detect an effect of the relevant size. The evidence that the N400 is sensitive to their discourse manipulation is complex: when collapsing across switches and non-switches, there was a small *word position* by *word class* interaction due to the fact that nouns in either language showed smaller N400s later in the discourse, while verbs did not. Thus, the critical evidence for a functional dissociation would be a modulation of this effect—namely, a three-way interaction between *word class*, *word position*, and *switching* such that the difference between code-switched nouns and non-switched nouns would be greater at the beginning of the story than at the end. It is unclear how large we would expect such a modulation to be, but presumably it would be smaller than the two-way interaction itself (since no crossover is predicted). Given that their two-way interaction was just within the standard limits of significance ($p = .04$), it seems likely that a fairly large modulation could be missed with this design.

Future studies could address this issue with larger samples or more powerful manipulations of context. Word position effects on N400 magnitudes are thought to be a side effect of predictability, i.e., the more context that precedes a target word, the stronger the lexical prediction for that word may become (e.g. Payne et al., 2015; Van Petten & Kutas, 1991; Van Petten & Luka, 2012). Thus, a more direct test of this functional distinction claim would be to test for differences in code-switching effects when the original target word is predictable or unpredictable. In Section 2.4.1, we presented evidence from an exploratory analysis of this kind using our own data: we found that the magnitude of the N400 response for strong-fitting code-switched words was reduced (relative to other violation conditions) when the target words were less predictable (see Figure 2.7). Moreover, we would also predict that the size of the N400 effects from both violation types (i.e. code-switched words vs. weak-fit words) should vary continuously across cloze probability. To test this, we conducted another set of post-hoc mixed models: The first model directly compared strong-fit English words to both Spanish conditions. We found a significant interaction between *cloze probability* and *language* ($b = -2.86$, $SE = 1.12$, $t = -2.551$, $p < .05$), suggesting that magnitude of the switch-related N400 effect (collapsing across contextual fit) increased as the predictability of the target word increased. In the second model, we directly compared the two English conditions and found a significant interaction between *cloze probability* and *contextual fit* ($b = -4.58$, $SE = 1.27$, $t = -3.60$, $p < .001$), suggesting that the N400 effects for weak-fitting, non-switched words also increase alongside predictability. Below, we summarize these findings by plotting the N400 effect sizes for each violation condition across cloze values (see Figure 2.9).

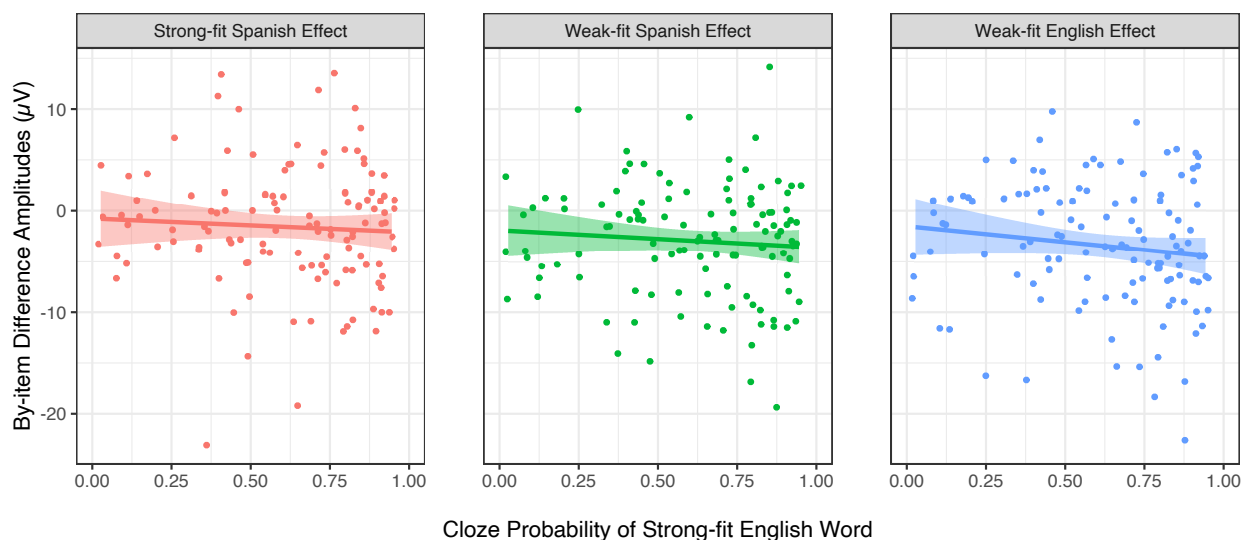


Figure 2.9. Average N400 effect amplitude across target predictability. Each observation represents the average N400 effect amplitude (violation – baseline) between 300–600 ms from midline electrodes Fz, Cz, and Pz. The x-axis is plotting the cloze probability values obtained from our audio cloze task. For each violation type, the N400 effect size becomes more negative as the predictability of the expected word (i.e. the strong-fitting English word) increases.

In sum, the prior evidence for a functional distinction between switch-related LANs and switch-related N400s is quite weak. There is no known set of factors that will reliably produce switch-related LANs as opposed to N400s. This problem extends to the broader psycholinguistic literature. In studies that are intended to elicit LANs, the observed effect is often not left-lateralized (e.g. Foucart & Frenck-Mestre, 2011, 2012; Hasting & Kotz, 2008; Lau et al., 2006; Nevins et al., 2007; Osterhout & Mobley, 1995; Tokowicz & MacWhinney, 2005) and sometimes has the morphology of a canonical N400 instead (Courteau et al., 2019; Fromont et al., 2020; Guajardo & Wicha, 2014; Nieuwland et al., 2013; Severens et al., 2008; Wicha et al., 2004).

A further reason to reject this hypothesis is that it fails to account for other ways in which switch-related negativities vary in their timing and distribution (e.g. *bilateral ANs*, Litcofsky & van Hell, 2017 on second code-switch; Zeller, 2020; *N1 effects*, Proverbio et al., 2002; 2004; *N200 effects*, Khamis-Dakwar & Froud, 2007; *left-occipital N250s*, Van Der Meij et al., 2011; *fronto-*

central negativities, Hut & Leminen, 2017; *broad negativities*, Zeller, 2020; *PMNs*, Liao & Chan, 2016; see Kutas et al., 2009; Payne et al., 2020 for reviews). Thus, it seems unlikely that code-switches in sentence contexts produce four or five categorically distinct effects that are elicited by differences across studies that we do not yet understand.

2.4.2.2. *The LAN and the N400 as a unitary phenomenon*

The second hypothesis regarding LANs and N400s is that both negativities reflect the same set of underlying processes—and thus belong to the same functional family, despite their differences in timing and scalp distribution. This functional family is often referred to as *mismatch negativities* (MMNs) because they are thought to reflect the degree to which top-down expectations mismatch the incoming sensory input. In a recent proposal, Bornkessel-Schlesewsky and Schlewsky (2019) argue that all language-related negativities can be understood as MMNs. They argue that differences in the timing and distribution of these negativities arise from differences in the specific stimuli that cause the mismatch—namely, differences in stimulus complexity, the window of temporal integration needed to detect the mismatch, and the nature of the representation that fails to match some top-down expectation. For example, the authors propose that ERP effects that are more LAN-like may arise when the top-down prediction involves more morphemic (rather than lexical) representations. But critically, all of these negativities (e.g. N400, LANs, ELANs, PMNs) reflect the same basic construct: *precision weighted prediction error* (i.e. an error signal that is inversely related to the uncertainty before the critical word, see Molinaro et al., 2011, for a related proposal, and Moreno et al., 2008 and Zeller, 2020 for discussions of how these same principles may apply to code-switching).

This proposal from Bornkessel-Schlesewsky and Schlesewsky (2019) is flexible enough to explain all of the existing data: it can account for both the LAN-like and N400-like effects; it can explain why these effects arise in response to code-switching without needing to posit two mutually-exclusive error detection processes; it can allow for intermediate data patterns; and it can encompass the host of other negativities that have been observed in code-switching paradigms (see list above). However, the flexibility of this proposal is also its greatest weakness—without a host of auxiliary hypotheses, it fails to predict when each pattern should occur. This weakness, however, may simply reflect the limits of our current knowledge. As we noted above, we do not yet have a set of factors that reliably give rise to these subtle differences in switch-related negativities.

2.4.2.3. The LAN as an epiphenomenon resulting from component overlap

There is a final hypothesis about the relationship between LANs and N400s. Some researchers argue that the LANs in most biphasic patterns may be epiphenomenal, arising in circumstances in which negative and positive components (typically N400s and LPCs) are both present and cancel out one another due to their temporal and spatial co-occurrence (see Guajardo & Wicha, 2014; Osterhout, 1997; Osterhout et al., 2004; Tanner, 2015; Tanner & Van Hell, 2014 for discussion; but also see Caffarra et al., 2019). In spoken language studies, the N400 is typically broadly distributed, not strongly lateralized, and emerges at centro-parietal electrode sites from 200–600 ms (Kutas & Federmeier, 2011). The LPC is often posteriorly distributed, can be right-skewed, and emerges at parietal electrode sites between 500–800 ms (Kuperberg et al., 2020; Tanner & Van Hell, 2014; Van Petten & Luka, 2012). Thus, the N400 response often begins prior to the LPC and can overlap in time and space with these late positivities.

In the ERP literature, component overlap may occur for various reasons: First, a single individual may generate both an N400 and an LPC in response to the same stimulus (e.g. a code-switched word). When both components are laid on top of each other, what remains is the left-most portion of the N400 in the earlier time windows (i.e. the LAN) and a positivity in the later windows at posterior electrode sites (see Osterhout & Mobley, 1995; Tanner, 2015). Second, component overlap may be artificially created when researchers average the ERP data across individual participants—a procedure that is standard in most ERP studies (Tanner & van Hell, 2014). At the individual-level, ERPs rarely show a true biphasic pattern with equally robust negativities and positivities (Tanner, 2015; Tanner & Van Hell, 2014; although, see Caffarra et al., 2019). Individual participants' data are often more negative (i.e. large early negativities with small late positivities) or more positive (i.e. small early negativities with large late positivities; see Osterhout et al., 2004; Tanner & Van Hell, 2014). Thus, when the study population contains both types of individuals, the averaging procedure can create a strong biphasic pattern.

This epiphenomenal hypothesis makes a few predictions about switch-related negativities: First, we expect that no study should find switch-related LAN effects without LPCs. To the best of our knowledge, this prediction holds true, as no code-switching study to date reports pure LAN effects and no late positivities. In the broader psycholinguistic literature, there are only a handful of studies that show pure LAN effects without late positivities (e.g. Coulson et al., 1998; Kim & Sikos, 2011; O'Rourke & Van Petten, 2011)—however, this pattern remains relatively rare. Second, in studies without LPC effects, we expect to find canonical N400 effects. In two prominent code-switching studies without LPCs, for example, the authors indeed report switch-related negativities with the canonical latency and distribution of N400 effects (see Fernandez et al., 2019; Proverbio et al., 2004).

Finally, we expect that the studies, experimental conditions, and individual participants that show larger LPCs should be more likely to have LANs and not N400s. There are two studies that find left-lateralized negativities in their most proficient L2 speakers but classic N400s in less proficient speakers. In both cases, the more proficient speakers also showed larger LPC responses to code-switched words, suggesting that the size of the LPC may contribute to the degree of lateralization (Ruigendijk et al., 2016; Van Der Meij et al., 2011). Tellingly, in Hut and Leminen (2017), both aspects of this pattern are reversed: less proficient speakers show greater left lateralization and a trend toward a larger LPC (though this interaction is not significant). This pattern suggests that the two components move in unison, just as this last hypothesis would predict.

Future studies could more directly assess how the presence (and strength) of late-arriving negativities influence the scalp distribution of earlier negativities. Many psycholinguistic studies have successfully suppressed late positivities over the course of an experiment by increasing the amount of surprising material—essentially, reducing the novelty of the violations (e.g. Hahne & Friederici, 1999). Thus, if one were to increase the number (or predictability) of the code-switches in a study, the LPC response should weaken over the course of the experiment. One could then observe how the distribution of switch-related negativities is affected by the strength of the LPC effects. Moreover, this study has the added benefit of evaluating switch-related LANs and N400s within the same individual, controlling for proficiency and the type of information that they may prioritize when processing the code-switches.

In sum, while there is considerable research to be done, we feel that the evidence to date does not provide strong support for the hypothesis that the LAN and the N400 are functionally distinct components. Instead, the data suggest that they reflect similar underlying processes with the differences in distribution reflecting either differences in the stimuli that give rise to the mismatch

(Section 2.4.2.2) or the effect of overlapping components on the observed morphology (Section 2.4.2.3). Thus, we will continue to discuss switch-related negativities as instances of N400s and to interpret them in light of the extensive work on the functional nature of this component.

2.4.3. When should we expect reduced or absent switch-related N400 effects?

On our hypothesis, the one-cost account, the larger N400s to code-switched words occur because comprehenders have predicted that they will encounter a particular word (or one of a small set number of candidate words), and that expectation is violated after encountering the translation equivalent instead. On this theory, the N400 effect is not driven by an increased N400 response to violations but rather by a decreased N400 to words that are expected due to the pre-activation of lexical and semantic information and the subsequent reduction in the processing load (e.g. Federmeier, 2007; Kuperberg et al., 2020; Kutas et al., 2006; Kutas & Federmeier, 2011; Lau et al., 2013; Van Berkum et al., 2005). A challenge for our hypothesis is explaining why some code-switching studies fail to find an N400 effect or any switch-related negativity (e.g. Moreno et al., 2002 with idioms; Ruigendijk et al., 2016 with intermediate-level speakers; Litcofsky & van Hell, 2017). On the one-cost account, there are two obvious explanations for the lack of N400 effects for code-switching: 1) the effect could be absent because the comprehender is failing to predict the upcoming word (or its features) in the matrix language; or 2) the effect could be absent because the comprehender is predicting both the expected word *and* the code-switched word to a similar degree. These two explanations seem to account for most (if not all) of the missing N400 effects.

The most common type of missing or reduced N400 effects occurs in studies where the matrix language is less familiar to the participants than the code-switched language. For example, Ruigendijk et al. (2016) investigated the comprehension of sentence-final German-to-Russian

code-switches in spoken sentences. They recruited German monolinguals (no knowledge of Russian) and Russian speakers with either high or intermediate German proficiency. Both groups with high proficiency in German (i.e. the matrix language) showed N400 effects for code-switching into Russian. The intermediate group showed no difference in their N400 responses to the German and Russian targets—in fact, the authors argue that the intermediate group showed large N400 responses to *both* the expected and code-switched conditions, suggesting a lack of pre-activation for any of the target words (see similar findings in our exploratory analyses above). This is consistent with our first explanation for missing N400 effects. Similarly, Liao and Chan (2016) only find switch-related N400 effects when switching from participants' dominant language into their weaker language. Finally, Van Der Meij et al. (2011) report delayed, less robust N400 effects in less proficient speakers. In semantic violation paradigms, participants also typically show weaker N400 effects in their second language than in their first language (see Ito, 2016; Ito et al., 2017b, 2018; Martin et al., 2013; van Hell & Tanner, 2012; Weber-Fox & Neville, 1996). Thus, the most parsimonious explanation of these data patterns is that the ability to predict words on-the-fly largely depends on fluency in the matrix language (e.g. Ito, 2016; Ito et al., 2017b, 2018; Kotz & Elston-Güttler, 2004).

Similar to being unable to use an unfamiliar language to make predictions, we can also have a challenging paradigm that makes it difficult to predict upcoming words. For example, in a study by Litcofsky and van Hell (2017), highly proficient Spanish-English bilinguals read sentences (word-by-word) that switched mid-sentence from one language to another. In this study, some sentences began in English, some began in Spanish, and half of the sentences contained code-switches. The authors did not find any differences between the N400 responses to switched and non-switched words in either switching direction. We believe that the lack of an N400 effect (in

either switching direction) reflects the fact that predicting *any* word in this study was difficult for participants. Support for this claim comes from the fact that both code-switched words *and* non-switched words elicited robust N400 responses (on the order of 2 μ V). This data pattern resembles the one from Ruigendijk et al. (2016) referenced above—i.e., when their intermediate L2 speakers could not predict in the matrix languages, there were robust N400 responses for their code-switched words and their baseline controls. The present study also finds increased N400 responses to our baseline controls when the target words are not easily predicted (see Figure 2.7 and Figure 2.8).

The challenge, however, for this theory is that Litcofsky and van Hell (2017) had highly proficient speakers and used sentences that do not seem highly unpredictable (although, they did not assess the predictability of their materials). For this reason, we believe that prediction has broken down because of the paradigm itself, perhaps due to their use of Rapid Serial Visual Presentation (RSVP) and/or their fast presentation rate (300 ms per word, 500 ms SOA). If the paradigm has made prediction more difficult, then we would expect that using the same stimuli with a more naturalistic presentation (i.e. auditory presentation where each word is presented at a speed that is correlated to the word's length) would allow prediction to occur more easily. Evidence in favor of this prediction comes from a study by Fernandez et al. (2019) who adapted the study by Litcofsky and van Hell (2017) using naturalistic auditory presentation and the same set of sentences. In contrast to the prior study, the authors find N400 effects for code-switched words in *both switching directions*, suggesting that prediction was enhanced in this more naturalistic presentation method. Moreover, the N400 response amplitudes for their baseline conditions appear to be reduced; however, this point remains speculative, as we should not readily compare N400 response amplitudes across written and spoken modalities.

Our second explanation for missing N400 effects is that, under some circumstances, bilinguals may predict the code-switched form in addition to the matrix form. Logically, this should happen most often when the location of the switch is highly predictable. The clearest example of such an effect is a study by Moreno et al. (2002), which presented a mix of regular and idiomatic sentences to English-Spanish bilinguals. In each sentence, the last word was manipulated to be the expected, within-language word, its translation equivalent, or an unexpected within-language word (e.g. “Out of sight, out of mind / brain / *mente* [mind].”). In regular sentences, the authors observed the typical N400 effect for the unexpected, within-language word and a biphasic pattern for the code-switched word consisting of a left-lateralized negativity (250–450 ms) and an LPC (450–850 ms). For the idiomatic sentences, however, the authors only report an N400 effect for the unexpected, within-language words (brain), but no negativity for the code-switched words (*mente*). We suspect that this reflects participants’ ability to predict the final word of the sentence well ahead of the time (since they know the idiom) and retrieve the relevant item in both languages (since they know that it is equally likely to end in either word). In this study, the average amplitude from 250 to 450 ms for the expected words in both regular and idiomatic sentences was about 5 μV . For code-switched words, the average amplitude in this time window for regular sentences was more negative (about 3 μV), reflecting the switch-related negativity reported in the study. However, in idiomatic sentences, the average amplitude for code-switched words was nearly identical to that of the expected word (about 5 μV). This evidence supports the idea above that participants were able to predict *both* the English and Spanish words in the highly predictable idioms but not in the regular sentences.

This raises an interesting question of whether code-switching in the wild ever becomes predictable enough to facilitate lexical processing in this way. Code-switching is not random,

instead it is argued to serve a range of discourse functions that might allow a listener to predict a switch (Auer, 1988; Gumperz, 1982; Heller, 2007; Poplack, 1980; Sebba et al., 2012). For example, there may be some words that are always code-switched, making the within-language word heavily dispreferred and arguably unexpected. In Spanish, the use of the English word *email* is preferred over the Spanish term *correo electrónico*. Similarly, many Spanish speakers will use *bar* instead of *la cantina*, *cervecería*, or *coctelería* when discussing where to meet up for drinks. In this scenario, the ‘expected’ code-switch (*email*, *bar*) should elicit smaller N400 effects relative to the more unusual, within-language word (*correo electrónico*, *cantina*). Another interesting question would be whether or not the bilinguals consider these borrowed words to be “code-switches” at all—as they are probably widely accepted in their language. This theory would make an interesting prediction: bilinguals that consider the words to be “code-switches” should show LPC effects, whereas those that do not consider the word as a language switch would not show LPC effects. To the best of our knowledge, there is no study that investigates these hypotheses, but clearly under our account, N400 effects for code-switches are predicted to disappear when code-switching is expected, and LPC effects should emerge whenever the word is interpreted as an unexpected switch in the language.

Taken together, the variability in the literature on switch-related N400 responses can be accounted for by a simple prediction account. In the next section, we address the implications of these findings for our theories of the N400 and its sensitivity to form-based predictions.

2.4.4. *What does this tell us about the functional significance of the N400?*

For decades, the N400 response has been used in psycholinguistic studies to determine the degree to which a particular context leads comprehenders to make lexico-semantic predictions,

easing the processing of a word once it is encountered (Federmeier, 2007; Kutas et al., 2006; Lau et al., 2013; Otten & Van Berkum, 2008; Van Berkum et al., 2005; Wlotko & Federmeier, 2015). There is an overwhelming amount of evidence showing that N400 responses are reduced when comprehenders are able to predict or pre-activate *semantic* features associated with upcoming words (e.g. Federmeier et al., 2002; Federmeier & Kutas, 1999; see Federmeier, 2007; Kuperberg, 2007; Kuperberg et al., 2020; Kutas & Federmeier, 2011 for reviews). In contrast, there is less evidence that the N400 response is sensitive to the pre-activation of features associated with a word's grammatical, phonological, or orthographic form (see Nieuwland, 2019). Nevertheless, there are a handful of studies that demonstrate that under certain circumstances comprehenders *can* anticipate these form-based features, resulting in reduced N400 responses (e.g. Brothers et al., 2015; DeLong et al., 2005; Ito et al., 2016, 2017; Van Berkum et al., 2005; Wicha, Bates, et al., 2003; Wicha et al., 2004; Wicha, Moreno, et al., 2003). We argue that the present study provides more evidence for this hypothesis: In our rich contexts, bilinguals seem to predict a particular word in a particular language. This pre-activation of a language-specific form leads to the reduction of the N400 response for that particular word and not its translation equivalent, which matches in semantic features. Our hypothesis makes the interesting prediction that, if the comprehender predicts form-based features for the expected word, any exact or near-cognate of that word would result in a reduced N400 effect. This is consistent with the literature on cognates showing facilitated lexical processing and N400 reductions as a function of form overlap (Christoffels et al., 2007; de Groot, 1993; de Groot & Nas, 1991; Dijkstra et al., 2015; Gollan & Acenas, 2004). Similar N400 reductions have been observed in monolingual populations when words are slightly misspelled (e.g. cake vs. *ceke*, see Kim & Lai, 2012). We take this as evidence that the N400

response is sensitive to some degree of form-based prediction, at least at the level of a particular lexical item, specified for its language.

However, we fully acknowledge that the evidence for prediction of pure word form features is limited and highly controversial at the moment (see Nieuwland, 2019; Nieuwland et al., 2018 for further discussion). Thus, an alternative explanation for our data could be that language is represented (and predicted) in a similar way to other form features like grammatical gender or number. There is ample evidence to suggest that comprehenders are capable of pre-activating features like grammatical gender (e.g. Wicha, Bates, et al., 2003; Wicha et al., 2004; Wicha, Moreno, et al., 2003). Under this account, bilinguals would need to predict both the semantic features *and* the language feature of an upcoming word in order to explain our pattern of results. If this proves to be the case, it might provide support for theories where lexical representations are divided into two levels: the lemma, which links conceptual and syntactic features, and the lexeme, which contains links to the phonological features (see Roelofs et al., 1998; cf. Caramazza, 1998). We would, however, need to specify that lemmas are sensitive to a language feature in the same way that they are sensitive to other grammatical and syntactic features like gender, number, and person (for similar proposals, see Bullock & Toribio, 2019; Poulisse & Bongaerts, 1994).

In order to test this hypothesis, we would need to have a manipulation in which two words share semantic *and* language features, but differ in word form features. Future work could investigate words that have acceptable alternative spellings (e.g. ax/axe, donut/doughnut, dialog/dialogue) in order to assess the degree to which semantic, language, and form-based features are predicted independently. To the best of our knowledge, there is no study that uses a manipulation of this kind. Thus, for the purpose of the present discussion, we will say that the most parsimonious way to interpret our data is to assume that word form prediction can occur and that

the limited evidence reflects the limits of experimental power, variability in the strength of predictive cues, the time available to make predictions, and/or the motivation or speed of processing in the research participants across studies.

2.4.5. *What does the present study say about LPC effects?*

The present study was specifically designed to explore the N400 rather than the LPC. As a result, our main hypothesis (the one-cost account) makes no differential predictions regarding the LPC effects for strong and weak-fitting code-switches. The code-switching literature suggests that LPC effects to code-switching are influenced by two main factors: the expectedness of the code-switching event and participants' language proficiency (e.g. Moreno et al., 2002, 2008; Proverbio et al., 2004; Ruigendijk et al., 2016; Van Der Meij et al., 2011; van Hell et al., 2015, 2018; van Hell & Witteman, 2009). LPC effects are often reduced (or missing) in experiments where the code-switching manipulation is highly predictable (see Proverbio et al., 2004). In the present study, our code-switching manipulation occurred at seemingly random intervals throughout the stories. Thus, participants were unable to guess when a code-switched word might appear—unlike in some prior studies that always manipulate the last word in the target sentence. We return to this point below in our discussion of the benefits and limitations of the *Storytime* paradigm. LPC effects can also be smaller and earlier when participants are more proficient in the code-switched language (Litcofsky & van Hell, 2017; Moreno et al., 2002; Ruigendijk et al., 2016; cf. Van Der Meij et al., 2011). In the present study, we did not manipulate proficiency levels in either English or Spanish. Our study population was largely dominant in English (the matrix language) but still reported high proficiency in Spanish (see Table 2.1). Nonetheless, there was still some variability in the levels of proficiencies across participants—thus, we conducted a set of exploratory analyses to see if

proficiency influenced the size of our LPC effects. However, these analyses did not reveal any significant effects of Spanish or English proficiency levels, perhaps due to the homogeneity of our study population. More information about these analyses can be found in our annotated analysis script on OSF (<https://osf.io/jwqpr/>).

Taken together, the present study provides critical information for understanding the functional significance of the LPC effects found in both monolingual and bilingual contexts. Most researchers agree that switch-related LPC effects index two aspects of comprehending a code-switch: First, the initial recognition of the switch, and then the subsequent reanalysis of the input and the prior context (Litcofsky & van Hell, 2017; Moreno et al., 2002; van Hell et al., 2018). This reanalysis process is argued to involve sentence or discourse-level restructuring, which is why LPC effects are seldom found in studies using single, isolated words (cf. Alvarez et al., 2003; Chauncey et al., 2008; Midgley et al., 2009; see Van Hell et al., 2015; 2018 for discussion). LPC effects, more broadly, have been argued to reflect a failure to update a participant's *mental model*, which is her, "high-level representation of meaning that is established and built during comprehension, based on the preceding linguistic and non-linguistic context, the comprehender's real-world knowledge, and her beliefs about the communicator and the broader communicative environment" (Kuperberg et al., 2020). This contrasts with the N400, which is seen as an index of lexico-semantic activation and integration. Thus, we expect that the LPC will not reflect the fit of the word within its context (weak vs. strong) but instead will reflect the degree to which the speech act fits into the broader communicative environment. This interpretation makes the prediction that using a design that supports code-switching events may reduce (or even eliminate) these posterior effects (see Moreno et al., 2002; 2008; Blanco-Elorrieta & Pykkänen 2016; 2017; 2018 for more discussion).

2.4.6. Methodological advantages and limitations to the *Storytime* paradigm

One goal of this study was to explore code-switching in a more natural context than is typically used in experimental studies. For this reason, we used auditory materials rather than written text (cf. Moreno et al., 2002; Ng et al., 2014) since code-switching is more common in spoken language (Fernandez et al., 2019; Litcofsky & van Hell, 2017; van Hell et al., 2018). To make the task more realistic and engaging, we used complete natural discourses rather than isolated words or sentences (cf. Alvarez et al., 2003; Liao & Chan, 2016; Moreno et al., 2002; Ruigendijk et al., 2016). Finally, we gave our participants no task beyond enjoying the story (cf. Ng et al., 2014), in an effort to remove potential strategic considerations.

One core characteristic of our *Storytime* paradigm is that our discourses are naturally produced, meaning that the speaking rate is not tightly controlled and arguably faster than those in traditional psycholinguistic experiments. One benefit of producing words naturally, however, is that each word is presented at a speed that is correlated to the word's length. The alternative is presenting all words at the same stimulus onset asynchrony (SOA), which may benefit shorter words but make longer words more challenging to process. Previous studies looking at form-based prediction in sentences found that at faster SOAs (e.g. 500 ms), the N400 effects are reactive (bottom-up) rather than predictive (Ito et al., 2016; 2017). In fact, the average SOA in our recordings is 521.9 ms ($SE = 8.0$) with a range of 20 ms to 2570 ms. This finding would suggest that the faster speaking rates in our paradigm would make it harder for listeners to make form-based predictions—which seems counter to our argument that naturalistic listening enhances prediction.

There are, however, two other temporal variables that are relevant to prediction, both of which we suspect will favor natural discourse. First, there is the time that passes between the

material that generates the expectation and the input in which the expectation is realized. For example, an expectation for an upcoming word could be formed many words (or even many sentences) before the word is actually produced.

Take, for example, this discourse: “Tim kept talking about wanting a cat for his birthday. So, when his birthday finally arrived, I went to the pet shop and bought him an adorable black and white **cat**.” In this case, a listener might have the time to make robust predictions even if the speech was quite rapid. This seems more likely to happen in a connected discourse in which ideas build upon each other over many sentences. Second, lexical prediction will depend on the amount of processing time needed to make the conceptual prediction and retrieve the relevant form. If a particular concept or lexical item is already active (because it is central to the discourse, appeared earlier, or is in the same semantic neighborhood as words that appeared earlier), then it may take less time to access that word for the purposes of form-based prediction—just like it would take less time to produce it. These factors would appear to favor prediction in rich connected discourse relative to isolated sentences. Nevertheless, we suspect that there might be even more prediction in a given discourse if the speech rate is slower. Future studies should address the interaction between these temporal variables to better understand the conditions under which we make form-based predictions.

Next, there are clearly ways in which our paradigm did not fully capture the rich context that code-switching typically occurs in. First, our design required that participants hear unexpected words in addition to code-switched words. This may have led participants to process unexpected items (e.g. code-switches) in a different way than they otherwise would have (see Van Berkum et al., 2005 for reasons to avoid implausible discourses). Second, our study involved single-word insertions rather than sentence alternations (i.e. where the sentence continues in the other language

following the code-switch). Single-word insertions are less frequent than sentence alternations (e.g. Litcofsky & van Hell, 2017; van Hell et al., 2015, 2018; cf. Poplack, 2018), especially in Spanish-English bilingual communities (Deuchar et al., 2014; Fernandez et al., 2019; Milroy & Gordon, 2008). We did this to minimize the differences across conditions such that effects of one trial would be unlikely to bleed into the next. Because we were primarily interested in the processes occurring immediately at the time of the switch—rather than downstream effects on subsequent words—this choice seemed optimal. But consistent use of single-word insertions may have made it more difficult for our bilinguals to adapt or may have introduced additional difficulties when switching back into the matrix language. Third, we selected the target words based solely on their cloze probabilities and then randomly assigned them to conditions. This meant that our code-switching events were not clearly motivated by the discourse, cultural practices, and/or language accessibility (see the *email* and *bar* examples in Section 2.4.3).

However, we still believe that these findings provide useful (and generalizable) information about how code-switching is processed in natural conversation. In most contexts, with most words, the discourse pressures and internal forces that lead one speaker to switch languages are unlikely to be completely transparent to the listener. Thus, much of the time, the within-language word will probably be expected by the listener, rather than its translation equivalent. When this is the case, we should expect the pattern of effects found here. Some initial support for this hypothesis comes from an MEG study by Blanco-Elorrieta and Pylkkänen (2017). They used naturally occurring dialogues between bilingual speakers and found an increased activation to code-switched words in auditory cortex, suggesting that listeners may predict the within-language target and have to put in extra effort to overcome this even in these natural dialogues (see Blanco-Elorrieta & Pylkkänen, 2016; 2017; 2018).

There is also a possibility that natural speech contains acoustic properties that could better signal code-switching—for example, the rate of speech may be faster for the matrix language relative to code-switched material, or there could be natural pauses prior to code-switching when produced naturally, or even subtle changes in articulators may provide natural cues rather than the unnatural transitions generated from splicing. Future work could address these issues with respect to the N400 by using dialogues between bilingual speakers as the base for stimulus creation, by implementing sentence alternations rather than single-word insertions, and by eliminating the weak-fit conditions.

2.5. Conclusion

In bilingual communities, speakers often switch between languages, and their listeners seem to readily follow them. Psycholinguistic research has suggested that these code-switches may be costly for listeners in some situations. The present study explored those costs by comparing them to the difficulties associated with hearing unexpected words within a single language context. Using our novel *Storytime* paradigm, we found three effects: an initial prediction effect (the N400), a post-lexical recognition of the switch in languages (the LPC), and a prolonged integration difficulty associated with weak-fitting words regardless of language (the sustained negativity). Together, these findings suggest that the difficulties that bilinguals encounter in understanding code-switched words can largely be understood within more general frameworks for understanding language comprehension. This work is consistent with other findings suggesting that a bilingual is not someone with two separate and competing languages living in their mind (e.g. Dijkstra & Van Heuven, 2002). Rather, bilinguals are individuals with a language system optimized to handle two coding systems, where a single lexical concept can be readily mapped onto two distinct forms (e.g.

Emmorey et al., 2008; van Hell & de Groot, 2008). As a result, the phenomenon of comprehending code-switched words in conversation can be understood as comprehending an unexpected word that just happens to be in another language.

Chapter 3

[Paper 2]

Let them eat *ceke*:

An EEG study of form-based prediction in rich naturalistic contexts

Anthony Yacovone, Briony Waite, Tanya Levari

& Jesse Snedeker

3.1. Introduction

When listening to a story or conversation, we use the sounds that we hear to reconstruct the message that the speaker is trying to convey. To do this, we must represent the incoming signal at multiple distinct levels (as phonemes, words, syntactic structures, and ideas). One of the central discoveries in psycholinguistics is that these representations are built (and refined) via both bottom-up and top-down processing. Bottom-up processing is when information from one level is used to build higher, more abstract representations (e.g. turning sounds into words and words into phrases). In contrast, top-down processing is when information from higher levels is used to influence which representations are being built at levels below (e.g. using world knowledge to interpret what someone just said or even to predict what someone is about to say).

Decades of research in psycholinguistics have focused on the role of top-down processing during comprehension—and in particular, the role of linguistic prediction.⁶ This work provides ample evidence for predictive processing of this kind, largely from reading time studies, visual world eye-tracking studies, and studies that use electroencephalography (EEG). In these studies, comprehenders read predictable words faster than unpredictable ones; they look towards particular referents in anticipation of them being mentioned; and they show reduced neural responses to words that are consistent with their top-down predictions (for reading studies, see Ehrlich & Rayner, 1981; Rayner & Well, 1996; Smith & Levy, 2013; for visual world studies, see Altmann & Kamide, 1999; Borovsky et al., 2012; Kamide et al., 2003; Milburn et al., 2016; and for EEG studies, see DeLong et al., 2005; Federmeier & Kutas, 1999; Van Berkum et al., 2005; Wicha et al., 2004).

Given these findings, there is now a general consensus that top-down prediction occurs during language comprehension (Pickering & Gambi, 2018). What remains unclear is whether prediction occurs at *all* levels of representation or only at higher ones (DeLong et al., 2021; Freunberger & Roehm, 2016; Ito et al., 2016; Nieuwland, 2019). One possibility is that, in general, comprehenders only make broad, high-level predictions about the gist of a sentence’s meaning as it unfolds. The other possibility is that, in addition to making these more general predictions, comprehenders might also make more precise predictions about the specific word(s) that will come next (Altmann & Mirković, 2009; Heilbron et al., 2022; Willems et al., 2016). Recent evidence from EEG suggests that comprehenders’ predictions can be lexically specific and occur at both

⁶ In the present study, we consider any pre-activation due to top-down processing as a form of prediction (for similar definitions, see DeLong, Troyer, et al., 2014; DeLong et al., 2021; Huettig et al., 2022; Kuperberg & Jaeger, 2016, Section 3, p. 39; Kutas & Federmeier, 2011; Pickering & Gambi, 2018). This contrasts with theorists who reserve the term *prediction* for a distinct form of processing in which an active commitment is made to upcoming material using mechanisms that are distinct from the incremental, top-down processes described above (see Kuperberg & Jaeger, 2016, Section 4, pp. 45–47; Kutas et al., 2011 for discussion).

levels of word meaning and word form (Brothers et al., 2015; DeLong et al., 2005; Ito et al., 2016; Laszlo & Federmeier, 2009; Wicha et al., 2004). But as we discuss below, form-based prediction appears to be more limited than semantic prediction, occurring primarily in tightly controlled experiments that use very predictable designs and/or slower-than-natural rates of presentation (for discussion, see Ito et al., 2016; but see, DeLong et al., 2021). In the present study, we try to better understand the scope of this phenomenon by asking whether form-based prediction occurs when people simply listen to a naturally produced story with no explicit task beyond understanding it.

In the remainder of this Introduction, we will do three things: First, we review the EEG literature on form-based prediction during comprehension (Section 3.1.1). Second, we consider the paradigms that are typically used, the limits of their ecological validity, and the questions that these limits raise about prediction in the wild (Section 3.1.2). Finally, we discuss how the present study is designed to explore form-based prediction in a natural listening context (Section 3.1.3).

3.1.1. Reviewing the EEG evidence for form-based prediction during comprehension

EEG studies have been central to our understanding of predictive processes during language comprehension (e.g. Beres, 2017; Federmeier, 2022; Kutas et al., 2006, 2014; Kutas & Federmeier, 2011; Payne et al., 2020; Swaab et al., 2012; Van Petten & Luka, 2012). In these studies, researchers typically record changes in participants' neural activity at the scalp as they comprehend a variety of sentences (Kutas et al., 2006; Kutas & Van Petten, 1994; Morgan-Short & Tanner, 2013; Swaab et al., 2012). These recordings are then time-locked to the onset of particular words, creating event-related potentials (ERPs). The interpretation of ERPs focuses on stable patterns of neural activity called *components*, which are typically distinguished from one another based on their voltage direction, peak latency, and scalp distribution, as well as their

sensitivity to particular variables (for reviews, see Kappenman & Luck, 2011; Luck, 2014). In the sections below, we review the ERP components that most commonly emerge in the study of form-based prediction in spoken language: the N400, the P600, and two early negativities known as the Phonological Mismatch Negativity (PMN) and the N200.

3.1.1.1. The N400 as an effect of lexicosemantic pre-activation during comprehension

One of the best understood and most replicated components in psycholinguistic research is the N400. This component is a negative-going deflection in the ERP waveform that typically emerges over centroparietal electrode sites, and peaks between 300–500 ms post-stimulus onset (Kutas & Federmeier, 2011). The N400 was first observed in studies in which participants read sentences with an anomalous ending, e.g., “He spread the warm bread with *socks*” (Kutas & Hillyard, 1980). For this reason, the N400 was initially characterized as a response to semantic anomalies. However, most contemporary theorists reject this characterization because it fails to account for the wide range of *plausible* contexts in which N400 effects also appear (Federmeier, 2007, 2022; Federmeier et al., 2007; Kuperberg, 2007; Kuperberg et al., 2020; Kutas & Federmeier, 2011). In fact, there is an N400 response for *every* word, whether it is presented in isolation or within a sentence (Kutas, 1993; Kutas & Federmeier, 2000, 2011; Payne et al., 2015; Rugg, 1990; Van Petten & Kutas, 1990).

In contemporary psycholinguistic theories, the N400 response is seen as an index of the relative difficulty of accessing the lexicosemantic features of a word (e.g. Federmeier, 2022; Kuperberg et al., 2020). There is ample evidence to support this interpretation: First, the N400 for a given word is larger when the word is presented in isolation and smaller when it is presented within a plausible sentence (Kutas, 1993). As we described above, when words are presented in a

broader context, they often become predictable to some degree, and comprehenders may be able to pre-activate representations associated with upcoming words. Thus, the reduction in the N400 response to a word within a sentence is often attributed to top-down prediction facilitating lexicosemantic processing (Federmeier, 2007; Lau et al., 2008, 2013; for computational descriptions of the N400, see Nour Eddine et al., 2022). Second, within a given sentence, the N400 to each subsequent word decreases as the cumulative context makes each word more and more predictable (Payne et al., 2015; Van Petten & Kutas, 1990, 1991). Third, N400 responses have an inverse correlation with cloze probability measures from offline sentence completion tasks, such that the N400 responses to words become smaller as the predictability of the words increase (Kutas et al., 1984; Kutas & Hillyard, 1984).

Subsequent research has built on this basic insight, using the N400 to explore which features of a word (or concept) are pre-activated by the context during language comprehension (e.g. DeLong et al., 2005, 2020; Federmeier & Kutas, 1999; Heilbron et al., 2022; Ito et al., 2016; Kim & Lai, 2012; Laszlo & Federmeier, 2009; Otten & Van Berkum, 2008; Wang et al., 2020; Wicha et al., 2004). For example, in a foundational study by Federmeier and Kutas (1999), participants read sentences like “The yard was completely covered with a thick layer of dead leaves. Erica decided it was time to get out the (*rake / shovel / hammer*).” In this context, the word *rake* is highly predictable, presumably leading to the pre-activation of its semantic features. By hypothesis, this should result in reduced N400s to *shovel* relative to *hammer* because the former shares more semantic features with *rake* (e.g. both are tools for yard work). As expected, the authors observed graded N400 responses such that expected words (*rake*) produced the smallest N400s followed by semantically similar words (*shovel*) and then dissimilar words (*hammer*).

Evidence for pre-activation of form features comes primarily from reading studies with violations that are orthographic neighbors of the predicted word. For example, Laszlo and Federmeier (2009) had participants read sentences with highly predictable endings like (1) and (2) below. In this study, the authors replaced these predictable sentence-final words with violations that either resembled or did not resemble the orthographic form of the original word. These orthographic violations were manipulated between items such that some sentences ended with form-similar conditions (*neighbors*, as in 1) and other sentences ended with dissimilar conditions (*non-neighbors*, as in 2). The authors also manipulated the lexical status of each violation, such that the violations were either unexpected words (*bark, clam*), nonwords (*pank, horm*), or illegal strings (*bxnk, rqck*).

- (1) *Neighbors*: Before lunch, he had to deposit his paycheck at the (*bank / bark / pank / bxnk*).
- (2) *Non-neighbors*: The genie was ready to grant his third and final (*wish / clam / horm / rqck*).

Similar to Federmeier and Kutas (1999), the authors observed graded N400 responses such that orthographic neighbors (regardless of lexical status) produced smaller N400s than non-neighbors. They interpreted this finding as evidence that, in sentences with strong contextual constraints, readers can rapidly predict upcoming words and pre-activate their orthographic features. These form-based predictions then facilitate the processing of expected words and form-similar violations. These findings have been replicated in a handful of other studies using both real-word and nonword manipulations (DeLong et al., 2019, 2020; Kim & Lai, 2012; Liu et al., 2006). And critically, these findings have been linked to top-down prediction, as the reduction in

N400s to form-similar words disappears when the original target word is less predictable (Ito et al., 2016).

3.1.1.2. *The posterior P600 as an effect of reprocessing strong violations of expectation*

Studies on form-based prediction often report another ERP component known as the P600 (see DeLong et al., 2019, 2020; Ito et al., 2016; Kim & Lai, 2012; Laszlo & Federmeier, 2009).⁷ This component typically emerges between 600–1000 ms over posterior electrode sites in response to anomalies or strong violations of expectation (see DeLong, Quante, et al., 2014; Kuperberg, 2007; Kuperberg et al., 2020; Van De Meerendonk et al., 2009; Van Petten & Luka, 2012). In the study above, Laszlo and Federmeier (2009) found P600s in response to their orthographic violations. Moreover, these P600s were sensitive to their manipulations of lexical status and orthographic similarity: First, they found that illegal strings (*bxnk*, *rqck*) produced the largest P600s, followed by nonwords (*pank*, *horm*), and then unexpected words (*bark*, *clam*). Second, regardless of lexical status, the violations that closely resembled the form of the original target words produced larger P600s than the dissimilar violations (see also Kim & Lai, 2012).

The precise interpretation of the P600 is still debated because it is observed in a wide range of psycholinguistic studies using various syntactic, semantic, and phonological violations (see Kuperberg et al., 2020; Ryskin et al., 2021; Van Petten & Luka, 2012). Researchers seem to agree, however, that the P600 reflects the initial failure to incorporate the bottom-up input into one's higher-level interpretation of the context, as well as the set of processes related to reprocessing that anomalous input (Brothers et al., 2020, 2022; Hagoort & Brown, 1999; Hahne & Friederici, 1999; Ito et al., 2016; Kim & Lai, 2012; Kuperberg et al., 2020; Laszlo & Federmeier, 2009;

⁷ This component has also been labeled as a Late Positive Component (LPC) and posterior post-N400 positivities (PNPs). For ease of discussion, we will use the term P600 in this paper to refer to all of these findings.

Osterhout et al., 1994, 2002; Osterhout & Holcomb, 1992; Van De Meerendonk et al., 2009; van de Meerendonk et al., 2010; Vissers et al., 2006). In line with this account, P600s tend to be larger in highly predictable contexts (Gunter et al., 2000; Ito et al., 2016; van de Meerendonk et al., 2010; Vissers et al., 2006) and in situations that promote deep comprehension (e.g. reading a discourse or listening to a narrative; see Brothers et al., 2020, 2022; Kuperberg et al., 2020).

According to this broad interpretation, the P600s in form-based prediction research could index comprehenders' attempts to gather more information about the nature of the violations, i.e., reflecting on whether they misperceived the input or whether someone produced a typo or speech error (Brothers et al., 2020, 2022; Kuperberg, 2007; Kuperberg et al., 2020; Van De Meerendonk et al., 2009; van Herten et al., 2005; Vissers et al., 2006). We return to these findings and interpretations in the General Discussion.

3.1.1.3. Early negativities as evidence for form-based prediction in spoken language contexts

Finally, form-based prediction research has also uncovered various ERP components that emerge *before* the N400 and P600 responses. These early components systematically differ across modalities, as some are only evoked by written language while others are only evoked by spoken language. Because the present study uses naturalistic speech, we will describe two of these early components from prior studies on spoken language comprehension: the Phonological Mismatch Negativity (PMN) and the N200. These early negativities, however, are difficult to replicate (Lewendon et al., 2020; Nieuwland, 2019; Poulton & Nieuwland, 2022), and many researchers simply interpret them as early-emerging N400 effects (e.g. Van Petten et al., 1999). But for the sake of completeness, we review the evidence for these two components below.

The first early component is the PMN, which was first reported in Connolly and Phillips (1994). In this study, the authors explored whether the initial stages of phonological processing could be influenced by listeners' top-down expectations about upcoming words. To do this, they had English-speaking adults listen to various sentences that strongly constrained for a particular sentence-final word, e.g. "At night, the old woman locked the *door*." For some sentences, the authors kept the expected sentence-final word (*door*, 3a). For other sentences, they replaced the expected word with a violation that overlapped with the expected word in its phonological onset (*eyes* → *icicles*, 3b), semantic features (*sink* → *kitchen*, 3c), or neither (*milk* → *nose*, 3d). Note, in the condition with different phonological onsets but similar semantic features (3c), the authors ensured that the violating word (*kitchen*) was always less predictable than the original sentence-final word (*sink*) from the target sentence.

- (3) a. At night, the old woman locked the ***door***. *Phoneme match, Semantic match* (door)
 b. Phil put some drops in his ***icicles***. *Phoneme match, Semantic mismatch* (eyes)
 c. They left the dirty dishes in the ***kitchen***. *Phoneme mismatch, Semantic match* (sink)
 d. Joan fed her baby some warm ***nose***. *Phoneme mismatch, Semantic mismatch* (milk)

Based on their prior work, the authors expected to find two distinct negativities: one related to processing unexpected phonological features at around 200–300 ms (the PMN) and one related to processing unexpected semantic features (the N400). Results indicated early negativities for conditions with an unexpected phonological onset (3c and 3d) relative to those conditions with the expected onset (3a and 3b). Then, in a later time window, there were greater negativities for the two semantic mismatch conditions (3b and 3d) relative to the conditions with semantically

congruent endings (3a and 3c). Taken at face value, this pattern suggests that there are two categorically distinct effects: an early PMN and a later N400.

But other features of the data pattern suggest that the effects are not discrete. The early negativity in the double mismatch condition (*milk* → *nose*, 3d) was greater than the negativity from the condition with just a different onset (*sink* → *kitchen*, 3c), suggesting that features beyond phonology influenced this early negativity. In fact, in the double mismatch condition, the PMN and the N400 blurred together to form one large, broadly distributed negativity that lasted from roughly 200–600 ms.

These findings are often contrasted with another set of findings from Van Petten et al. (1999). In this study, English-speaking participants listened to constraining sentences, e.g., “It was a pleasant surprise to find that the car repair bill was only seventeen *dollars*.” The authors manipulated the sentence-final word to be either the expected word (*dollars*), a semantically incongruous word that shares initial phonemes with the expected word (*dolphins*), or a semantically incongruous word that rhymes with the expected word (*scholars*). Results showed increased negativities for both incongruous conditions, which the authors labeled as N400 effects; however, similar to the findings above, the negativity for the condition with an unexpected onset (*scholars*) emerged earlier than the other violation condition with an expected onset (*dolphins*). Similar to the study above, the authors observed one broadly distributed negativity (rather than two distinct effects) for the condition with the unexpected onset and unexpected meaning (*scholars*).

The second early component is the N200, which is best characterized by van den Brink et al. (2001). In this study, Dutch-speaking adults listened to sentences with highly predictable sentence-final words, e.g., “De schilder kleurde de details in met een klein *penseel*” (English

translation, “The painter colored the details with a small paint *brush*”). Similar to Van Petten et al. (1999), the authors manipulated the target words to be either the expected word (*penseel*, “brush”), a semantically anomalous word with the same initial phonemes (*pensioen*, “pension”), or a semantically anomalous word with different initial phonemes (*doolhof*, “labyrinth”). In contrast with Van Petten et al. (1999), the authors found two distinct negative peaks in the ERP waveforms for all three conditions—one around 200 ms (the N200) and the other around 400 ms (the N400). The double mismatch condition (*doolhof*) produced a larger amplitude for the N200 peak than the two conditions with the expected initial phonemes (*penseel*, *pensioen*), which both produced similarly small N200 responses. The N400 effects showed a pattern similar to those found in the written studies above: the double mismatch violation (i.e. the phonologically dissimilar word, *doolhof*) had the largest N400; the shared onset violation (i.e. the phonologically similar word, *pensioen*) had a reduced N400 effect relative to the double mismatch violation; and finally, the expected word (*penseel*) had the smallest N400 effect. These effects are slightly earlier than prior findings—but the authors argued that because all of their target words began with plosives, their phonological effects may have been better aligned in time, producing more robust peaks in an earlier time window than seen in prior studies.

Although some authors consider early negativities to be categorically distinct from the N400, others interpret such effects as early modulations of the N400 (e.g. Van Petten et al., 1999; for discussion, see Lewendon et al., 2020; Nieuwland, 2019). On these accounts, the negativities produced by the unexpected conditions in the studies above simply reflect the degree of facilitated access at the form level due to the prior pre-activation of phonological features. There are two patterns that favor this hypothesis. First, as the findings above suggest, these early negativities vary considerably in their timing and scalp locations and are often continuous (in time and space)

with the later N400 effects. Second, the same factors that influence the N400 also influence these early negativities, suggesting a similar functional interpretation for these effects (see Lewendon et al., 2020; Nieuwland, 2019). Given the wide range of interpretations and the instability in these early ERP components, we focused our analyses on the N400 and P600 in the present study.

3.1.2. Reviewing the typical paradigms in form-based prediction studies and their limitations

The findings above demonstrate that comprehenders can, under some circumstances, predict the form of upcoming words. Most of this evidence, however, comes from studies with paradigms that are very similar to one another and quite different from the typical contexts of language comprehension. Thus, it is unclear how broadly these findings generalize to more real-world settings. Specifically, all of the studies to date ask participants to attend to a stream of unrelated sentences with either no clear purpose or with a goal that is independent of comprehension (e.g. monitoring for errors). Often, but not always, these studies present their sentences in ways that diverge from how language is normally produced. For example, in EEG reading studies, the sentences are presented word-by-word and often at a rate that is slower than typical reading (Ito et al., 2016; Kim & Lai, 2012; Laszlo & Federmeier, 2009; Vissers et al., 2006; but see DeLong et al., 2021).

Several recent studies have explored how presentation rate in particular might influence form-based prediction. For example, Ito et al. (2016) conducted two experiments in which they directly manipulated the predictability of sentences and the rates of presentation. Participants read sentences with highly predictable target words (e.g. “The student is going to the library to borrow a *book* tomorrow”) and moderately predictable target words (e.g. “The family went to the sea to catch some *fish* together”). In these experiments, the target words were manipulated to be one of

the following word types: the expected word (*book*), a semantically similar word with no form overlap (*page*), a semantically dissimilar word with form overlap (*hook*), or a word with no overlap in semantics or form (*sofa*). To test for an effect of presentation rate, they conducted the same experiment twice: In the first experiment, they presented sentences word-by-word with 500 ms in between each word—a rate similar to other studies in the literature (e.g. Kim & Lai, 2012; Laszlo & Federmeier, 2009). In the second experiment, the time between words was increased to 700 ms.

As expected, Ito and colleagues found N400 effects in all violation conditions (*hook*, *page*, *sofa*) in both experiments. The N400 effects for semantically similar words (*page*) were smaller than those for unrelated words (*sofa*) at both presentation rates. This reduction, however, was only present in the highly predictable sentences, suggesting that pre-activation of semantic features occurred regardless of presentation rate as long as the target word was predictable in its context. In contrast, the N400 effects for words with form overlap (*hook*) were only reduced when the sentence was both highly predictable *and* presented at a slower rate. In all other conditions, the N400 effects for words with overlapping forms (*hook*) patterned with those for the unrelated words (*sofa*). The authors concluded from these findings that it is easier to pre-activate semantic features relative to lower-level features (e.g. phonological form), perhaps because top-down predictions are initially made at the higher level of meaning, and thus activate semantic representations before they can trickle down to lower levels and activate representations of form.

Thus, Ito et al. (2016) found that, although semantic prediction occurred at both slower and faster presentation rates (1.5 words per second and 2 words per second respectively), form-based prediction only occurred at the slower rate. Given that typical adults read 3–5 words per second (Brysbaert, 2019) and speak at a rate of about 3–4 words per second (Tauroza & Allison, 1990), many theorists have concluded that form-based prediction is unlikely to occur in most ordinary

language comprehension contexts (Freunberger & Roehm, 2016, 2017; Indefrey & Levelt, 2004; Ito et al., 2016; Pickering & Garrod, 2007). This conclusion, however, has been challenged by reading studies that *do* find evidence of form-based prediction at rates of 2 words per second (DeLong et al., 2019; Kim & Lai, 2012; Laszlo & Federmeier, 2009) or even 4 words per second (DeLong et al., 2021). Critically, however, in all four of these studies the target word was always highly predictable (unlike the manipulation of predictability in Ito et al., 2016). One possibility is that all of these additional constraints might facilitate prediction—but we postpone further discussion of these findings until the General Discussion.

To the best of our knowledge, there has been no work that directly explores whether form-based prediction persists in a rich discourse context where the primary goal is comprehension. A richer discourse could facilitate prediction by introducing stronger constraints that unfold gradually over time, ensuring that the relevant lexical items are highly active long before the word itself is uttered. Alternatively, natural discourse may well be more variable than typical psycholinguistic stimuli, making prediction more complex and potentially less advantageous. After all, we usually speak because we have something *new* and potentially *unexpected* to convey.

3.1.3. *The present study*

To explore the degree to which form-based prediction occurs in naturalistic contexts, we used a novel comprehension task called the *Storytime* paradigm (Yacovone et al., 2021). Rather than using hundreds of unrelated sentences as stimuli, this paradigm uses coherent, naturally produced stories. These stories can be presented intact in a correlational design that explores how naturally occurring variation affects the ERP signals at each word (Brennan et al., 2016; Brennan

& Hale, 2019; Li et al., 2021; Levani & Snedeker, 2018) or alternatively, we can use these stories as a substrate into which experimental manipulations are spliced (Yacovone et al., 2021).

In the present study, we adopt the latter approach and splice in manipulations that resemble those from the form-based prediction studies described above: Participants will occasionally hear nonwords with varying degrees of similarity to the original word from the story (e.g. *ceke* when the original word is *cake*). Unlike prior studies, however, we did not include a condition in which the nonword shares no features with the target (e.g. *tont* when the original word is *cake*) because we did not want to completely disrupt the flow of the story. Instead, we compared the highly similar nonwords (*ceke*) with nonwords that have a different onset but the same rime (e.g. *vake* when the original word is *cake*). These rime conditions are similar to those in Van Petten et al. (1999). Finally, like Ito et al. (2016), we wanted to investigate form-based prediction in both predictable and unpredictable environments, which we identified using the cloze procedure described below. Thus, our experiment had a factorial design with three word types (*cake*, *ceke*, *vake*) in two predictive contexts (*predictable*, *unpredictable*).

Given the design of our study, we predicted that there will be a reduction in the N400 effect for similar nonwords (*ceke*) relative to the less similar rime nonwords (*vake*) in the most predictable contexts. In less predictable contexts, there should be no difference between the N400 effect for the two nonwords. This data pattern would provide strong evidence that form representations were pre-activated in predictable contexts, leading to easier processing of the similar nonword once it is encountered in the input. Such evidence would also support the idea that form-based prediction *does* occur during naturalistic listening tasks.

In addition to our main prediction, we also had a set of general expectations and secondary predictions about other data patterns that may emerge: First, for non-manipulated baseline words

(*cake*), the N400 responses should become smaller as the predictability of the word increases—i.e., predictable words should have more *positive* N400 responses than unpredictable ones. This is because the N400 is inversely correlated with the predictability of a word given its context (Kutas et al., 1984; Kutas & Hillyard, 1984). Second, the N400 effects in the more predictable contexts may emerge earlier than those in less predictable ones (e.g. Brothers et al., 2015). Finally, we anticipated that the P600 responses to nonwords would be larger in higher constraint relative to lower constraint contexts (Gunter et al., 2000; Ito et al., 2016; Kuperberg et al., 2020; Vissers et al., 2006).

3.2. Method

3.2.1. Participants

We recruited 38 native English-speaking adults from the Greater Boston area. Four adults self-reported speaking a non-American dialect of English (British English). All participants provided consent and received two study credits or cash payment for their time. We excluded eight participants from our final analyses following our pre-registered exclusion criteria: three for having more than 25% trial loss after data processing, four for poor attention to the task (e.g. falling asleep), and one for researcher error. After these exclusions, we had 30 participants for our final analyses (*Mean age* = 22 years, *Range* = 16–40 years).

3.2.1.1. Sample size calculation and power analyses for mixed-effects models

A priori power analyses were conducted using the *mixedpower* package in R (Kumle et al., 2021; R Core Team, 2021). This package uses a simulation-based approach to estimate the power for each effect of interest across a range of sample sizes. To determine our optimal sample size,

we first collected data from 30 pilot participants. Then, we implemented a set of mixed-effects models outlined in our pre-registration on OSF (<https://osf.io/ymv84>). We found that the subtlest, reliable effects in our pilot data were the interaction terms. Thus, we wanted to determine the number of participants needed to achieve at least 80% power for all relevant interaction effects with $\alpha = .05$.

For ease of computation, we dropped the random slopes in our simulations. Additionally, we did not want model convergence issues to contribute inaccurate model estimates into the simulated distribution of effect sizes. To account for dropping slopes (and to be more conservative with our power calculations), we also implemented a *Smallest Effect Size of Interest* (SESOI) approach by reducing our observed effect sizes by 20% and then calculating the sample size necessary to achieve 80% power for these SESOIs (see Kumle et al., 2021). Results indicated that we needed a final sample size of 30 participants.

3.2.2. Stimuli

The present study used a novel EEG task called the *Storytime* paradigm, which involves taking naturally produced stories and splicing in carefully controlled experimental manipulations (see Yacovone et al., 2021). Specifically, we used an abridged version of a children's book called *Mystery of the Turtle Snatcher* by Kyla Steinkraus as the substrate for our experimental design. We also created a cartoon for this story, which participants watched while listening to the story narration. To preview our design, we selected 180 target words within the story to create a 2×3 manipulation of predictability and word type. First, we selected target words with high or low predictability given their preceding story contexts (as determined by a cloze probability task). Predictable target words had higher cloze probabilities whereas unpredictable target words had

lower cloze probabilities (see Section 3.2.2.3 for more details). Then, we created three alternative productions for each target word: the original word, a form-similar nonword, or a less similar rime nonword. This resulted in the six conditions shown below in Table 3.1. In the remainder of this section, we provide additional details about how these words were selected and how the audio and cartoon stimuli were created.

Table 3.1: *Example sentences from the story.*

Predictability	Word Type	Example Sentence
High Cloze	Baseline	These hairs prove you were at the scene of the crime [kraɪm].
High Cloze	Similar	These hairs prove you were at the scene of the crame [kreɪm].
High Cloze	Rime	These hairs prove you were at the scene of the nime [naɪm].
Low Cloze	Baseline	They are only found in a certain river [rɪvər] in Texas.
Low Cloze	Similar	They are only found in a certain ruver [rʌvər] in Texas.
Low Cloze	Rime	They are only found in a certain piver [pɪvər] in Texas.

3.2.2.1. *Selecting our story substrate*

We selected *Mystery of the Turtle Snatcher* for two reasons: First, we plan to conduct a parallel experiment with young children. Thus, we needed a story with age-appropriate language and a simple, child-friendly plot. Second, we wanted to use a story that participants were unlikely to be familiar with, so that the cloze measures would reflect the predictability of the target word given the discourse context rather than prior knowledge of the story itself. Because the original story was too long to present in a single EEG recording session, we created an abridged version by eliminating non-essential passages. This version was roughly 30 minutes when read aloud.

3.2.2.2. Characterizing the predictability of every word in the story

To find predictable and unpredictable target words for our study, we conducted a cloze task (e.g. Taylor, 1953) in which we determined the predictability of every word in our story. Specifically, we recruited 541 participants on Amazon Mechanical Turk (<https://www.mturk.com>) and asked them to complete sentences from the story by guessing each word, one after another (e.g. *The...*, *The cat...*, *The cat was...*, *The cat was hungry*). Participants guessed around 300 words from a single section of the story and read the remainder of the text sentence-by-sentence. Occasionally, participants would see illustrations from the original story. We excluded 91 participants for failing data quality checks, resulting in 450 participants in the final sample. After these exclusions, we had 30 observations for each word in the story, which we used to calculate cloze probabilities.

In the present study, we define a word's cloze probability as the proportion of trials in which participants correctly guessed that word—for example, if 30 participants provided a guess for the word, and 27 participants guessed it correctly, that word would have a cloze probability of 90% (i.e. 27/30) given its preceding context. This approach, however, is slightly different from approaches that use cloze tasks to characterize whether participants converge on *any* word given the context. One could imagine a situation in which the context is highly constraining but leads participants to guess a word that was not actually used in the story itself. For example, the sentence “I like my coffee with cream and cinnamon” is apt to be completed with *sugar* instead of *cinnamon*. Thus, cinnamon would be a low cloze word in a highly constraining context. In the present study, however, cloze and constraint were tightly linked: our low cloze target words primarily occurred in less constraining contexts, and our high cloze targets necessarily occurred in highly constraining contexts.

3.2.2.3. *Selecting the predictable and unpredictable target words*

To select our predictable and unpredictable target words, we first calculated the cloze probabilities for all common nouns in the story. Then, we sorted these nouns from highest to lowest cloze and removed all nouns that started with vowel sounds (because we could only create the rime condition in a consistent way if the target word began with a consonant). We also removed words from the list if they appeared in the same sentence as another noun with more optimal cloze probabilities (i.e. higher for predictable targets or lower for unpredictable targets). Next, we removed the additional tokens of nouns that occurred more than three times in the story (e.g. turtle) to ensure that participants never heard the same manipulation more than once. In choosing which tokens of a given noun to keep, we preferentially selected those with the most extreme cloze values (either high or low). Finally, we had to remove nouns that could not be changed into nonwords following the process described below. For example, *pan* could not be turned into a form-similar nonword by changing the first vowel because all possible candidates are in fact real words (e.g. pawn, pain, pin, pen). After all of these exclusions, we selected the top 90 words for the high cloze targets and the bottom 90 words for the low cloze targets. The high cloze targets had an average cloze probability of 81.2% ($SD = 14.2\%$, $Range: 53\text{--}100\%$) and the low cloze targets had an average of 7.2% ($SD = 13.8\%$, $Range: 0\text{--}50\%$).

We also characterized the lexical frequency of each word and its length in syllables. We collected standardized word frequencies (per million words) from the SUBTLEX_{US} corpus (Brysbaert & New, 2009), which contains roughly 51 millions words from American English subtitles between 1990–2007. The high cloze targets had an average frequency of 200.37 per million words ($SD = 352.66$) whereas the low cloze targets had an average frequency of 120.91

($SD = 357.62$). For word length, the high cloze targets had an average of 1.37 syllables ($SD = 0.59$) whereas low cloze targets had an average of 1.68 ($SD = 0.76$).

3.2.2.4. *Creating the nonword violations*

We created three conditions for each of the 180 final target words: the original word, the form-similar nonword, and the less similar rime nonword. The original word was the intended target word from the story. To create the form-similar nonword, we changed the first vowel sound of each target word, ensuring that the vowel change did not result in another word of English (e.g. *nap* [næp] became *nupe* [nup] and not *nip* [nip]). Note, some of these changes may have resulted in extremely low frequency words (e.g. *beal*) or words from non-American dialects of English (e.g. *lud*, *mooth*). To create the rime nonwords, we changed the first consonant of each target word (e.g. *cage* became *nage*). We wanted the consonant change to be maximally different—so, we implemented changes on three dimensions: place of articulation, manner of articulation, and voicing. For example, the [k] in *cage* is a voiceless, velar stop whereas the [n] in *nage* is a voiced, alveolar nasal.

3.2.2.5. *Creating the spliced recordings*

After constructing all 540 target sentences (180 total target words \times 3 word types), the first author, who is a native English speaker, recorded the materials. First, he recorded the 30-minute story in its entirety. We used this story as the substrate into which we spliced our manipulations. Next, he recorded the 540 target sentences in isolation, making sure to replicate (to the best of his ability) the intonational and prosodic contours of the original recording. We then extracted the critical target words from these isolated sentence recordings and spliced them into the base

recording. In some sentences, we needed to extract a few words before and/or after the target word to avoid issues with co-articulation and prosody when splicing.

This splicing procedure ensured two properties of our stimuli: First, all conditions, regardless of being a nonword or the intended word, had been spliced in from a different audio file; and second, the auditory context before and after the target word (or region) was held constant across experimental lists. We constructed three experimental lists using a pseudo-Latin Square design, ensuring that no two target words from the same condition appeared back-to-back. We also ensured that all repeated target words appeared in different conditions in each list. To meet these criteria, we needed to have a slight difference in the number of observations per cell in each list. For example, one list had the following distribution: for high cloze targets, 27 baseline words (*cake*), 30 similar nonwords (*ceke*), and 33 rime nonwords (*vake*); for low cloze targets, 33 baseline words, 27 similar nonwords, and 30 rime nonwords.

3.2.2.6. Creating the cartoon for the story

To encourage participants to pay attention to the story, we created a cartoon to accompany it. Examples of stills from the cartoon can be found in Figure 3.1. The cartoon was created using Vyond software (Vyond.com), and the first author was unaware of which target words would be selected from the story at the time of making the cartoon. Thus, we did not design the cartoon to alter the predictability of the target words.



Figure 3.1: Stills taken from the cartoon stimulus. This cartoon was presented alongside the story narration to promote attention to and understanding of the discourse context.

3.2.2.7. *Describing the filler structure and the presentation of our target words*

In the *Storytime* paradigm, we do not construct specific sentences to serve as fillers. Instead, we rely on the sentences in the text that have not been manipulated to serve the functions of fillers (e.g. making the manipulations less predictable and reinforcing the expectation that most sentences do not have errors). Our story had roughly 500 sentences, 180 of which contained target words. Thus, there was a ratio of roughly 2:1 between fillers and targets. Moreover, only two thirds of the target sentences contained a violation, making the ratio of correct-to-incorrect sentences 3.5:1. The stimulus onset asynchrony (SOA), which represents the amount of time from the onset of one target to the onset of the next, was 10.25 s on average (*Range*: 1.33–47.81; *SD* = 8.62). To estimate the speech rate, we divided the total number of words in our story by the total amount of time spent speaking (i.e. the total phonation time). Specifically, we had 4,634 words in our story and an average phonation time of 1,427.32 s, as calculated by a PRAAT script from de Jong and

Wempe (2009). Thus, on average, our story versions were produced with an average speech rate of 3.25 words/s.

3.2.3. Procedure

3.2.3.1. Experimental set-up

Participants listened to the story while watching the cartoon in a single 30-minute EEG recording session. The cartoon was presented using PsychoPy (Peirce et al., 2019). Participants sat roughly 100 cm from a TV monitor, and they were encouraged to minimize movement and to keep their faces relaxed.

3.2.3.2. EEG recording

Participants were fitted with an electrode cap (actiCAP SnapCap) containing 31 active Ag/AgCl electrodes that were connected to the EEG equipment, Brainvision's actiCHamp Standard 64 System. Two external mastoid electrodes (TP9 and TP10) were placed directly behind participants' ears. The EEG data were recorded at a sampling rate of 500 Hz using Brainvision's Recorder (BrainVision Recorder, Version 1.23.0001, Brain Products GmbH, Gilching, Germany). On average, electrode impedances were kept below 20 K Ω . During recording, the ground electrode was FPz, and the reference electrode was FP1.

3.2.4. Data pre-processing steps and data exclusion criteria

We used the EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) toolboxes in MATLAB (The MathWorks Inc., 2020) to pre-process the EEG data. Our procedure for pre-processing is as follows: First, we re-referenced the data to the average of

the left and right mastoid electrodes. Then, we applied a high pass filter of 0.1 Hz and downsampled the data from 500 to 200 Hz. Next, we extracted 2000 ms epochs from the continuous data between –500 and 1500 ms relative to stimulus onset (without baselining). We then conducted an Independent Component Analysis (ICA) with these epochs in order to identify and correct EEG artifacts (including blinks and horizontal eye movements). ICA components were classified using ICLabel (Pion-Tonachini et al., 2019) and then corrected if they received at least 75% probability of belonging to the following artifact groups: Muscle, Eye, Heart, Line Noise, or Channel Noise. Then, we extracted our target epochs from –200 to 1500 ms with a 200 ms pre-stimulus baseline. These epochs were subjected to an automatic artifact rejection procedure, which removed trials with voltages exceeding $\pm 100 \mu\text{V}$. If necessary, electrode channels with greater than 5% trial loss were interpolated; however, we never interpolated more than 10 channels for a single participant. In fact, we only interpolated 19 channels in total, and on average, less than one channel per participant (*Range* = 0–8 channels interpolated per participant). Of these interpolated channels, only two were used in our final analyses. Finally, we applied a low pass filter of 30 Hz.

All participants with more than 25% of their trials rejected after these cleaning procedures were excluded, and their data were replaced. On average, we rejected 4.75% of participants' total trials (*SD* = 5.29%, *Range*: 0–22.8%). In total, we rejected eight participants: three for data loss, four for failing to attend to the task (e.g. falling asleep), and one for researcher error.

3.2.5. Statistical analyses

3.2.5.1. Determining our spatial and temporal regions of interest

To determine our spatial regions of interest (ROIs), we relied on information from the prior literature and our pilot data. For N400 effects, we selected a centroparietal ROI with eight

electrodes: *Cz, C3, C4, CP1, CP2, Pz, P3, P4*. For P600 effects, we selected a parietal ROI with three electrodes: *Pz, P3, P4*. To determine our temporal ROIs, we used a collapsed localizer technique in which all the conditions being compared were collapsed into one grand average waveform, which was then used to determine the ROI (for discussion, see Luck & Gaspelin, 2017). Researchers vary in how they use the grand average—some rely on visual inspection, some conduct cluster-mass permutation tests on these averages, and some select a time window based on the highest peak within a window (Luck, 2014; Luck & Gaspelin, 2017). We used this last approach: Specifically, we used a 200 ms time window centered on the most negative peak (for N400 effects) and the most positive peak (for P600 effects) in the grand average waveforms.

For analyses that directly compared high cloze and low cloze conditions, we defined two temporal ROIs, creating a grand average for each cloze condition but still collapsing across word type. This is because prior work has shown that predictability can influence the timing of ERP effects (Kutas & Federmeier, 2011; Swaab et al., 2012). For example, N400 effects can emerge earlier for words in predictable contexts relative to unpredictable ones (Brothers et al., 2015).

3.2.5.2. *Linear mixed effects model specifications*

All of the linear mixed effects models were implemented using the *lme4* and *afex* packages in the R statistical computing environment (Bates et al., 2014; R Core Team, 2022; Singmann et al., 2023). All pairwise comparisons were implemented using the *emmeans* package (Lenth et al., 2021). We initially fit our models with the maximal random effects structures justified by our data. If we encountered convergence issues, we simplified the random effects structure until the models properly converged (Baayen et al., 2008; Barr, 2021). Most issues were resolved by constraining the covariance parameters for the random effects to zero (i.e. removing the correlations between

them). But, if this step did not resolve the issues, we began to incrementally drop random slopes (while trying to preserve the slopes for the highest order effects of interest) until the models converged. All effects with an absolute value of t greater than 2 are considered significant (Gelman & Hill, 2006). We follow this convention due to the uncertainty in the field about how to best calculate the appropriate degrees of freedom in linear mixed effects models (Baayen et al., 2008). For the sake of completeness, however, we also report the p -values as calculated by the *lmerTest* package (Kuznetsova et al., 2017). Note, in all of our models, both methods of evaluating significance arrived at the same conclusions.

3.2.5.3. *Outline of our analyses*

For the present study, we pre-registered a set of primary and secondary hypotheses. In the Results section below, we report the findings from five linear mixed effects models that aimed to test those hypotheses.⁸ First, we tested whether the N400 effects to form-similar nonwords (*ceke*) were reduced relative to less similar nonwords (*vake*) in high but not low cloze contexts. Second, we tested whether our P600 effects were sensitive to predictability and word type. Third, we investigated our secondary hypotheses about item-level differences in our ERP effects. To do this, we first analyzed how our N400 effects changed as a function of a word's predictability, e.g., did baseline N400 responses show an inverse linear relationship with cloze probability? Then, we

⁸ In our pre-registration, we outlined a series of analyses that modeled by-participant and by-item ERP averages for each electrode in our region(s) of interest. To do this, we planned to implement mixed effects models with random effects for participants or items, as well as random effects for electrode site. During an external review of this work, however, some questions were raised about whether this approach could properly account for the correlations between individual electrodes. To address these questions, we had two options: First, we could collapse across electrodes, leaving only one average ERP value per condition for each participant or each item. Then, with these averages, we could implement simpler regressions. Alternatively, rather than initially collapsing across participants (for the by-item analyses) or items (for the by-participant analyses), we could just collapse across electrodes and implement linear mixed effects models with random effects for *both* participants and items. This latter option is more commonly used in psycholinguistic research, and so we report the results from these trial-level analyses below. Importantly, however, all statistical approaches resulted in the same pattern of findings and theoretical conclusions.

conducted a set of parallel analyses for our P600 effects. The results of these analyses are briefly discussed in the section below, as well as more thoroughly in the General Discussion.

3.3. Results and Discussion

3.3.1. Visualizing grand average waveforms and topographic maps

In Figure 3.2, we present the grand average waveforms from all centroparietal electrodes (*Cz, C3, C4, CP1, CP2, Pz, P3, P4*) for each word type in both cloze conditions. These waveforms were calculated by first collapsing across items to get six waveforms for each participant, and then collapsing across those by-participant averages. Visual inspection shows robust N400 and P600 effects for all violation conditions regardless of predictability. Critically, in the high cloze condition, the N400 effect is reduced for the form-similar nonwords (*ceke*) relative to the less similar rime nonwords (*vake*) across all electrodes (see Figure 3.2, top panel). In contrast, in the low cloze condition, there is no evidence of a reduction for the form-similar nonwords (see Figure 3.2, bottom panel). Finally, for the baseline conditions (*cake*), the N400 responses are smaller for high cloze words than for low cloze words, as predicted.

In Figure 3.3, we present the topographic maps for the effects of our form-similar and rime manipulations. In high cloze contexts, the reduced N400 for form-similar nonwords can be seen as a weaker negative effect in the 400 ms and 600 ms time windows relative to the rime condition (see Figure 3.3, top panel). Both violations in high cloze conditions also produced large P600s that were similar in magnitude and latency (see 1000 ms and 1200 ms). For low cloze contexts, both violations seemed to elicit similar N400 and P600 effects (see Figure 3.3, bottom panel). In addition, the N400 effects for the high cloze conditions are smaller and emerge slightly earlier than those in the low cloze conditions, as predicted.

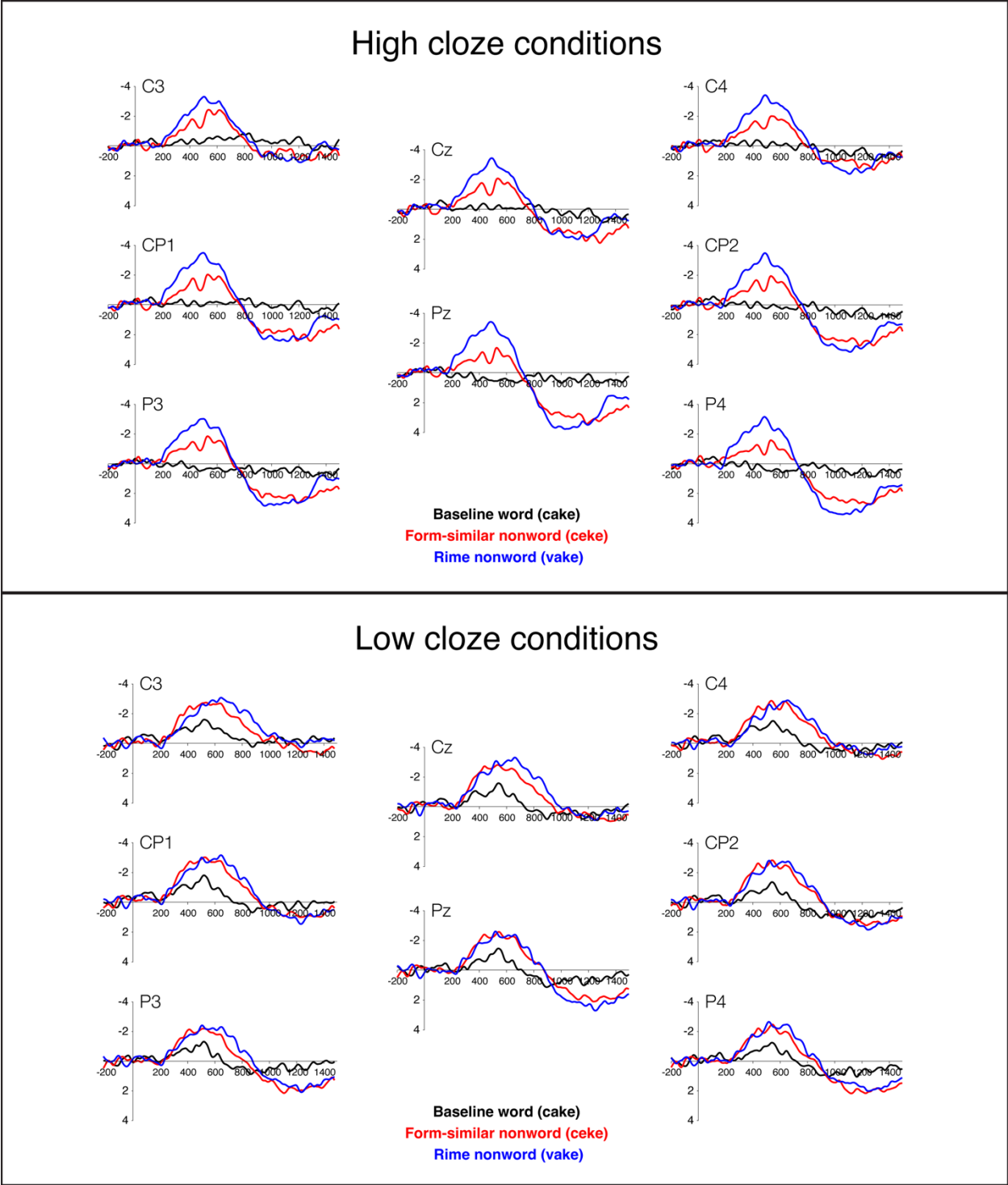


Figure 3.2: Grand waveforms for all word types by cloze condition. The averages (μV) for the centroparietal electrodes of interest are presented for both high cloze (top panel) and low cloze (bottom panel) conditions. The black lines represent the baseline condition (*cake*), while the red and blue lines represent the form-similar (*ceke*) and rime (*vake*) conditions respectively. All waveforms were subjected to an additional low-pass filter of 15 Hz for plotting purposes.

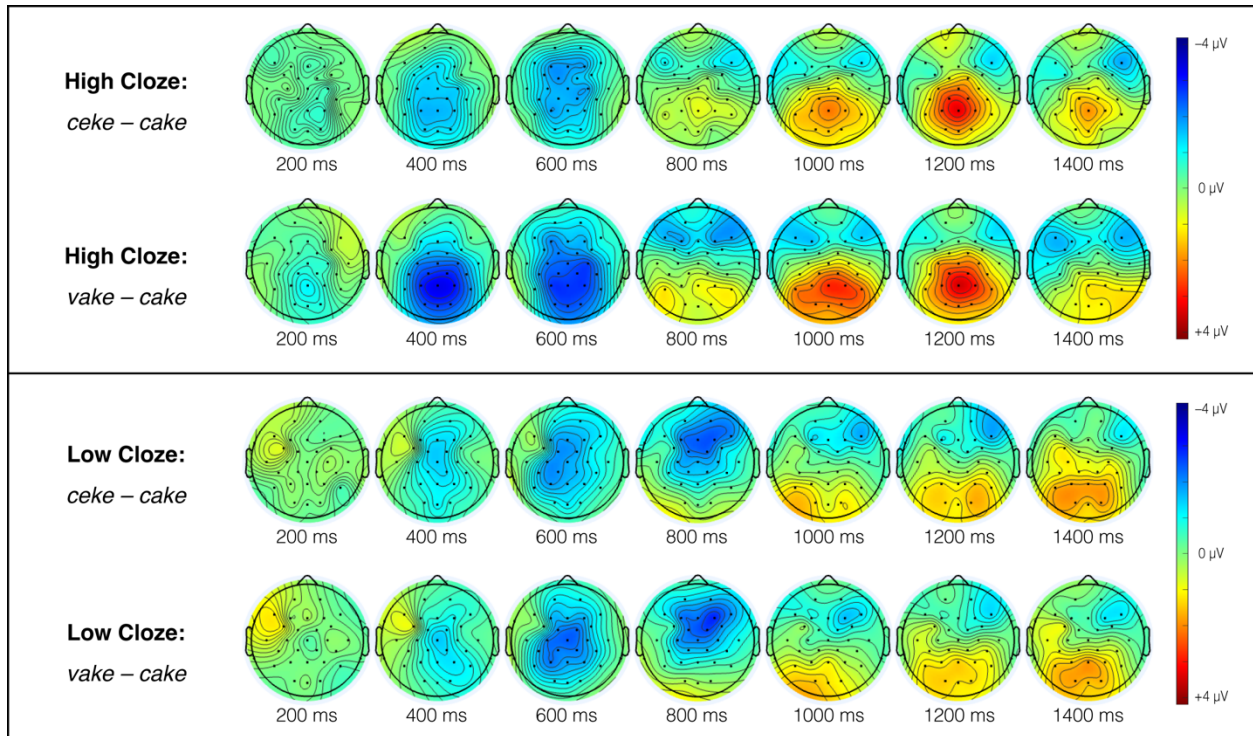


Figure 3.3. Topographic maps of the ERP effects across cloze conditions. These topographic maps depict the isolated effects of the form-similar (*ceke*) and rime (*vake*) nonwords in both predictable (high cloze, top panel) and unpredictable (low cloze, bottom panel) contexts. These effects are calculated by subtracting the baseline ERP activity from the activity evoked by each violation.

3.3.2. Are N400 effects reduced for form-similar errors in predictable contexts?

To demonstrate form-based prediction, we would need to show a significant reduction in N400 effects for form-similar nonwords (relative to the less similar rime nonwords)—but only when participants are actually able to predict the original word from the story. Figure 3.2 provides initial evidence that form-based prediction is occurring, as *ceke* evoked smaller N400 responses than *vake* in high cloze but not low cloze conditions. To confirm that these differences are statistically reliable, we first calculated mean N400 responses by averaging the amplitudes from our pre-registered centroparietal electrodes in the time windows identified by the localizers. These localized time windows were 420–620 ms and 430–630 ms for high and low cloze words respectively. We then modeled these mean N400 responses using a linear mixed effects model.

This model had fixed effects of word type (*cake*, *ceke*, *vake*) and predictability (*high cloze*, *low cloze*), as well as their interaction. For word type, we used contrast coding to test for successive differences between *cake* and *ceke*, and then between *ceke* and *vake*. For predictability, we used contrast coding to test for differences between high and low cloze conditions (*high cloze* = $-.5$, *low cloze* = $.5$). The remaining pairwise comparisons were tested (and corrected for multiplicity) using the *emmeans* package (Lenth, 2021). Finally, this model had random intercepts and maximal slopes for participants and items. To reach convergence, we constrained the covariance parameters for the random effects to zero (see Baayen et al., 2008; Barr, 2021).

Results indicated a significant main effect for the word-type contrast between *cake* and *ceke* ($b = -1.61$, $SE = .40$, $t = -4.01$, $p < .001$). There were no main effects for the contrast between *ceke* and *vake* ($b = -0.67$, $SE = .36$, $t = -1.89$, $p = .067$) nor predictability ($b = 0.51$, $SE = .42$, $t = 1.22$, $p = .23$). There was, however, a significant interaction between word type and predictability, but only for the contrast between the two nonwords, *ceke* and *vake* ($b = -1.52$, $SE = .71$, $t = -2.15$, $p = .032$). To unpack this interaction, we conducted planned pairwise comparisons within predictability conditions, which revealed that the N400 responses for *ceke* were significantly smaller than those for *vake* in the high cloze condition ($b = 1.43$, $SE = .50$, $t = 2.85$, Tukey-adjusted $p = .016$) but not in the low cloze condition ($b = -0.08$, $SE = .50$, $t = -0.17$, Tukey-adjusted $p = .98$). All of the remaining pairwise comparisons within predictability groups were significant.

These results confirmed that there were robust N400 effects for all error types across both high and low cloze conditions; however, the N400 effects for *ceke* were only significantly smaller than those for *vake* in high cloze environments. In contrast, there was no reduction in the N400 effects for *ceke* in the low cloze environments, which suggests that participants were unable to predict the form of the original word in those less predictable contexts.

3.3.3. Do the P600 effects differ across error type and predictability?

To determine whether P600 effects are sensitive to word type and predictability, we followed a similar procedure to the one outlined above: First, we calculated mean P600 amplitudes from our pre-registered parietal electrodes in the time windows from the localizers. Those time windows were 1085–1285 ms and 1160–1360 ms for high and low cloze respectively. Then, we implemented a linear mixed effects regression using the same fixed effects of word type (*cake*, *ceke*, *vake*), predictability (*high cloze*, *low cloze*), and their interaction. For word type, we used contrast coding to test for differences between *cake* and *ceke*, and then between *cake* and *vake*. Note, this last comparison is different from the one used in the N400 analysis above, reflecting a difference in the hypothesis being tested. For predictability, we again compared the two cloze conditions (*high cloze* = $-.5$, *low cloze* = $.5$). Finally, this model had random intercepts and maximal slopes for participants and items. To reach convergence, we constrained the covariance parameters for the random effects to zero.

Results indicated main effects for both word-type contrasts, confirming the robust P600 effects seen in Figures 3.2 and 3.3 for both *ceke* ($b = 1.81$, $SE = .44$, $t = 4.15$, $p < .001$) and *vake* conditions ($b = 2.08$, $SE = .41$, $t = 5.07$, $p < .001$). There were no main effects of predictability ($b = 0.58$, $SE = .56$, $t = 1.04$, $p = .30$) nor any interactions, suggesting that the observed differences in magnitude across high and low cloze conditions were not statistically reliable. We return to this point in Section 3.4.3 of the General Discussion.

3.3.4. Investigating how the N400 and P600 effects vary with word-level predictability

In addition to our primary analyses, we also pre-registered a set of secondary analyses that explore whether word-level predictability modulates our observed effects. In the first analysis, we

sought to replicate the finding that N400 responses for non-manipulated words (*cake*) are inversely and linearly correlated with the predictability of that word given its particular context (Kutas et al., 1984; Kutas & Hillyard, 1984). In the second analysis, we explored how the N400 effects for our violations changed as a function of the original word's predictability. For example, does the N400 reduction for *ceke* increase linearly with predictability (similar to the N400 reduction for *cake*)? In the final analysis, we investigated parallel questions about the relationship between P600s and word-level predictability.

3.3.4.1. Visualizing grand average waveforms across three cloze probability bins

Before addressing these questions, we first visualized how our grand average waveforms changed across cloze probability by categorizing items into one of three bins: Lowest Cloze, Middle Cloze, and Highest Cloze (see Figure 3.4). To create these waveforms, we first averaged across participants to get three waveforms per item. Then, we averaged across these by-item averages within the same cloze bin. Figure 3.4 provides some insight into our three secondary hypotheses above: First, with respect to baseline conditions, the N400 responses become less negative across the cloze bins. Second, the reduction in N400 responses for *ceke* becomes more robust as predictability increases. Interestingly, the size of the N400 response to *vake* seems to slightly increase across predictability, moving in the opposite direction of the form-similar effect. Finally, the magnitude of the P600 effects appears to be similar across error types and across cloze bins, supporting our findings from the primary analyses above.

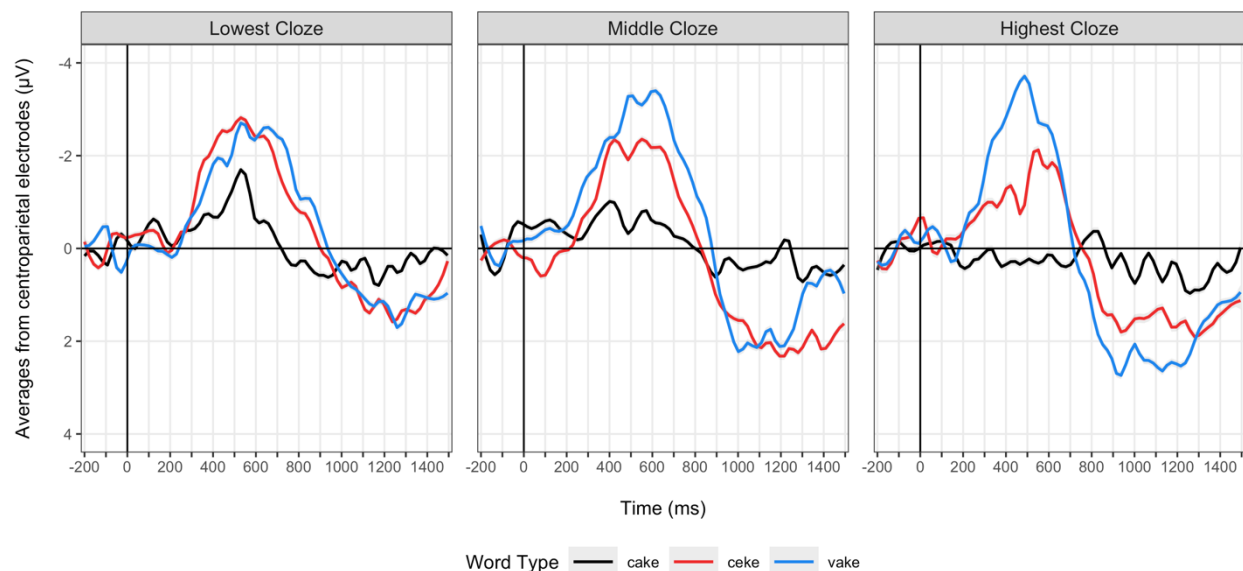


Figure 3.4: Visualization of the ERP effects across cloze probability bins. The grand averages are plotted in three cloze probability bins, increasing in predictability from left to right. The black lines represent the baseline condition (*cake*), and the red and blue lines represent the form-similar (*ceke*) and rime (*vake*) conditions respectively. These waveforms were produced in R and then smoothed using local regression (loess) smoothing techniques.

3.3.4.2. Do baseline N400 responses become smaller as predictability increases?

It is well-documented that the N400 responses to non-manipulated words become smaller (or more positive) as a function of the word's predictability in its particular context (Kutas et al., 1984; Kutas & Hillyard, 1984). Our design allows us to investigate whether this data pattern is present in naturalistic contexts and with stimuli that contain frequent violations. To do this, we modeled the trial-level data from above using a linear mixed effects model with a single continuous fixed effect of cloze probability. For random effects, the model had random intercepts for both participants and items (as justified by the data). Results confirmed a significant effect of cloze probability such that the N400 amplitude for baseline words becomes less negative as predictability increases ($b = 1.77, SE = .76, t = 2.33, p = .021$). Figure 3.5 below shows the change in N400 amplitudes across cloze probability values for all word types.

3.3.4.3. How do the N400 effects for our two error types change across predictability?

Next, we wanted to investigate how predictability modulates the size of the N400 for the form-similar (*ceke*) and rime (*vake*) nonwords. To do this, we originally planned to isolate the effects by calculating difference waves, i.e., subtracting the baseline (*cake*) from both error conditions as in Figure 3.3. However, in constructing these difference waves, we realized that the substantial effect of cloze on the baseline condition makes it difficult to interpret this analysis in isolation. So, although this approach deviates slightly from our pre-registration, we thought it would be informative to test all word types in a single linear mixed effects model to quantify their similarity. Specifically, we implemented a new model with fixed effects of word type (*cake*, *ceke*, *vake*), word-level cloze probability (continuous, 0–100%), and their interaction. As in the primary models, we used contrast coding to test for successive differences between *cake* and *ceke*, and then between *cake* and *vake*. For random effects, the model converged with random intercepts and maximal slopes for participants and items.

Results indicated a significant effect between *cake* and *ceke* ($b = -1.42$, $SE = .58$, $t = -2.45$, $p = .015$), as well as a significant interaction between word type and cloze probability for the contrast between *ceke* and *vake* ($b = -1.81$, $SE = .90$, $t = -2.02$, $p < .05$). To follow up on this interaction, we estimated the slopes for each word type individually using the *emtrends* function (Lenth et al., 2021). For *cake*, we found identical results to the analysis above. For the two error conditions, the slopes did not reach statistical significance—although, *ceke* was estimated to become slightly more positive across cloze ($b = 1.34$, $SE = .71$, $t = 1.89$, $p = .06$) whereas *vake* was estimated to become slightly more negative ($b = -0.47$, $SE = .76$, $t = -0.62$, $p = .54$). These findings tentatively suggest that *cake* and *ceke* experience similar degrees of N400 reduction as

word-level predictability increases. Figure 3.5 shows the relative changes in N400 amplitudes for each word type across cloze probability.

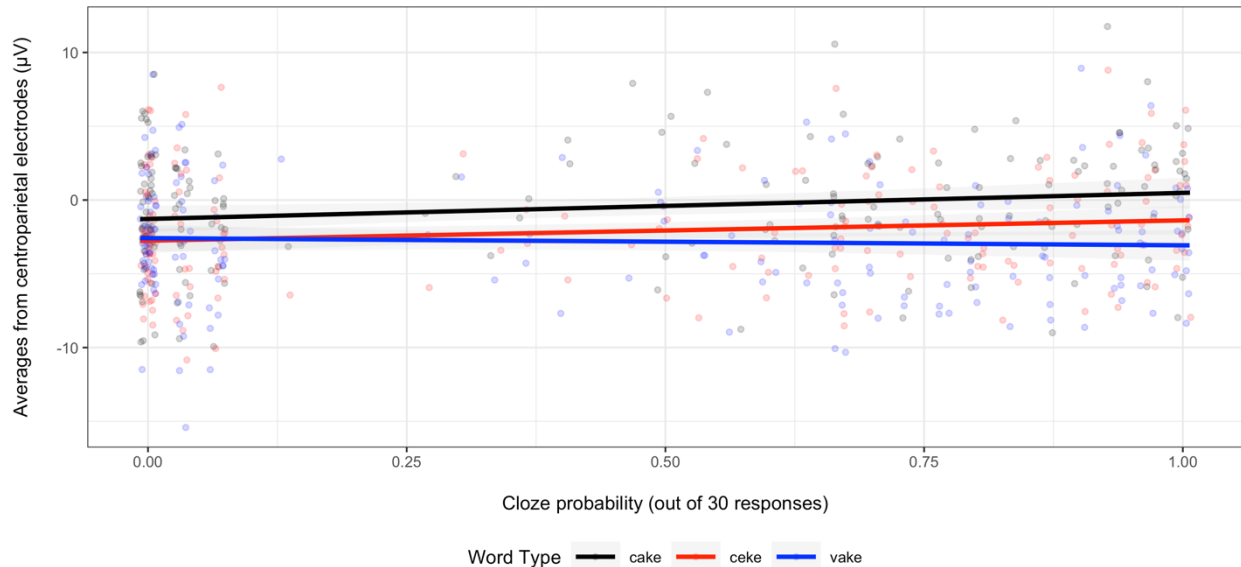


Figure 3.5. Visualization of how N400 amplitudes change across cloze probability for each word type. Each line represents the linear trend in N400 amplitudes as cloze probability increases from left to right. The black line represents the baseline condition, and the red and blue lines represent the form-similar and rime conditions respectively. Each dot represents the average N400 amplitude (μV) for each item (180 items \times 3 word types = 540 observations). Note, we only collapsed across participants for plotting purposes. In the statistical analyses, we preserved the participant and item-level structures.

3.3.4.4. How do the P600 effects for our two error types change across predictability?

We conducted parallel analyses on our P600 effects; however, none of these analyses revealed any significant insights. Specifically, across all of the analyses above, we found robust P600 effects for both violations, with no differences in magnitude between them. Moreover, we found no effects of predictability nor any interactions between predictability and word type. But as we mentioned in the Introduction, the cloze probability of a target word is not necessarily indicative of the constraint of the target sentence. For example, in the Introduction, we used the sentence “I like my coffee with cream and *cinnamon*.” This particular sentence is apt to be completed with *sugar* instead of *cinnamon*, meaning that *cinnamon* is a low cloze completion for

this highly constraining context. In the General Discussion, we present an exploratory analysis of P600 effects in high constraint contexts with low cloze target words. To foreshadow these findings, we show robust P600 effects for all high constraint contexts, regardless of the cloze probability of the target word from the story. These findings are consistent with the proposal that P600s reflect the recognition of a conflict or failure to incorporate the bottom-up input into the comprehenders' higher-level interpretation of the unfolding context—but only when the context is sufficiently constraining such that higher-level interpretations are actually built (e.g. Brothers et al., 2022; Ito et al., 2016; Kuperberg, 2007; van de Meerendonk et al., 2010; van Herten et al., 2005; Vissers et al., 2006).

3.4. General Discussion

In the present study, we directly explored whether adults predict the form of upcoming words while listening to a rich, naturalistic discourse. To do this, we manipulated a set of target words within a children's story such that participants would either hear the original word from the story (*cake*), a form-similar nonword (*ceke* instead of *cake*), or a less similar rime nonword (*vake* instead of *cake*). In highly predictable contexts, we found that form-similar nonwords (*ceke*) elicited smaller N400 responses than rime nonwords (*vake*). This finding suggests that participants had predicted the form of the intended word in high cloze conditions, resulting in facilitated processing of the form-similar nonword. In less predictable contexts, both kinds of nonwords elicited similarly sized N400 responses. To the best of our knowledge, these findings are the first to demonstrate that form-based prediction occurs not only in tightly controlled experimental settings but also in the context of a rich natural discourse.

In the remainder of the General Discussion, we will do four things: First, we reconcile our findings with the prior literature suggesting that form-based prediction should *not* readily occur in naturalistic comprehension (Section 3.4.1). Second, we explore how our work relates to prior studies on phonological mismatch effects in auditory language comprehension (Section 3.4.2). Third, we further examine our P600 effects and integrate our findings with the broader literature on these posterior positivities (Section 3.4.3). Finally, we discuss several open questions about form-based prediction and outline potential avenues for future research on this phenomenon using the *Storytime* paradigm (Section 3.4.4).

3.4.1. Should form-based prediction be expected during naturalistic comprehension?

There is an ongoing debate in the literature about the limits of form-based prediction and its role in everyday comprehension. Some researchers argue that, while form-based prediction can occur in certain experimental contexts, it is unlikely to occur in most naturalistic settings (Freunberger & Roehm, 2016; Ito et al., 2016; Ito, Martin, & Nieuwland, 2017). Specifically, there are two features that are unusual about the experimental contexts in which form-based prediction has been studied:

First, the manipulations that are used in such studies can help create a context in which making predictions about an upcoming word is unusually useful for the comprehender. If all the target sentences contain a highly predictable word, and that word is the one that is manipulated or replaced, it would make sense to try to anticipate these words (rather than passively process them) in order to reconstruct the intended (or original) meaning of the utterance. Second, in form-based prediction studies, the target sentences are often presented at slower-than-natural rates, which may provide comprehenders with more time to process the unfolding words and use them to generate

predictions (for discussion, see Ito et al., 2016; but see, DeLong et al., 2021). In the sections below, we ask whether these two features are also true of the stimuli in the present study. To foreshadow our results, we do not find any evidence that our story is particularly slow or predictable. Thus, we end with a proposal about why form-based prediction was possible in this rich but highly variable naturalistic context.

3.4.1.1. How predictable was our story?

Prior work has found that comprehenders engage in more predictive processing when they are in highly predictable contexts (e.g. Brothers et al., 2015; Lau et al., 2013). For example, in a study by Lau et al. (2013), participants simply read pairs of words and indicated when they saw the name of an animal. The authors manipulated these word pairs to either be semantically related to one another (e.g. salt–PEPPER) or not related at all (e.g. salt–UNCLE). They also manipulated the overall proportion of trials in which the word pairs were related: In one experimental block, 50% of the word pairs were related, so that activating close semantic associates would be helpful about half of the time. In the other block, only 10% of the word pairs were related, and thus, participants who activated semantic associates would be unlikely to correctly predict the second word given the first. As expected, the authors observed a reduction in the N400 for words that were preceded by semantically-related words (e.g. salt–PEPPER) relative to words preceded by unrelated words (e.g. salt–UNCLE). Critically, however, this reduction was greater in the experimental block with a higher proportion of related trials relative to the block with a lower proportion, demonstrating that pre-activation based on semantic association is greater when semantic predictability is high.

In the present study, we explored comprehension in the context of a written narrative that was read aloud. Like most written narratives, this story had been carefully crafted and edited with the goal of ensuring that it would be easily understood. Edited language like this is unnatural and yet pervasive—unnatural in the sense that it is quite different from the language environment that existed for most of humans’ evolutionary history, and pervasive in the sense that it now makes up a sizeable portion of our participants’ linguistic input on any given day (e.g. podcasts, movies, news broadcasts, articles, novels, and many social media posts). One might wonder if, on average, edited text is more predictable than spontaneously produced language—especially when the edited text is intended for children, like the story used in the present study.

Fully exploring this question goes beyond the scope of the present paper. Nevertheless, to better understand the scope and generalizability of our findings, we revisited our cloze data to characterize the overall predictability of our story. In our original cloze task (see Section 3.2.2.2), each person would read the story up until they had to guess, and then they would guess each word in their small section, word-by-word (e.g. *The...*, *The cat...*, *The cat was...*, *The cat was hungry*). Participants saw the correct answer on the screen after making each guess. For example, a person might have guessed that a sentence started with *The*. After guessing, they learn that it started with *She*, and so they update their expectations about how the sentence will unfold and then guess the word *went*.

This procedure allowed us to collect incremental cloze values for every word in our 30-minute story. Using these values, we characterized the overall predictability of all the content words (nouns, verbs, and adjectives) in our story. Given the design of the present study, we knew that there would be some words that were highly predictable and others that were not. The question for this analysis, however, was how often the content words were *highly* predictable. Figure 3.6

shows the distribution of cloze probabilities for all 1,947 content words in our story. The median cloze value was 6.7% (out of 30 total guesses). There were 639 words (roughly 33% of all words) that no participant correctly guessed. Thus, even in this simple edited narrative, most words could not be predicted before they were encountered.

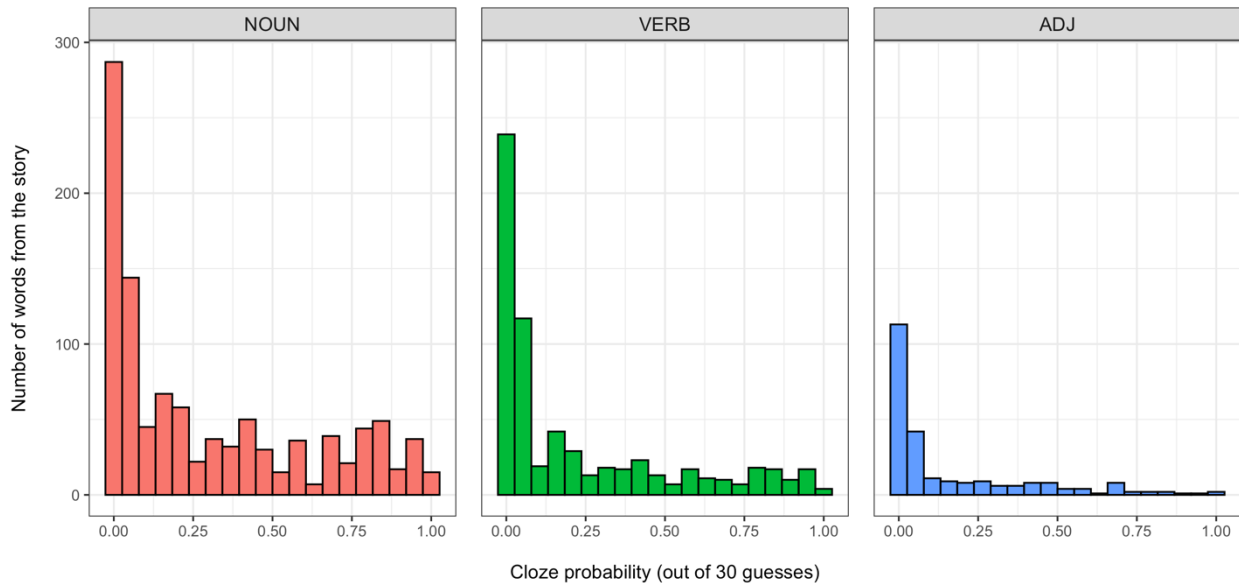


Figure 3.6: *Distribution of the cloze probability values for all nouns, verbs, and adjectives in our story.* We calculated the cloze probability for all 1,947 content words in our story stimulus. These words are presented by syntactic category: noun (left), verb (middle), or adjective (right). In each panel, the bars represent the total number of words within each cloze bin (ranging from 0–100%).

In light of these findings, we conclude that form-based prediction can emerge for highly predictable words despite these words occurring in a broader, more variable environment with generally unpredictable words. Although we cannot ascertain whether our story is more or less predictable than other forms of natural language, we can compare these cloze values with those from the stimuli used in prior psycholinguistic studies. Prior studies that manipulate predictability typically have the following classification: words with cloze values under ~10% are low cloze, words with values up to ~65% are medium cloze, and words with values over 65% are high cloze (Block & Baldwin, 2010; Brothers & Kuperberg, 2021; Kutas & Hillyard, 1984). By this criterion,

the content words in our story were mostly low cloze (52.2%) with many medium cloze words (31.2%) and a smaller number of high cloze words (16.6%). The distribution of cloze values in our stimuli also appears to be broadly similar to the values observed in cloze studies using written passages intended for adults (Lowder et al., 2018; Luke & Christianson, 2016; Smith & Levy, 2013). In sum, these analyses suggest that our story was not unusual in its degree of predictability, and thus, it is unlikely that the effects we observed were driven by a strategy that is specific to the materials that we used.

3.4.1.2. How slowly was our story read?

Most researchers would agree that making top-down predictions during comprehension takes some amount of time (e.g. DeLong et al., 2021; Freunberger & Roehm, 2016; Ito et al., 2016; Pickering & Gambi, 2018; Pickering & Garrod, 2007). To predict the form of an upcoming word, comprehenders must do several things: First, they must perceive the earlier words in an utterance and use them to make inferences at higher conceptual levels about what is likely to come next. Then, after making those inferences, they must transmit information back down to lower levels in order to pre-activate 1) the relevant lexical concepts and 2) the form features associated with them (in that order). These steps constitute a feedback loop in which information from the perceived input is propagated to higher levels and then back down again (Dell, 1986; Pickering & Garrod, 2007).

Decades of research on incremental processing has provided insight into how (and when) we should expect to see substantial effects of feedback loops: First, feedback signals should emerge gradually over time, becoming stronger as more time passes (Dell, 1986). Second, in a system with hierarchical, multi-layered representations, it should take longer for top-down

information to reach lower levels than higher ones (Elman & McClelland, 1988; Indefrey & Levelt, 2004; Pickering & Garrod, 2013; Rumelhart & McClelland, 1982). On this account, we should expect predictions about upcoming forms to emerge later in time than predictions about upcoming meanings or concepts, as form-based prediction requires the same steps as semantic prediction, plus the additional step of pre-activating phonological and perceptual features. Taken together, these insights have generated skepticism about whether form-based prediction readily occurs during ordinary comprehension (e.g. Freunberger & Roehm, 2016; Indefrey & Levelt, 2004; Ito et al., 2016; Ito et al., 2017a; Pickering & Gambi, 2018; Pickering & Garrod, 2013). If we assume that predictions are largely driven by the word immediately before the target, and that this prior word lasts between 200–400 ms, then comprehenders have only a few hundred milliseconds to identify the word, generate expectations about the next word, and pre-activate its form.

As we noted in the Introduction, there is some empirical evidence to support this claim. Specifically, Ito et al. (2016) found reduced N400 responses to unexpected, form-similar words at a slow presentation rate (1.5 words per second) but not at a faster one (2 words per second). Given that natural speaking and reading rates are between 3–5 words per second (Brysbaert, 2019; Tauroza & Allison, 1990), these findings suggest that form-based prediction should rarely occur in everyday contexts.

The stimulus in the present study was an audio recording of a children’s story that we intend to also use with child participants. Because child-directed speech is often, but not always, slower than speech directed to adults (e.g. Biersack et al., 2005; Fernald et al., 1989; Ratner, 2013), one might wonder whether our story was so slow that it allowed adults to pursue a predictive strategy that would not ordinarily be available to them. In Section 3.2.2.7, we calculated the average speech rate for our stories to be 3.25 words per second, which falls within the range of 3–

5 words per second for natural adult-directed speech (Tauroza & Allison, 1990) and well within the range for studies of spoken and written language comprehension in adults (see Dambacher et al., 2012; DeLong et al., 2020; Ito et al., 2016; Wlotko & Federmeier, 2015).

Moreover, other studies using similar designs and paradigms to Ito et al. (2016) have also found strong evidence for form-based prediction at faster rates (DeLong et al., 2019, 2021; Kim & Lai, 2012; Laszlo & Federmeier, 2009). For example, DeLong et al. (2021) had adults read sentences word-by-word at a rate of 4 words per second. All of their sentences originally had a highly predictable noun, which they either kept or replaced with an orthographically similar word, a semantically similar word, or an unrelated word (e.g. “The Doberman stood its ground and bared its (*teeth* / *tenth* / *dentist* / *report*) to the mailman”). The authors found smaller N400s for the orthographically and semantically similar words relative to the unrelated words—despite presenting their sentences at nearly twice the speed of Ito et al. (2016).

Despite this clear pattern of findings, we are inclined to agree with the skeptics: it seems implausible, given what we know about the mind, that a listener can identify a word as it unfolds, integrate it into higher level discourse structure, make a prediction for the next word, and pre-activate the form of that next word—all in the span of ~300 ms. So, how can we explain the findings from our study, as well as those from the prior studies that report form-based prediction at fast presentation rates?

3.4.1.3. Getting a head start on form-based prediction

One way to reconcile these findings with our understanding of the temporal properties of feedback loops is to assume that form-based prediction is not typically triggered by the immediately preceding word. Instead, the words that appear earlier in the context may generate

specific lexical expectations that will be fulfilled several words downstream—we will call this *long-distance* prediction. If these long-distance predictions can be made in parallel with the bottom-up processing of each subsequent word, then such a system could allow predictions to emerge gradually during comprehension. In the rest of this section, we review the preliminary evidence for long-distance prediction and then explore the degree to which this phenomenon may account for the divergent findings in prior work. To do this, we investigate whether long-distance prediction might have been possible for the critical words in the present study, as well as two prior studies on form-based prediction.

3.4.1.3.1 Preliminary evidence for long-distance prediction

The clearest evidence for long-distance prediction comes from a recent study exploring form-based prediction in the visual world paradigm. Li et al. (2022) asked native Mandarin speakers to look at visual displays while listening to highly constraining sentences like “After school, I put my pencil case and notebooks into my **schoolbag** and get ready to go home.” The visual displays always contained four objects, positioned in the four quadrants of the screen. Three of these objects were unrelated distractor objects that shared no semantic or phonological features with the highly predictable noun (e.g. *schoolbag* in the sentence above). The fourth object was either the highly predictable noun (e.g. a schoolbag), a semantic competitor (e.g. an eraser), a phonological competitor (e.g. in Mandarin Chinese, *comb* and *schoolbag* have the same first syllable and tone), or an unrelated distractor (e.g. funnel).

Li et al. (2022) found that participants began looking to the semantic and phonological competitors (over the distractors) well before the target word was produced (see also Ito et al., 2018). Specifically, the authors observed increased looks to both competitors starting ~1400 ms

(or two words) before the target word onset. This pattern was interpreted as evidence that form-based predictions were made well in advance of the target word. To determine whether long-distance predictions were possible in the context, Li et al. (2022) conducted an exploratory cloze task with a different set of Mandarin speakers. These participants heard the original target sentences; however, the authors truncated them ~1400 ms before the target word. Participants were then asked to complete each of the truncated sentences. Results indicated that participants included the target words in their completions at rates well above chance (e.g. average target cloze probability was 33% with a range of 20–45%). Thus, it is clear that, under some circumstances, listeners can make predictions about an upcoming word on the basis of information presented much earlier in the sentence—and moreover, these predictions can result in the pre-activation of form features well over a second before the predicted word is produced.

3.4.1.3.2 Assessing long-distance prediction in the present study and two prior studies

This finding raises the possibility that the reduced N400s in form-based prediction studies sometimes result from predictive processing that occurs well before the pre-target word. Specifically, we might expect that the studies showing form-based prediction at faster stimulus presentation rates have stimuli that support long-distance prediction, while the studies showing form-based prediction only at slow presentation rates may not. To explore this, we revisited the stimulus sets used in DeLong et al. (2021) and Ito et al. (2016). DeLong et al. (2021) reported form-based prediction at presentation rates of 4 words per second, whereas Ito et al. (2016) found that form-based prediction broke down at rates of 2 words per second. In addition, we also explored a subset of our target sentences to see if they provided support for long-distance prediction.

To investigate this systematically, we conducted a set of exploratory cloze tasks that presented participants with truncated versions of the high cloze sentences from these three studies (160 high cloze sentences from DeLong et al., 2021; 88 from Ito et al., 2016; and 25 from the present study).⁹ For the two reading studies, we presented each target sentence three times, providing additional context each time. For example, we first showed participants a truncated sentence that stopped four words before the target (*The lumberjack chopped...*) and asked them to complete the sentence. Then, we revealed two words (*The lumberjack chopped the wood...*) and asked for another completion. Finally, we presented the entire sentence up to the target word (*The lumberjack chopped the wood with his...*) to determine whether participants could guess the target word *ax*. For the materials from the present study, we implemented a similar procedure: the cartoon would play until reaching four words before the target and then it would pause. Participants were instructed to then complete the current sentence before receiving more context from the story. This procedure was identical to the one for the two reading studies; thus, we were able to collect comparable incremental cloze values for these various high cloze contexts.

Figure 3.7 shows the by-item cloze probabilities from each study at the three different distances. In all three studies, participants readily predicted the target words right before they appeared in the input: the present study had an average cloze of 92.4% ($SD = 16\%$); the DeLong study had 91.8% ($SD = 12.5\%$); and the Ito study had 88.7% ($SD = 16.4\%$). Long-distance prediction rates varied across the studies. The target words from the present study could often be predicted at distances of four words ($M = 38.9\%$ cloze, $SD = 29.7\%$) and two words ($M = 69.8\%$, $SD = 27.8\%$) before they appeared. The two reading studies showed less long-distance prediction;

⁹ Given the nature of our spoken story, we selected 25 high cloze target words (out of 90) for participants to complete online. For this truncated cloze task, all target sentences needed to have a minimum of 5 words before the target word. We did not use all viable sentences from our stimulus because participants needed to watch an entire cartoon rather than simply read the sentence and having them guess 90 target words dramatically extended the duration of the study.

however, the target words in both studies were still predicted at rates above 20% for all distances. The DeLong targets were numerically more predictable across all distances (4 words prior, $M = 23.3\%$, $SD = 30.4\%$; 2 words prior, $M = 51.9\%$, $SD = 35.5\%$) than the Ito targets (4 words prior, $M = 20.1\%$, $SD = 27.9\%$; 2 words prior, $M = 48.8\%$, $SD = 33.1\%$); however, we do not believe that this constitutes a categorical difference in the degree of long-distance prediction in the reading studies.¹⁰

In short, these findings provide additional evidence for long-distance prediction, generating a slew of questions about the types of systems that can generate predictions in advance and maintain them as more words are being perceived. We suspect that this kind of long-distance prediction is common in rich naturalistic contexts like our story. Narratives typically follow characters across events and revisit the same topics, making it useful to track (at many levels) the ideas, objects, people, and places that are likely to be mentioned. The cues that allow us to make these predictions may occur a few words, a few sentences, or even a few paragraphs in advance of the target word, allowing for prediction to occur even at quite rapid presentation rates.

¹⁰ Given these findings, one might wonder why the participants in the Ito study did not show form-based prediction at faster presentation rates, while those in the DeLong study did. We see three possible explanations: First, the small difference in both immediate predictability and long-distance predictability led to more rapid prediction of the critical items in DeLong; Second, because the high predictability sentences formed a much smaller portion of the Ito stimulus set, participants may have engaged in less predictive processing in the Ito study. Third, the discourse constraints allowing for prediction in the Ito study may be more complex and involve higher-level conceptual relations that are only calculated when more attentional resources are available (as in our offline cloze task) while the discourse constraints in DeLong might be more associative and less dependent on attention.

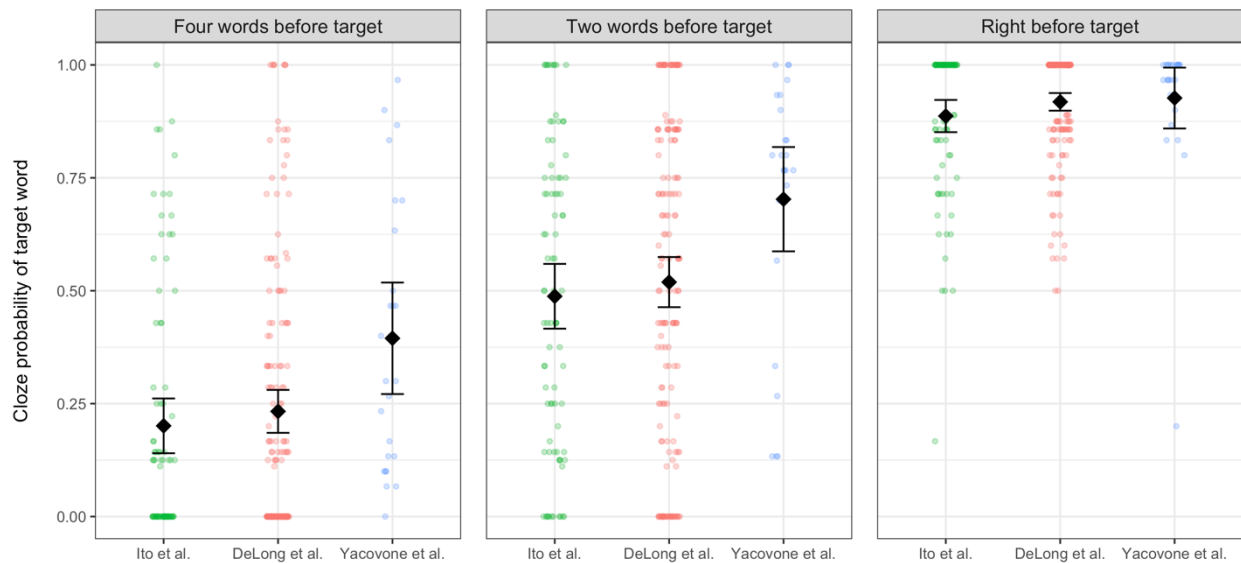


Figure 3.7: *Investigating long-distance prediction in three studies.* The predictability of target words across three time points from Ito et al., 2016; DeLong et al., 2019, 2021; and Yacovone et al. (the present study). The three time points correspond to the amount of context encountered before that target word. If the sentence was *I like my coffee with cream and...sugar*, participants would get “I like my...” (four words before target), “I like my coffee with...” (two word before target), and then “I like my coffee with cream and...” (right before target). Cloze probabilities were then determined by calculating how often participants provided the target word (sugar) in their sentence completions at each time point.

3.4.2. How do our findings relate to prior work on phonological mismatch effects?

In the Introduction, we reviewed a handful of studies that investigated the effects of phonological mismatch during spoken language comprehension. In these studies, participants listened to constraining sentences with highly predictable sentence-final words, e.g., “It was a pleasant surprise to find that the car repair bill was only seventeen *dollars*.” On some trials, these predictable words appeared as expected. On other trials, these words were replaced with semantically incongruous words that either shared the same initial phonemes with the expected word (*dolphins*) or did not (*scholars*). Across these studies, there are three main findings that must be reconciled with the present study, each of which we address below.

First, a handful of studies report early negativities that are argued to be distinct from the N400 for two reasons: 1) they typically emerge and peak earlier than canonical N400s (200–300

ms following the onset of a violation with unexpected initial phonemes) and 2) there is sometimes an apparent temporal discontinuity between an early peak and the peak of the N400 (e.g. Connolly & Phillips, 1994; van den Brink et al., 2001). In the present study, we did not show an early negativity with a peak that was spatially and temporally distinct from our N400 responses. This is not terribly surprising, as auditory ERPs rarely show distinct early effects due to the variability in how spoken words are produced and how they unfold over time (Holcomb & Neville, 1991; Kutas et al., 1987, 2006; Kutas & Van Petten, 1994; Swaab et al., 2012). In a naturalistic stimulus like ours, there is a high level of variability in the onsets of each word, the prosody of each utterance, and the timing between each subsequent word. This variability has the potential to blur together any early effects by increasing noise and interfering with participants' abilities to precisely predict when a word will begin and how it might sound. If we were to use these materials but tightly control the timing of each spoken word (e.g. playing sentences word-by-word or introducing gaps between them), it is possible that we might find distinct early effects (see Kutas & Federmeier, 2011). In our naturalistic *Storytime* paradigm, however, this level of control is not possible, and thus, one limitation of this particular approach is its reduced sensitivity to small, early, and/or short-lived ERP effects.

Second, the studies on phonological mismatches during auditory language comprehension typically find a delay in the mismatch effect for violations that share initial phonemes with the expected word (e.g. Liu et al., 2006; Petten et al., 1999). For example, Van Petten et al. (1999) found a later N400 effect for unexpected words that shared initial phonemes with the predicted word (e.g. *dolphins* when expecting *dollars*) relative to those that did not (*scholars* when expecting *dollars*). This pattern is consistent with a large body of evidence demonstrating that people incrementally interpret spoken words, restricting their hypotheses about the word's identity as it

unfolds (e.g. Allopenna et al., 1998; Marslen-Wilson, 1987; Marslen-Wilson & Zwitserlood, 1989). In a predictive system with incremental interpretation, we should expect to see early violation effects when the initial phonemes violate our predictions (*scholars*), and later effects when the initial phonemes are consistent with our predictions (*dolphins*).

Thus, it is somewhat surprising that we did not find any latency differences between our form-similar (*ceke*) and less similar rime (*vake*) violation effects in either cloze condition. In high cloze environments, the N400s for both the *ceke* and *vake* conditions began to diverge from the baseline at ~200 ms (see Figure 3.2). The only difference between these two effects was in overall amplitude, as the N400 for *ceke* was always smaller than the N400 for *vake* in predictable contexts. In low cloze environments, we did not see any significant differences in amplitude, latency, or scalp distribution for the nonword N400s.

We suspect that these patterns reflect two features of the present study: First, as we mentioned above, the inherent variability when using the *Storytime* paradigm limits our ability to detect small, short-lived effects. Second, because the present study focused on form-based prediction rather than incremental interpretation, both nonword conditions diverged from the target word quite early. Studies of incrementality in spoken language have generally used cohort competitors that have prolonged phonological overlap with the target or expected word (e.g. Allopenna et al., 1998; Liu et al., 2006; Petten et al., 1999). For example, Van Petten et al. (1999) had an onset-overlap condition that shared an entire syllable with the expected word (*dollars* vs. *dolphins*). In contrast, their rime-overlap condition immediately mismatched the expected word in initial phonemes (*dollars* vs. *scholars*). Thus, we estimate that Van Petten et al. had a period of roughly 300 ms (out of a total word duration of ~585 ms) in which the rime-overlap violation was detectable, but the onset-overlap violation was not.

In contrast, the present study did not have long periods of time in which listeners could detect the rime nonwords but not the form-similar nonwords. The form-similar violations only shared an initial consonant (or consonant cluster) with the target word and then diverged at the initial vowel (e.g. *cake* vs. *ceeke*). The rime violations had the opposite pattern, diverging from the target word at the initial consonant(s) but sharing an initial vowel (e.g. *cake* vs. *vake*). As a result, the form-similar violations could presumably be detected shortly after the release for stop consonants (or even on the basis of co-articulatory cues for non-stop consonants). In other words, the time needed to disambiguate *cake* from *ceke* and *cake* from *vake* probably differs by tens of milliseconds rather than hundreds like in prior studies. Thus, any delay in violation effects for *ceke* would have been short-lived, making it difficult to detect using the present paradigm.

The final issue to explore is why some spoken language studies demonstrate reduced N400s for onset-overlap violations (e.g. van den Brink et al., 2001) while others do not (Van Petten et al., 1999). In contrast to the present study, Van Petten et al. (1999) did not find reduced N400s to onset-overlap violations (relative to rime-overlap violations) after controlling for differences in disambiguation points. This prior study and our study used nearly identical violation designs but found different patterns of results—thus, we must attempt to reconcile these two studies to better understand the conditions in which form-based prediction occurs.

One clear difference between these two studies is the use of nonword vs. real-word violations (in addition to the clear differences in disambiguation points described above). When a listener encounters a nonword like *ceke* in an environment that is highly constraining for the word *cake*, they may be apt to simply interpret this input as being consistent with *cake*. In contrast, if a listener encounters a real-word violation like *dolphins* in a context that predicts *dollars*, it may be more difficult to recover the intended meaning of the utterance and re-cast *dolphins* as meaning

dollars. Thus, if the size of the N400 reflects the amount of additional processing needed to override a prior prediction, activate the newly perceived lexicosemantic features, and integrate the unexpected word into the context, then it makes sense that *dolphins* and *scholars* evoke similarly sized N400s. However, if the violation neither brings new lexicosemantic features nor strongly disconfirms a prior prediction (e.g. *ceke*), then it should be processed more similarly to the expected word than something unexpected. This hypothesis predicts that nonword and real-word violations should be processed differently in spoken language comprehension.

Preliminary support for this hypothesis comes from a study of spoken language comprehension in Mandarin Chinese. In this study, Liu et al. (2006) had two experiments: the first used real-word violations similar to Van Petten et al. (1999), whereas the second used nonword violations similar to the present study:

In Experiment 1, native Mandarin speakers listened to sentences with highly predictable endings, e.g., “The sound in the radio became weaker and weaker. It seems that I must buy several new sets of *batteries*.” Similar to prior work, the authors manipulated the last word to be the expected word (*batteries*, in Mandarin /*dian4*-/*chi2*/) or one of three real-word violations: an onset-overlap violation (*electric stove*, /*dian4*-/*lu2*/), a rime-overlap violation (*water pool*, /*shui3*-/*chi2*/), or a no-overlap violation (*illness*, /*bing4*-/*tai4*/). They found increased N400 amplitudes for all violations relative to the expected word—and critically, the timing of these effects differed from one another, but the overall amplitude did not. Specifically, the N400 effect for the onset-overlap (/*dian4*-/*chi2*/ vs. /*dian4*-/*lu2*/) emerged later than the N400 effects for the other violations.

In Experiment 2, another set of native Mandarin speakers listened to both predictable and unpredictable sentences. The authors manipulated the last word in these sentences to be one of

four conditions: *the original word* from the sentence; a *minimal-onset-mismatch* (a nonword with an onset that mismatches the original word in one or two features); a *maximal-onset-mismatch* (a nonword with an onset that mismatches the original by two or more features); and a *first-syllable-mismatch* (a nonword with a completely different first syllable than the original word, i.e., the rime condition from prior studies). Similar to the present study, Liu et al. found N400 effects to all nonword violations—however, in predictable contexts, the size of the N400 effect depended on the degree of form similarity to the expected word: the N400s for *original word* < *minimal-onset-mismatch* < *maximal-onset-mismatch* < *first-syllable-mismatch*. In the unpredictable contexts, all three nonword violations produced similarly sized N400 effects. Taken together, these findings tentatively support the hypothesis that real-word and nonword violations produce different ERP effects in spoken language comprehension.

3.4.3. How do our findings advance understanding of late posterior positivities?

In the Introduction, we mentioned that posterior P600s often accompany N400 effects in studies of form-based prediction. We replicated this finding in the present study; however, our results slightly diverge from the prior literature in two ways:

First, we observed P600 effects for *both* form-similar (*ceke*) and less similar (*vake*) violations, and these effects were similar in magnitude. Prior work has shown that the degree of form similarity between a violation and an expected target word influences the size of the P600 (e.g. Ito et al., 2016; Laszlo & Federmeier, 2009; Ryskin et al., 2021; Vissers et al., 2006). So, why might our two violations elicit similarly sized P600s? While *cake* may seem more similar to *ceke* than *vake* when processed incrementally, both violations only differ from the target word by one or two phonemes. After all, we decided to use rime violations in order to allow listeners to

recover the intended meaning of utterances with violations in them. Thus, a simple explanation for why the P600s may be similar in size for both *ceke* and *vake* could be that the P600 reflects a process that occurs *after* bottom-up processing is complete (for similar discussion, see Ito et al., 2016). If the processes indexed by the P600 are *reactive* rather than predictive, both *ceke* and *vake* can be construed as slight deviations from a more congruent continuation (*cake*). On this account, we might expect that comprehenders will face similar difficulties when attempting to incorporate these two nonwords into their higher-level interpretations, and when reprocessing these violations in order to assess the nature of the anomalous input.

Second, we observed P600 effects to all violations regardless of the predictability of the target word. Prior work has demonstrated that P600s are more robust in high constraint contexts (Gunter et al., 2000; Kuperberg et al., 2020; Van De Meerendonk et al., 2009; van de Meerendonk et al., 2010). This finding has led some researchers to argue that the P600s in the form-based prediction literature reflect comprehenders' interpretation of the violations as misspellings of the predicted word (which also explains why P600s are not readily observed in low constraint contexts, see Vissers et al., 2006). So, why did we observe robust P600 effects in our less predictable conditions? We see two possible explanations for this pattern.

The first possibility is that the P600 simply reflects the initial failure to incorporate the bottom-up input into one's high-level interpretation of the context, as well as the subsequent disruption to ongoing comprehension caused by reprocessing the anomalous input (Brothers et al., 2020, 2022; Kuperberg, 2007; Kuperberg et al., 2020). On this account, comprehenders do not need to have strong top-down expectations about what is coming next in the sentence—however, they must be actively engaged in deep comprehension to experience a disruption from the violation (Brothers et al., 2020). Typically, low constraint contexts do not readily provide rich details about

an unfolding event; thus, deep comprehension may be difficult to achieve in these kinds of sentences. In the present study, however, the unpredictable sentences are embedded within a larger discourse. So, although a sentence may be unpredictable at the local level, it is still contributing to the understanding of the broader context. On this account, we would expect P600s to violations in these unpredictable sentences because comprehenders are deeply processing the linguistic material, and the violations are disrupting the ongoing construction of the narrative.

The second possibility is that there are some sentences in the low cloze group that are actually highly constraining for a different word than the one in the story. For example, if someone said, “For my birthday, I am going to bake a large *pie*,” you would not consider the sentence to be anomalous despite it violating your expectations. In these scenarios, the context is generating strong constraints for a particular continuation (e.g. *cake*); however, the speaker never produces the predicted word. Thus, if the P600 is primarily sensitive to anomalous (or highly implausible) continuations in high constraint contexts, then we should expect to see robust P600s in our low cloze environments when they strongly constrain for an unobserved alternative word.

To investigate this, we characterized the constraint of all sentences containing target words (regardless of the cloze probability of the word from the story). For this analysis, we focused on low cloze target words (however, Figure 3.8 presents these exploratory findings alongside the high cloze condition). All of our low cloze items had target words with values less than 50% (out of 30 total responses). To measure the constraint for each item, we calculated the cloze probability of the most frequent response produced by participants in our cloze task. We then grouped these items using a median split approach: Sentences with constraint values greater than 40% were classified as high constraint, and those with values less than 40% were classified as low constraint.

In Figure 3.8, we visualize the grand waveforms for three groups of words: high cloze target words in high constraint contexts, e.g., “I like my coffee with cream and *sugar*” (left panel); low cloze target words in high constraint contexts, e.g., “For my birthday, I wanted to bake a large *pie*” (middle panel); and finally, low cloze target words in low constraint contexts, e.g., “They are only found in a certain *river*” (right panel). These waveforms revealed robust P600 effects for all violation conditions in high constraint contexts (regardless of the cloze probability of the target word). These observations tentatively support the claim that the P600 effects in our low cloze conditions (see Figure 3.2, bottom panel) were largely driven by sentences with high cloze competitors.

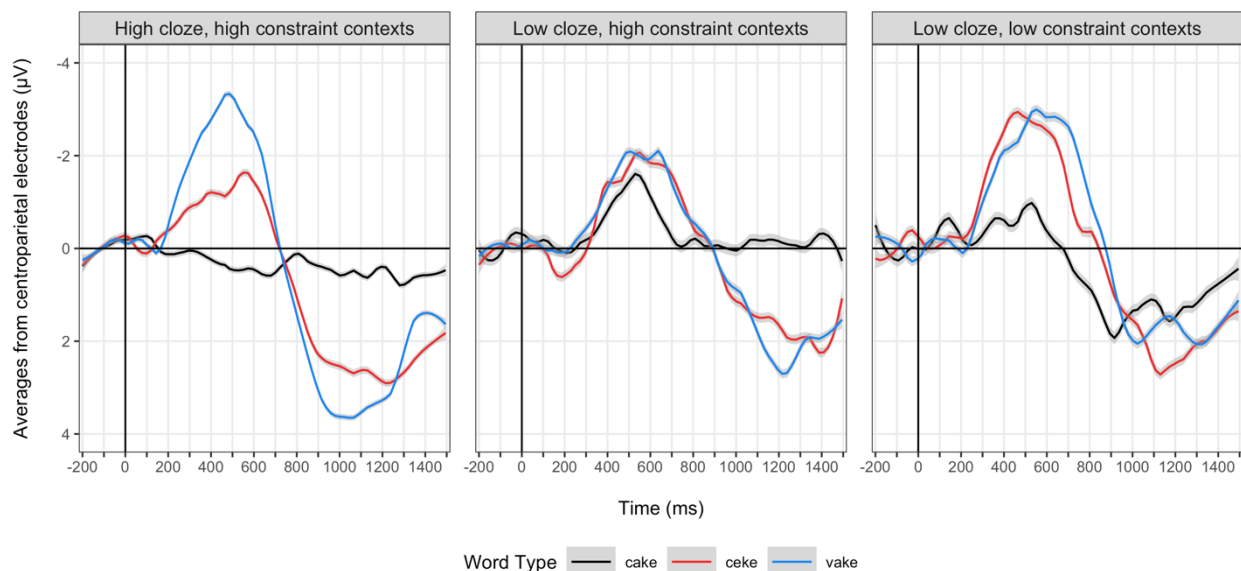


Figure 3.8: Visualization of the ERP effects across sentence constraint and target predictability. The grand average waveforms from parietal electrodes (Pz, P3, P4) are plotted in three groups, depending on the target word’s cloze probability (high or low cloze) and the overall constraint of the context (high or low constraint). The black lines represent the baseline condition (*cake*), and the red and blue lines represent the form-similar (*ceke*) and rime (*vake*) conditions respectively. These waveforms were produced in R and then smoothed using local regression (loess) smoothing techniques.

3.4.4. What are the open questions and what should we (collectively) do next?

The present study demonstrates that form-based prediction is a widespread phenomenon in language comprehension, occurring in both tightly controlled experiments, as well as in more variable naturalistic contexts. If prediction is as ubiquitous as our data suggest, a number of unanswered questions take on new importance: How is prediction carried out in linguistically diverse populations such as signers and bilinguals? How is prediction affected by cognitive variability (autism, ADHD)? How is it affected by aging? What is the relationship between variation in predictive abilities and variation in other measures of linguistic ability? How do we represent a form-based prediction while processing the form of the current (perceived) word? This latter point seems like a particularly tricky problem if we engage in long-distance predictions, as this could well involve the simultaneous prediction of a number of different lexical items.

One unanswered question, however, seems particularly urgent: How and when does form-based prediction develop? Most of the prior work on form-based prediction has recruited highly competent language users (e.g. adults with high levels of education and literacy). As we noted above, form-based prediction requires the rapid coordination of several processing steps: Comprehenders must leverage contextual information to make inferences at the highest conceptual levels and then send information back down to pre-activate lower-level representations—all before the critical word is produced and perceived. Thus, effective prediction should require considerable knowledge about one's language (and the world, more generally), as well as a rapid and efficient processing system. Given that young children are both less knowledgeable and slower at basic cognitive tasks (Kail, 1991; Kail & Salthouse, 1994; Kail & Ferrer, 2007), one might expect that they would be less apt to make form-based predictions. This deficit could resolve gradually, as processing speed increases and children fill in the gaps in their knowledge. Or there could be a

sudden shift in the system, as children adopt different processing strategies—either as a side effect of their effectiveness (e.g. ignoring predictions until they have been found to be accurate) or based on their experiences with literacy (Huettig & Pickering, 2019; Mani & Huettig, 2012).

To date, most of the research on predictive processing in children has focused on semantic prediction using the visual world paradigm. For example, Mani and Huettig (2012) presented German-speaking two-year-olds with visual displays that had two objects (e.g. a cake and a bird). On each trial, the children would hear a sentence that either contained a neutral verb or a highly constraining verb. For example, “The boy eats (sees) the big cake” (German translation, “Der Junge ißt (sieht) den großen Kuchen”). The authors found that the children made predictive eye-movements to the critical object (the cake)—but only after hearing the highly constraining verb (eats). These and other similar findings suggest that young children can use contextual cues to generate predictions (see Borovsky et al., 2012; Kidd et al., 2011; Lew-Williams & Fernald, 2007).

These findings, however, have two general limitations: First, it is unclear from the dependent measure in these studies whether children are making predictions about what is going to be said next or simply making inferences about the event under discussion. Prior work has shown that adults and children will look towards objects that are contextually relevant—even when they presumably know that this object is unlikely to be explicitly mentioned. For example, when given *wh*-questions like “What did the man eat?” people will look longer at edible objects in the scene (Atkinson et al., 2018; Golinkoff et al., 2013; Goodwin et al., 2012; Jyotishi et al., 2017; Seidl et al., 2003; Sussman & Sedivy, 2003; Yuan et al., 2011). Second, even if predictive eye-movements are generated by a linguistic prediction, it is unclear whether the prediction is at the level of meaning and/or form, i.e., are two-year-old children activating the concept of cake, the lexical item cake, or the phonological form of that lexical item?

To the best of our knowledge, there is only one study that directly explores form-based prediction in young children. Gambi et al. (2018) conducted a visual world eye-tracking study with English-speaking adults and children (aged 2–5 years). On the critical trials, the visual display consisted of two objects, each one positioned on a different side of the screen. One object would typically be labeled by a word beginning with a vowel (e.g. an ice cream cone). The other object would typically be labeled by a word beginning with a consonant (e.g. a soccer ball). In the test sentence, the two pictures were labeled with an indefinite determiner that provided predictive information about the identity of the upcoming noun (e.g. “Can you see *an*...ice cream?”). Participants’ performance on these form-based prediction trials were contrasted with trials in which different numbers of objects appeared on each side of the display and a number word was used in the test sentence (e.g. “Can you see *two*...ice creams?”). To provide more time for making predictions, the authors inserted a pause after the determiner such that the target word was produced roughly 1200 ms later.

In the number trials, results indicated that all age groups were able to shift their gaze to the correct referent after hearing the number word. Gambi et al. (2018) interpreted this finding as semantic prediction—although, it could also be interpreted as a product of incremental semantic analysis (i.e. participants shift to looking at sets of objects after hearing *two*, just as they might shift to look at green objects after hearing *green*). In the form-based prediction trials, the youngest children failed to shift their gaze on the basis of the indefinite article, suggesting that they were unable to use the phonological cues from the determiner to predict the phonological form of the upcoming noun and then correctly infer the referent. Three- to five-year-old children showed fragile effects of form-based prediction across their analyses, which led the authors to conclude that form-based prediction was only reliably observed in adults.

Based on this finding, one might conclude that, even in the most supportive of contexts, form-based prediction is absent until at least five years of age (see Pickering & Gambi, 2018). We suspect, however, that this conclusion is premature, and we see two reasons why additional research is needed: First, the phenomenon at the heart of Gambi et al. (2018) appears to be a rather weak one that is atypical of form-based prediction more broadly. In their study, words were only predictable because of arbitrary phonological rules that allowed participants to predict the onset of an upcoming word and then infer the correct referent. In contrast, when a word was predictable in the present study, it was because the content of the discourse constrained the possibilities of which words could or should come next, which in turn constrained expectations for possible word forms. Thus, form-based prediction in our study was based on a top-down flow of information, whereas prediction in the Gambi study relied on the form of one word constraining the form of the next.

We suspect that this latter pathway is rarely useful in the wild: phonological constraints are often highly local cues (e.g. the *a/an* distinction requires attention to immediately adjacent phonemes). Thus, there is little time to use these constraints to generate predictions in real time. Furthermore, predictions that are based solely on phonological information would generally be quite coarse, as there are typically just two or three alternative forms of the predictive cue (resulting in thousands of possible lexical candidates). Gambi et al. (2018) carefully designed their study to overcome these real-world impediments by reducing the discourse context to just two objects and using artificially long prosodic breaks to afford more time for prediction. It is remarkable that adults in this study were able to flexibly adapt to these circumstances, but it is not surprising that children were less flexible. While the *a/an* constraint is probably the best known test case for form-based prediction, the findings have been variable and the effects, if they exist, appear to be quite weak (for additional context, see DeLong et al., 2005, 2017; Ito, et al., 2017a, 2017c). We suspect

that these inconsistent findings have less to do with the fragility of form-based prediction, and more to do with the minimal incremental value of this particular predictive cue.

Second, the wider literature on children's language processing strongly suggests that, similar to adults, children have robust predictive abilities. While none of these studies provide direct evidence for form-based prediction, each of the relevant findings suggest that lexical prediction is common in children, and that it is not radically different from lexical prediction in adults. Thus, if lexical prediction in adulthood involves form-based prediction, then these findings suggest that children engage in form-based prediction as well.

For example, like adults, children show robust N400 effects during comprehension, which suggests that some degree of lexicosemantic pre-activation emerges in adolescence (e.g. Friedrich & Friederici, 2006; Henderson et al., 2011; Juottonen et al., 1996). Prior work using the *Storytime* paradigm has shown that N400 effects in adults and children (5–10 years old) are best predicted by the cloze probability of a word (Levari & Snedeker, 2018). This finding suggests that both age groups are sensitive to the predictability of a word given its context. Finally, developmental researchers have long used versions of the cloze task to assess children's comprehension and morphological productivity (Berko, 1958; Brown & Berko, 1960; Carroll, 1971; Shanahan et al., 1982; Skarakis-Doyle & Dempsey, 2008). In these studies, children must correctly predict a word, including its precise phonological form, in order to accurately complete the sentence.

These studies, however, clearly stop short of demonstrating that children often (or ever) use top-down contextual information to predict upcoming words in real time. For this reason, we are currently conducting a study parallel to the present experiment with young children (5–7 years old) to see if they also show reduced N400 effects to form-similar nonwords in highly predictable

contexts. The findings of this study will place hard constraints on our theory of how this central skill develops.

3.5. Conclusion

The present study demonstrates that form-based prediction is a widespread phenomenon in language comprehension, occurring in both tightly controlled experiments, as well as in more variable naturalistic contexts. To do this, we relied on a novel naturalistic listening task in which participants simply listened to a narrative with experimental manipulations spliced into it. Our findings suggest that adults were actively predicting the phonological form of upcoming words, as evidenced by reduced N400 responses to form-similar violations in highly predictable contexts. In addition, we observed robust late posterior positivities to all violations in our task, suggesting that adults engage in deep comprehension while listening to naturally produced stories. The success of this paradigm in eliciting robust prediction opens the door for testing predictive abilities in a wide range of populations and age groups. With these findings, we conclude that form-based prediction is a common aspect of comprehension in the wild, and future research must begin to characterize how this phenomenon emerges across development and across a wider range of contexts.

Chapter 4

[Paper 3]

Studying linguistic prediction in American Sign Language narratives

Anthony Yacovone, Jessica Moore, Annemarie Kocab,

Kathryn Davidson & Jesse Snedeker

4.1. Introduction

Prediction research has long been dominated by studies of written and spoken languages. These two language systems are tightly intertwined—after all, written text was largely created to record what people were saying aloud (Bloomfield, 1984, Chapter 2; but see, Olson, 1993). One consequence of this relationship is that written and spoken languages make use of the same lexical items, syntactic structures, and semantic representations (e.g. Caramazza, 1997; Cleland & Pickering, 2006). As a result, the experiences that we have in one modality can influence our comprehension and production in the other modality (Braze et al., 2007; Huettig & Pickering, 2019; Monzalvo & Dehaene-Lambertz, 2013; Storch & Whitehurst, 2002).

Recently, Huettig and Pickering (2019) proposed that learning to read enhances the ability to predict upcoming words in written *and* spoken modalities. Specifically, they argue that reading provides optimal conditions for the development of prediction (e.g. reading can unfold in a stable,

self-paced manner, and written text is largely invariant in form across contexts). Thus, as reading proficiency improves over time and with ample experience, people begin to predict which written words are likely to come next in a sentence *before* they encounter them on the page. Once these predictive skills are well-established in reading, they can also be used to predict upcoming spoken words when listening to the same language.

On this account, comprehenders' abilities to predict upcoming words (and their form features) is not a central characteristic of the language system but rather a cultural artifact that emerges under particular circumstances. Specifically, the transfer of predictive skills from the written to the spoken modality occurs due to their shared systems of representation. Preliminary support for this claim comes from psycholinguistic experiments showing that higher reading proficiency in a language is associated with making form-based predictions about upcoming spoken words (and phonological forms) in the same language (Favier et al., 2021; Mani & Huettig, 2012, 2014; Mishra et al., 2012; Ng et al., 2018).

This proposal sparks an interesting question—what happens when a language does not have a primary written form? For example, visual-manual (or sign) languages like American Sign Language (ASL) are fully realized, autonomous linguistic systems with the same semantic, syntactic, and lexical complexity as spoken languages (Bellugi & Fischer, 1972; Hickok et al., 2001; Stokoe, 1980; for discussion, see Emmorey & Lane, 2013)—however, unlike most spoken languages, they do not have a commonly used written form (Grushkin, 2017). As a result, deaf signers who use a visual-manual language as their primary mode of communication often learn to read and write in another language like English (Hoffmeister & Caldwell-Harris, 2014; Musselman, 2000). The present study explores whether such signers, who do not engage directly

with their primary language through written text, still demonstrate form-based prediction when comprehending a narrative in the visual-manual modality.

In the remainder of this Introduction, we first address the process of learning to read in hearing and deaf signing populations, motivating the question of whether predictive skills from reading should transfer to the deaf signers' processing of sign languages. Next, we discuss how researchers typically demonstrate prediction in written and spoken languages, drawing on research that uses electroencephalography (EEG). Then, we review the handful of EEG studies in ASL, noting that these studies do not provide definitive evidence for prediction in the visual-manual modality. Finally, we end by describing the present study and how it is designed to assess prediction in ASL.

4.1.1. How do hearing and deaf signing populations learn to read?

The process of learning to read is fundamentally different for hearing children and deaf signing children with limited to no hearing ability (Goldin-Meadow & Mayberry, 2001; Lederberg et al., 2012). For hearing children, learning to read largely consists of mapping new written symbols onto their existing phonological and lexical representations. For example, in an alphabetic language like English, hearing readers must learn how graphemes map onto the phonemes of their spoken language (Ehri, 1987; Gibson et al., 1963). After learning these mappings, they can successfully decode written words into the spoken words that they already know. As hearing children become more proficient readers, they begin to build up direct associations between written word forms and their lexical concepts, allowing them to access a word's meaning without necessarily sounding it out (Rayner, Foorman, et al., 2001; Seidenberg, 2017).

In contrast, deaf children do not benefit from a direct correspondence between written text and a primary language that is fully accessible. As a result, the process of learning to read for deaf signers involves simultaneously learning the written word forms, the lexicalized concepts that they encode, and the syntactic system of a new language (see Goldin-Meadow & Mayberry, 2002). For example, a deaf child learning to sign in ASL *and* read in English will need to acquire the lexicalization patterns and syntactic structures unique to both languages (and modalities). The mapping between sign and written languages is not as direct as the mapping between spoken and written languages—however, there is some evidence that deaf readers, like hearing readers, make connections between written words and the corresponding lexical representations in their primary language.

For example, Treiman and Hirsh-Pasek (1983) demonstrated that native English speakers struggled more when silently reading sentences with similar sounding words (e.g. “She chose three shows to see at the theater”) relative to sentences with dissimilar sounding words (e.g. “She picked two movies to see with her friend”). In contrast, native deaf ASL signers did not show difficulties with these sound-similar sentences, but they did experience difficulties when the sentences contained English words that corresponded to signs with similar articulatory features (e.g. “I ate the apples at home yesterday” in which the ASL signs for EAT, APPLE, HOME, and YESTERDAY all involve closed handshapes articulated near the face).

At this point, it may be tempting to assume that acquiring a sign language before learning to read may present challenges, as these language systems do not readily reinforce one another in the ways that spoken and written languages do. This assumption, however, would be inconsistent with the body of evidence showing that sign languages do not interfere with the acquisition of a written language. In fact, numerous studies have shown that being *more* proficient in a sign

language is associated with stronger reading abilities (Andrew et al., 2014; Chamberlain et al., 1999; Strong & Prinz, 1997; see Goldin-Meadow & Mayberry, 2002). Unsurprisingly, this pattern also holds true for hearing populations: the more proficient the speaker, the stronger the reader in that same language (see Nation & Snowling, 2004).

While the process of learning to read in hearing and deaf populations is quite distinct—at the end of these processes, both populations seem to make clear connections between their primary languages and the written language that they are trying to acquire. As a consequence, the processing skills that hearing and deaf children slowly built and refined while learning their first language are likely to also be relevant for comprehending written languages. Both spoken and sign languages afford knowledge about the kinds of events that are likely in the world, the particular details that are relevant to mention, and the range of linguistic devices available to convey those details at many levels of abstraction (e.g. syntax, morphology, the existence of sublexical structure). This knowledge surely helps comprehenders when trying to decode written words on a page, allowing them to make top-down expectations about (at least) the gist of what is likely to come next.

In light of this, the open questions in the present study are whether the predictive skills developed from experience with written text (i.e. the prediction of upcoming word forms) transfer to the visual-manual modality. According to Huettig and Pickering (2019), the primary reason that predictive skills transfer from written to spoken language is that these language systems act on shared representations. Thus, the process of reading develops strong associations between particular words and broader linguistic contexts, and those associations are also directly present in that language's spoken analog. On this account, predictive skills should transfer from written to spoken languages but not necessarily to sign languages.

As a starting point for addressing these open questions, the present study seeks to understand predictive processing in sign comprehension. In particular, we are interested in whether signers make lexical predictions and anticipate the form of upcoming signs in an ASL narrative. In the sections below, we summarize the prior literature using EEG to demonstrate prediction in written and spoken languages. Then, we describe how EEG has been used in sign language research, noting that there is very little evidence to date that prediction of any kind occurs in the visual-manual modality.

4.1.2. Predictive processing in written and spoken languages—evidence from EEG

Psycholinguists have long used EEG and event-related potentials (ERPs) to study the nature of prediction during comprehension (Federmeier, 2007; Kuperberg & Jaeger, 2016; Kutas et al., 2006; Kutas & Van Petten, 1994; Swaab et al., 2012). To do this, researchers typically record comprehenders' neural responses to words presented in various ways: in prime-target pairs (e.g. salt–PEPPER), in sentential contexts, or even in larger discourse narratives (Kutas & Hillyard, 1984; Lau et al., 2013; van Berkum, 2004; Yacovone et al., 2021). The responses to each target word are then averaged together, revealing systematic patterns of neural activity known as ERP components (Kappenman & Luck, 2011; Luck, 2014). The most common and best understood ERP component in prediction studies is the N400. This component is a negative-going deflection in an ERP waveform that typically peaks over centroparietal regions between 300–500 ms after word onset (Kutas & Federmeier, 2011).

Contemporary theorists generally interpret the N400 as an index of the relative difficulty in accessing the lexicosemantic features associated with a word (Federmeier, 2007, 2022; Kuperberg, 2007; Kuperberg et al., 2020; Kutas et al., 2014). There are at least three pieces of

evidence to support this interpretation: First, when a word is presented in a plausible context, it evokes smaller N400 responses than when the same word is presented in isolation (Kutas, 1993). This pattern is largely attributed to comprehenders' abilities to use contextual information to pre-activate (or predict) representations associated with upcoming words, leading to easier lexicosemantic processing. Second, N400 responses to words at the beginning of sentences are often larger than those to words at the end, as the unfolding context generally makes each new word more predictable than the last (Payne et al., 2015; Van Petten & Kutas, 1990, 1991). Finally, the size of the N400 has an inverse correlation with behavioral measures of predictability (e.g. cloze probabilities) such that the greater a word's predictability, the smaller the N400 response (Kutas et al., 1984; Kutas & Hillyard, 1984; for a computational review of N400, see Nour Eddine et al., 2022). Taken together, these findings demonstrate that the N400 is sensitive to a word's predictability given its context—but how do we know which specific aspects of a word are predicted during comprehension?

To understand the nature of comprehenders' predictions, researchers often compare the N400 responses to violations that share or do not share certain features with an expected word. For example, Federmeier and Kutas (1999) had participants read English sentences with predictable endings like “Chris moped around all morning when he discovered there was no cream cheese. He complained that without it he couldn't eat his *bagel*.” Then, they either kept the expected word (*bagel*) in the sentence or replaced it with a word from the same semantic category (*toast*) or a word from a different one (*cake*). The authors found N400 effects for both violations; the effects, however, were reduced for the same-category word (*toast*) relative to the different-category word (*cake*). Moreover, this N400 reduction for same-category words is not observed in sentences where the ending is less predictable, e.g., “Amy was very anxious about traveling abroad for the first

time. She felt surprisingly better, however, when she actually boarded the *plane* | *helicopter* | *gondola*.” Thus, the authors concluded that readers predict the semantic features of an upcoming word when the sentence is predictable, leading to smaller N400s when those predicted semantic features appear in the input.

This kind of semantic prediction, however, is not the relevant predictive processing for assessing the aforementioned literacy hypothesis. Semantic prediction has been demonstrated with pre-literate children as young as two (Borovsky, Elman, & Fernald, 2012; Mani & Huettig, 2012), as well as low-literacy adults (e.g. Ng et al., 2018). Thus, a stronger case study for demonstrating robust, lexically specific predictions during comprehension is *form-based prediction*. More specifically, when a violation resembles the orthographic or phonological form of a highly predictable word, its N400 response is reduced. To demonstrate this, Kim and Lai (2012) created highly constraining sentences like “She measured the flour so she could bake a *cake*.” Then, they either kept the predictable final word or replaced it with a nonword that either shared orthographic features (*ceke*) or did not (*tont*). Results indicated smaller N400 responses to the form-similar nonwords (*ceke*) relative to the form-dissimilar nonwords (*tont*)—in fact, the size of the N400 responses for *ceke* and *cake* were not significantly different from one another. Numerous studies have replicated this finding using both real words and nonwords (e.g. DeLong et al., 2019, 2021; Ito et al., 2016; Laszlo & Federmeier, 2009), and critically, this pattern of findings has been directly linked to predictive processing, as form-based reductions in N400s are not found in less predictable contexts (Ito et al., 2016).

4.1.3. Using EEG to assess signers' sensitivity to form features during comprehension

Numerous studies have shown signers' sensitivity to semantic violations in sentence contexts (Capek et al., 2009; Grosvald et al., 2012; Gutierrez et al., 2012; Neville et al., 1997). But as we mentioned above, semantic prediction does not provide the strongest evidence for predictive processing, and researchers often explain these effects as reflecting bottom-up perception and integration rather than prediction (for discussion, see Cates et al., 2013). Thus, the strongest evidence for prediction in sign language would involve form-based prediction. To the best of our knowledge, there are no EEG studies with explicit demonstrations of form-based prediction in sign languages. To do this, a study would need to show patterns akin to those in prior written and spoken language studies (i.e. reduced N400s to form-similar violations in predictable but not unpredictable contexts). There are, however, various studies that use EEG to assess signers' sensitivity to manipulations of form features (i.e. the *handshape*, the *location*, and the *movement* of a sign) during comprehension. These studies are particularly relevant to the present investigation, as they demonstrate signers' awareness of form similarity between signs and their ability to detect form violations in predictable sentence contexts. Thus, we review those findings below.

A handful of EEG studies have used priming paradigms to show that N400 responses to target signs are reduced when preceded by a form-similar sign relative to an unrelated sign (Meade et al., 2018, 2022). For example, the ASL signs for SUMMER and UGLY share *handshape* and *movement* features but differ in their *location*—SUMMER is articulated across the forehead, whereas UGLY is articulated across the face. In contrast, SUMMER and MEDICINE do not share any features. Thus, SUMMER evokes smaller N400s when preceded by UGLY than when preceded by MEDICINE (see Meade et al., 2018). These priming studies do not demonstrate form-based prediction per se,

but they do show that signers are sensitive to form manipulations and experience smaller N400 effects when form features are primed by the prior bottom-up context.

To date, there is only one EEG study that provides insight into form-based prediction in ASL. Gutierrez et al. (2012) presented native deaf ASL signers with videos of sentences that contained a highly predictable sign, e.g., “ALASKA THEIR ANIMAL WORST BIG THAT **BEAR** EXTREME” (English translation, “*The biggest animal in Alaska is a massive **bear***”). The authors either kept the expected sign (BEAR) in the sentence or replaced it with one of four violations. These violations either matched or mismatched the expected sign in their semantic and form features. For example, BAT is semantically related to bear and has the same *location* feature [+semantics, +location]; MONKEY is semantically related but shares no form features [+semantics, –location]; BACKPACK is semantically unrelated but has the same *location* [–semantics, +location]; and finally, DOLLAR is a sign with no semantic or form overlap with BEAR [–semantics, –location].

Results indicated robust N400 effects to all violation conditions between 450–600 ms, consistent with prior demonstrations of semantic incongruity effects in ASL (e.g. Capek et al., 2009; Grosvald et al., 2012; Neville et al., 1997). Interestingly, the N400 effect for the [+semantics, +location] violation seemed to be reduced relative to the others. It is tempting to interpret this finding as an effect of semantic and form-based prediction; however, the one-feature overlap and the no overlap violations evoked similarly sized N400s. Thus, these findings are most consistent with the N400 reflecting the bottom-up processing of these signs and the relative difficulty of integrating them into the context (for discussion of these findings, see also Cates et al., 2013).

In sum, the prior literature on sign language processing has not clearly demonstrated semantic nor form-based prediction. To do this, a study would need to contrast the processing of

violations in both predictable and unpredictable conditions. In the section below, we outline the design of the present study and discuss how it aims to assess prediction in ASL.

4.1.4. *The present study*

The long-term goal of this project is to better understand how prediction unfolds in sign languages. Our immediate goal, however, is to determine whether signers of ASL make lexically specific, form-based predictions in ordinary contexts. To do this, we used a naturalistic comprehension task known as the *Storytime* paradigm (see Yacovone et al., 2021). In this paradigm, people simply comprehend coherent, naturally produced stories for the sole purpose of understanding the narrative. We embed our experimental manipulations within these stories by splicing in carefully recorded audio for spoken languages or videos for sign languages. There are many advantages to the *Storytime* paradigm: it makes traditional psycholinguistic experiments accessible to a wider range of populations and age groups; it promotes attention to contextual information and arguably enhances predictive processing; and it provides evidence that traditional findings generalize to more naturalistic settings.

In the present study, native deaf signers simply watched someone sign an hour-long story in ASL. Throughout the story, the narrator occasionally produced non-signs with varying degrees of resemblance to the original sign from the story. Specifically, we created two types of non-signs: *handshape-changes* and *all-changes*. In the handshape-change condition, the non-sign matched the original sign from the story in location and movement but mismatched in handshape. For example, if the original sign had a handshape with an extended index finger, we changed the extended finger when producing the non-sign but kept the location and movement constant (see Figure 4.1 below). In the all-change condition, the non-sign differed from the original sign in

handshape, location, and movement. Importantly, these target signs appeared in both predictable and unpredictable contexts, allowing us to attribute our effects to predictive processing (i.e. predictive effects can only emerge when signers can anticipate upcoming signs and their features).



Figure 4.1: *Examples of the three sign manipulations in our story context.* In these panels, the narrator is producing the target sign TIME (left panel) and the *handshape-change* and *all-change* variations (middle and right panels, respectively). The handshape-change (or HS-change) non-sign shares location and movement with TIME but uses the “I” handshape instead of the “X” handshape. The all-change non-sign shares no form features with TIME, and it was created by producing FALL with a “7” handshape.

Given our design, we had three main predictions: First, if predictive processing is occurring, we should expect to see baseline differences in the N400 responses to non-manipulated, baseline signs depending on their predictability. Specifically, we should see smaller N400 responses to signs that are predictable relative to signs that are unpredictable (see Kutas et al., 1984; Kutas & Hillyard, 1984). Second, we should expect the non-signs with more phonological overlap, i.e., handshape-changes, to elicit smaller N400 responses compared to all-changes, but only in predictable contexts. Third, we also expect to find another ERP component known as the P600 to all our non-signs. These late posterior positivities are evoked by nonwords in reading studies (Kim & Lai, 2012; Laszlo & Federmeier, 2009), auditory studies (Yacovone et al., Chapter 3), and even sign language studies (Grosvald et al., 2012). P600s are often larger in high constraint contexts and in contexts with violations that are similar in form to the original word in the sentence

(Ito et al., 2016; Kuperberg et al., 2020; Laszlo & Federmeier, 2009; Ryskin et al., 2021). Thus, we expect to find P600s for all non-signs, and perhaps evidence that they are greater for handshape-change non-signs, which have a closer resemblance to the original sign from the story.

4.2. Method

4.2.1. Participants

We recruited 24 native deaf ASL signers; however, we excluded three participants from the final analyses following our pre-registered criteria on OSF (<https://osf.io/3w2yq>). Seventeen participants in the final sample were early signers (exposed to ASL before age 4) and four were late signers (*Mean age* = 31 years, *Range* = 19–52 years). Two participants were left-handed. Participants varied in their highest level of education: four participants had obtained high school or vocational degrees; nine held bachelor's degrees; and eight held a master's degree or higher. Seven participants reported current or prior experience using hearing aids and/or cochlear implants. All participants provided consent and received cash payment for their time. Our final sample size will be 30 signers, as determined by a priori power analyses. Thus, this is a preliminary analysis, and our findings may change once the full sample is obtained.

4.2.2. Stimuli

The present study used a naturalistic comprehension task called the *Storytime* paradigm (see Yacovone et al., 2021). We presented a coherent story in ASL in which the narrator occasionally produced a non-sign in ASL (see Figure 4.1). The story was an ASL translation of an abridged English children's story called *Mystery of the Turtle Snatcher* by Kyla Steinkraus. This story stimulus has many useful features: it uses simple language; it has predictable plotlines that

are not widely known to the public; and it is accessible to a wider age range, paving the way for future developmental research. Moreover, this story has been used in prior studies on form-based prediction with hearing adults and children (Yacovone et al., Chapter 3). We selected 180 target signs from the translated story to create a 2×3 manipulation of sign type (*baseline, handshape-change, all-change*) and predictability (*higher cloze, lower cloze*). We also made sure that no lexical items were used more than three times in the story to ensure that no manipulations were repeated.

Sign type manipulations. We created our handshape-change non-signs by changing the dominant handshape of our baseline target signs. For example, if the citation form of the target sign included a handshape with an extended index finger, we changed the extended finger when producing the non-sign (see the change from the “X” handshape to the “I” handshape in Figure 4.1). All-change non-signs were created by finding an existing sign that differed from the intended sign in location and movement. Then, we changed its dominant handshape to make a non-sign, ensuring that this new sign obeyed the phonological constraints of ASL but was not an actual sign.

Predictability manipulations. To ascertain the predictability of our target signs, we conducted two cloze tasks with native deaf ASL signers via Zoom. In the first cloze task, participants watched the story with no lexical violations. The video paused before a target sign, and the signers were instructed to provide the next sign in the sentence. The second cloze task was identical to the first, except the story contained the non-sign manipulations described above. We used this second cloze task to characterize the predictability of our target signs in the exact video stimuli used in the EEG study (i.e. the story with lexical violations). For our analyses, we averaged the cloze probabilities from these two tasks. The median cloze probability was 32%, which we used as the boundary for separating higher and lower cloze target signs. Higher cloze targets had

an average cloze probability of 58.5% ($SD = 16.6\%$, $Range = 33\text{--}100\%$). Lower cloze targets had an average cloze probability of 12.6% ($SD = 10.9\%$, $Range = 0\text{--}32\%$).

Sign frequency measures. We also characterized the frequency of our target signs using the subjective frequency ratings from ASL-LEX 2.0 (Sehyr et al., 2021). These frequency values were obtained from native deaf ASL signers, who used a 7-point scale to indicate how often each sign appeared in everyday conversation (1 = very infrequently, 7 = very frequently). Higher cloze targets had an average frequency of 5.41 ($SD = 1.04$) and lower cloze targets had an average frequency of 4.98 ($SD = 1.05$). Twelve of our target signs, or at least the dialectal variations that our narrator used, did not appear in ASL-LEX 2.0. Thus, we removed these items from our final analyses, leaving 159 out of 180 items (some target signs were used more than once).

4.2.3. Procedure

Participants watched an hour-long video of a native deaf person signing our story stimulus. Video recordings were presented using PsychoPy (Peirce et al., 2019). Participants sat roughly 100 cm from a TV monitor, and they were encouraged to minimize movement and to keep their faces relaxed. Participants were given a break in the middle of the story to rest. We asked comprehension questions during the break and at the end of the story. Participants were fully debriefed at the end of the study.

4.2.3.1. EEG recording and pre-processing procedure

Participants wore an electrode cap (actiCAP SnapCap) containing 31 active Ag/AgCl electrodes that were connected to the EEG equipment, Brainvision's actiCHamp Standard 64 System. We attached two mastoid electrodes (TP9 and TP10) directly behind participants' ears.

The EEG data were recorded at a sampling rate of 500 Hz using Brainvision's Recorder (BrainVision Recorder, Version 1.23.0001, Brain Products GmbH, Gilching, Germany). On average, electrode impedances were kept below 20 K Ω . During recording, the ground electrode was FPz and the reference electrode was FP1.

4.2.4. Analysis plan

4.2.4.1. Determining our spatial and temporal regions of interest

To determine our spatial regions of interest (ROIs), we relied on information from the prior literature and our pilot data. For N400 effects, we selected a centroparietal ROI with eight electrodes: *Cz, C3, C4, CP1, CP2, Pz, P3, P4*. For P600 effects, we selected a parietal ROI with three electrodes: *Pz, P3, P4*. To determine our temporal ROIs, we used a collapsed localizer technique in which all the conditions being compared are collapsed into one grand average waveform, which is then used to determine the ROI (for discussion, see Luck & Gaspelin, 2017). The grand average can be used to determine ROIs in various ways: simple visual inspection, cluster-mass permutation tests on grand averages, or centering ROIs on peak amplitudes (Luck, 2014; Luck & Gaspelin, 2017). We adopted this last approach and centered a 200 ms time window on the most negative peak (for N400 effects) and the most positive peak (for P600 effects) in the grand average waveforms.

Following our pre-registration, we defined temporal ROIs for each cloze condition separately. We took this step because prior work has shown that predictability can influence the timing of ERP effects (Kutas & Federmeier, 2011; Swaab et al., 2012). For example, N400 effects can emerge earlier in predictable contexts relative to unpredictable ones (Brothers et al., 2015).

4.2.4.2. Linear mixed effects model specifications

All of the linear mixed effects models were implemented using the *lme4* and *afex* packages in the R statistical computing environment (Bates et al., 2014; R Core Team, 2022; Singmann et al., 2023). All pairwise comparisons were implemented using the *emmeans* package (Lenth et al., 2021). We initially fit our models with the maximal random effects structures justified by our data. If we encountered convergence issues, we simplified the random effects structure until the models properly converged (Baayen et al., 2008; Barr, 2021). Most issues were resolved by constraining the covariance parameters for the random effects to zero (i.e. removing the correlations between them). But, if this step did not resolve the issues, we began to incrementally drop random slopes (while trying to preserve the slopes for the highest order effects of interest) until the models converged.

All effects with an absolute value of t greater than 2 are considered significant (Gelman & Hill, 2006). We follow this convention due to the uncertainty in the field about how to best calculate the appropriate degrees of freedom in linear mixed effects models (Baayen et al., 2008). For the sake of completeness, however, we also report the p -values as calculated by the *lmerTest* package (Kuznetsova et al., 2017). Note, in all of our models, both methods of evaluating significance arrived at the same conclusions.

4.3. Results

For the present study, we pre-registered a set of primary and secondary hypotheses related to the nature of predictive processing in ASL. These hypotheses are similar to those explored in Chapter 3 of this dissertation. Below, we report the findings from a set of linear mixed effects

models that aimed to test those hypotheses.¹¹ First, if signers in our task are predicting upcoming material, we should see smaller N400 responses to baseline signs in more predictable relative to less predictable contexts. This pattern would be consistent with prior findings that the N400 is reduced as a function of the target word's predictability (Kutas et al., 1984; Kutas & Hillyard, 1984). Second, if signers are engaging in form-based prediction, we should see reduced N400s to handshape-change non-signs relative to all-change non-signs—but only in higher cloze contexts. Third, if signers are sensitive to our non-sign manipulations, we should see P600s in response to all violations. In the prior literature, P600s seem to be sensitive to the constraint of a sentence (e.g. Gunter et al., 2000; van de Meerendonk et al., 2010) and the similarity in form between a violation and the intended word (e.g. Ito et al., 2016; Kim & Lai, 2012; Laszlo & Federmeier, 2009; Ryskin et al., 2021; Vissers et al., 2006). Thus, we should expect larger P600s in more constraining contexts, as well as larger P600s to handshape-change relative to all-change non-signs. We investigate each of these predictions below.

4.3.1. Visualizing the grand average waveforms and topographic maps

Figure 4.2 presents the grand average waveforms from centroparietal electrodes (*Cz*, *C3*, *C4*, *CP1*, *CP2*, *Pz*, *P3*, *P4*) for all experimental conditions. To calculate these waveforms, we

¹¹ In our pre-registration, as discussed in Chapter 3, we outlined a series of analyses that modeled by-participant and by-item ERP averages for each electrode in our region(s) of interest. To do this, we planned to implement mixed effects models with random effects for participants or items, as well as random effects for electrode site. During an external review of this work, however, some questions were raised about whether this approach could properly account for the correlations between individual electrodes. To address these questions, we had two options: First, we could collapse across electrodes, leaving only one average ERP value per condition for each participant or each item. Then, with these averages, we could implement simpler regressions. Alternatively, rather than initially collapsing across participants (for the by-item analyses) or items (for the by-participant analyses), we could just collapse across electrodes and implement linear mixed effects models with random effects for *both* participants and items. This latter option is more commonly used in psycholinguistic research, and so we report the results from these trial-level analyses below. Importantly, however, all statistical approaches resulted in the same pattern of findings and theoretical conclusions.

collapsed across individual items to create six waveforms for each participant. We then further collapsed these participant-level averages to create the grand averages. Figure 4.3 presents the topographic maps for the isolated effects of our handshape-change and all-change manipulations.

Visual inspection of these figures reveals robust N400 effects for the two non-sign conditions in higher but not lower cloze contexts. Moreover, these N400 effects are similarly sized, suggesting that our signers did not engage in form-based prediction. There is, however, evidence for our remaining two predictions: First, there are robust P600 effects for all non-signs, suggesting that signers detected our violations within the larger story context. These P600s appear larger to handshape-change non-signs than to all-change non-signs, as expected. We do not, however, see any evidence for larger P600s in our higher cloze contexts overall. Second, there is evidence that the N400 responses to the baseline signs are smaller in more predictable contexts. This pattern is consistent with the prior work from written and spoken language studies, suggesting that signers are engaging in predictive processing at some level of representation during naturalistic comprehension.

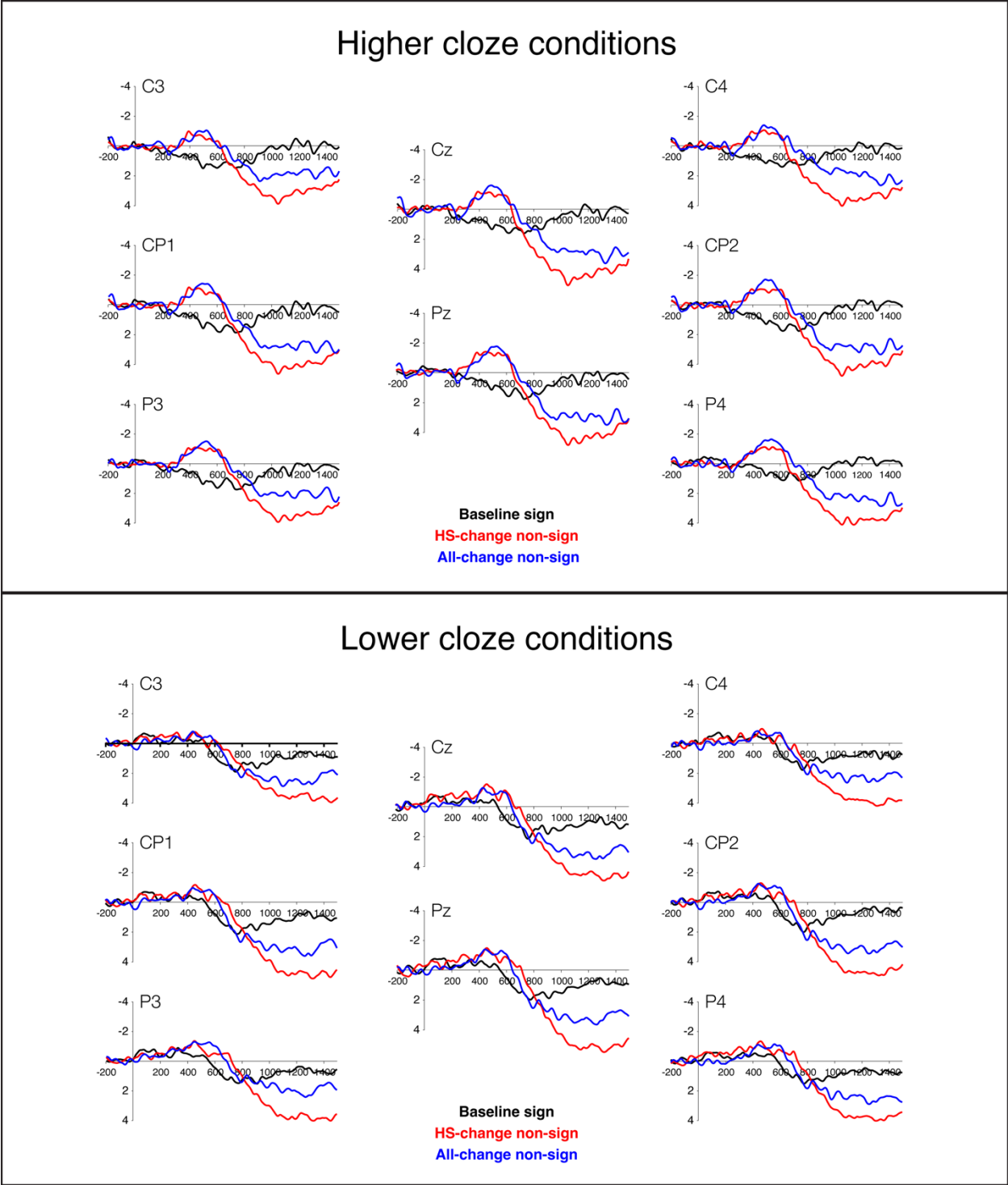


Figure 4.2: Grand average waveforms for all sign types by predictability. The grand averages (μV) for centroparietal electrodes are presented for higher cloze (top panel) and lower cloze (bottom panel) conditions. The black lines represent the *baseline signs*, while the red and blue lines represent the *handshape-change* and *all-change non-signs* respectively. All waveforms were subjected to an additional low-pass filter of 15Hz for plotting purposes.

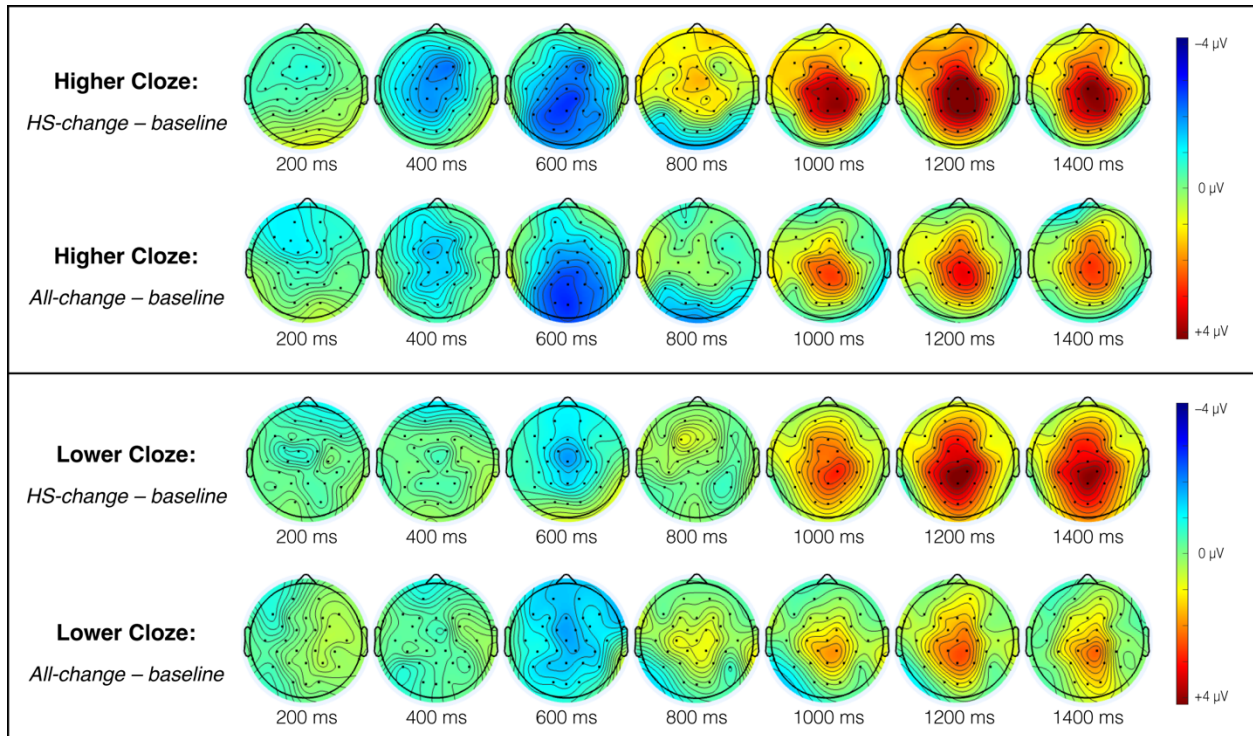


Figure 4.3. Topographic maps of the ERP effects across predictability. These topographic maps depict the isolated effects of the handshape-change and all-change non-signs in both higher and lower cloze contexts. These effects are calculated by subtracting the baseline ERP activity from the activity evoked by each violation.

4.3.2. Are N400 effects reduced for handshape-change non-signs in higher cloze contexts?

If signers were making form-based predictions in our study, we should see smaller N400 responses to handshape-change relative to all-change manipulations—but only when signers were able to predict the original target sign. There is little evidence for form-based prediction in Figures 4.2 and 4.3, as the magnitudes of the N400 effects for both violations in higher cloze contexts are nearly identical. To statistically confirm this finding, we calculated mean N400 amplitudes from centroparietal electrodes between 435–635 ms and 345–545 ms for higher and lower cloze conditions respectively. These time windows were selected using the localizer technique described in Section 4.2.4.1. We then modeled these averages using a linear mixed effects model with fixed effects of sign type (*baseline, handshape-change, all-change*), predictability (*higher cloze, lower*

cloze), and their interaction. For sign type, we tested for successive differences using two contrasts: *baseline* vs. *handshape-change*, and then *handshape-change* vs. *all-change*. For predictability, we simply tested for differences between higher and lower cloze conditions (*higher cloze* = $-.5$, *lower cloze* = $.5$). All relevant pairwise comparisons were tested (and corrected for multiplicity) using the *emmeans* package (Lenth, 2021). Finally, we had random intercepts and maximal slopes for participants and items. To reach convergence, however, we needed to constrain the covariance parameters for the random effects to be zero.

Results indicated a significant main effect of sign type for the contrast between *baseline* and *handshape-change* conditions ($b = -1.37$, $SE = .43$, $t = -3.19$, $p = .003$). There was no main effect of sign type for the contrast between non-signs ($b = -0.08$, $SE = .39$, $t = -0.21$, $p = .83$), confirming that *handshape-change* and *all-change* non-signs evoked similarly sized N400s. There was also no main effect of predictability ($b = -0.33$, $SE = .39$, $t = -0.84$, $p = .41$); however, there was a significant interaction between sign type and predictability, but only for the contrast between *baseline* and *handshape-change* conditions ($b = 1.89$, $SE = .78$, $t = 2.42$, $p = .016$). Planned pairwise comparisons revealed that this interaction was driven by the presence of N400 effects in higher cloze but not lower cloze contexts: in higher cloze contexts, *baseline* and *handshape-change* conditions significantly differed ($b = 2.31$, $SE = .57$, $t = 4.03$, Tukey-adjusted $p < .001$) as did the *baseline* and *all-change* conditions ($b = 2.38$, $SE = .58$, $t = 4.08$, Tukey-adjusted $p < .001$). In lower cloze contexts, there were no significant differences in any of the pairwise analyses.

4.3.3. Are the P600 effects sensitive to the non-signs and predictability manipulations?

Given the prior literature, we were interested in three particular patterns for our P600 effects: First, we expected to find P600s to all non-signs. Second, we expected to see larger P600s

to violations that more closely resembled the form of the intended sign. Thus, the P600s to our handshape-change non-signs should have evoked larger P600s than our all-change non-signs. Third, P600s are often larger in high constraint relative to low constraint sentences—so, we might expect larger P600s overall in higher vs. lower cloze contexts.

To address each of these predictions, we conducted a parallel analysis to the one described in Section 4.3.2. In this analysis, however, we modeled average P600 amplitudes calculated from parietal electrodes and time windows of 960–1160 ms and 1225–1425 ms for higher and lower cloze conditions respectively. As in the prior analysis, these time windows were ascertained using the localizer approach described above. The linear mixed effects model for P600 effects had identical fixed and random effects structures and tested the same set of contrasts as the model for the N400 effects. Results indicated significant main effects for both sign type contrasts, confirming P600 differences between *baseline* and *handshape-change* conditions ($b = 4.15$, $SE = .70$, $t = 5.91$, $p < .001$) and between *handshape-change* and *all-change* conditions ($b = 1.80$, $SE = .52$, $t = 3.47$, $p = .002$). There was no main effect of predictability nor any significant interactions.

These results provide support for two of our three predictions: First, there are robust P600 effects for all non-signs (regardless of predictability); Second, the P600s to handshape-changes are more positive than those to all-changes, suggesting that signers are sensitive to the similarity between the encountered violation and the intended sign given the context. In contrast, there was no evidence for larger P600s in higher cloze relative to lower cloze conditions. This finding is most likely due to the fact that our higher cloze contexts were not extremely constraining, as the average cloze probability for higher cloze target signs was roughly 60%.

4.3.4. Do the baseline N400s change as a function of the sign's predictability?

Although we do not see evidence for form-based prediction, there is some evidence for predictive processing in our study. Specifically, there appears to be smaller N400 responses to non-manipulated baseline signs in higher relative to lower cloze contexts. In our pre-registration, we planned to assess how our ERP effects changed as a function of sign-level predictability. To do this, we simply used the baseline N400 mean amplitude values from the analysis above, and then implemented a mixed effects model with a single categorical fixed effect of predictability (*higher cloze* = .5, *lower cloze* = -.5). Note, the range of cloze probability values in our items was not well distributed between 0–100%, thus we decided to use a categorical predictor (rather than a continuous predictor) in this analysis. As in prior analyses, we also had a maximal random effects structure for participants and items. Results confirmed that the N400 amplitudes for higher cloze items were significantly smaller than those for lower cloze items ($b = -1.70$, $SE = .58$, $t = -2.96$, $p = .004$). Figure 4.4 shows the average waveforms for higher and lower cloze baseline signs (left panel) and the trial-level mean N400 amplitudes by predictability (right panel). This finding confirms that signers were able to pre-activate at least some features of upcoming signs in our task, clearly demonstrating for the first time that prediction occurs during sign comprehension.

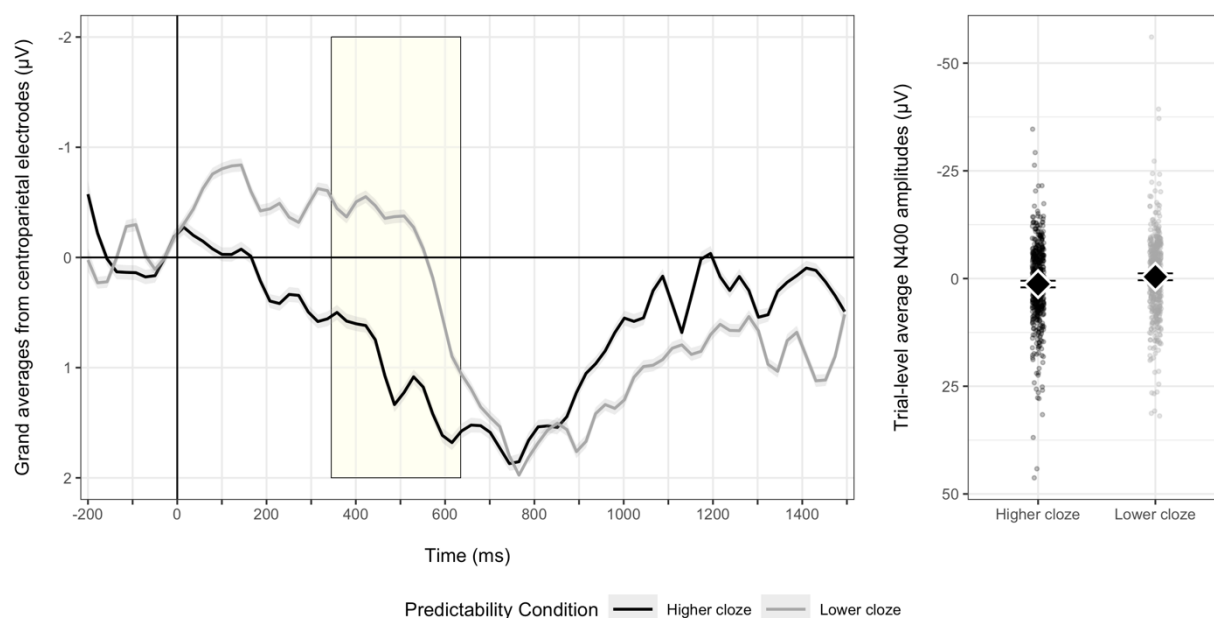


Figure 4.4: *Average waveforms and N400 mean amplitudes for all baseline conditions.* Grand average waveforms from centroparietal electrodes are plotted for higher and lower cloze baseline signs (left panel, black and gray lines respectively). These waveforms were produced in R and then smoothed using local regression (loess) smoothing techniques. The time window highlighted in yellow is the combined, localized windows for both cloze conditions from the primary N400 analysis (345–635 ms). Trial-level mean N400 amplitudes are plotted by predictability (right panel, black observations). Each dot represents the average N400 amplitude (μV) for each baseline sign (for each participant).

4.4. Discussion

It is tempting to interpret these findings as evidence that semantic prediction—but not form-based prediction—occurs during sign language comprehension. This conclusion would be quite surprising, as prior work has demonstrated that form-based prediction is readily observed in both written (e.g. DeLong et al., 2021) and spoken language comprehension (see this dissertation, Chapter 3). In the remainder of this Discussion, we outline two hypotheses that might explain why form-based prediction is less likely to appear in ASL (Section 4.4.1). There are, however, a few limitations to the present study, which make us hesitant to accept these conclusions without additional research. Thus, we end by exploring these limitations (Section 4.4.2) and discussing

ways in which future research should assess prediction during sign language comprehension (Section 4.4.3).

4.4.1. Two hypotheses about the emergence of form-based prediction in spoken and sign languages

In the Introduction, we described a recent hypothesis from Huettig and Pickering (2019) that proposed that learning to read enhances comprehenders' abilities to predict upcoming words in both written *and* spoken modalities. On this account, learning to read provides the optimal conditions for learning associations between words and broader linguistic contexts (i.e. predicting which words are likely to appear next in a sentence), as well as associations between lexical items and their specific orthographic forms in a language (i.e. predicting what the upcoming word will look like on the page). The authors argue that form-based prediction emerges first in reading because 1) reading is a self-paced process, and there are natural motivations to learn to read faster, and 2) written text is largely invariant across different contexts, making it easy to learn the specific orthographic shapes associated with a particular lexical item. Then, once these predictive skills are established in reading, they begin to emerge in spoken language, acting on the same associations that were strengthened by exposure to the language via text.

As we discussed in the Introduction, this hypothesis states that form-based prediction emerges in spoken language because of the shared systems of semantic, syntactic, and lexical representations. Thus, there is a tacit assumption that predictive skills should *not* transfer from reading to another linguistic modality (or language, more broadly) if it does not share representational systems. In the present study, we tested this hypothesis by looking at the predictive skills in deaf signers with considerable exposure to reading and writing in English. At face value, our findings are consistent with the literacy hypothesis, as we did not find strong evidence for prediction of form in our ASL comprehension task. To provide definitive support for

this prediction, however, we would need to demonstrate that deaf signers actually make form-based predictions when reading in English (see future directions in Section 4.4.2).

There is another hypothesis, however, that might account for predictive differences across modalities *without* invoking literacy—namely, sign languages may be less predictable than written and spoken languages in general, resulting in weaker expectations about which signs are likely to appear next. It is well-established that sign languages have access to a wide range of linguistic and depictive devices: *classifier predicates*, *constructed actions*, *lexical signs*, and *fingerspelling* (e.g. Beal et al., 2021; Cormier et al., 2013; Goldin-Meadow & Brentari, 2017; Keane & Brentari, 2015; Padden, 1998). Moreover, sign languages also enjoy freer word orders than most spoken languages (Cheng & Mayberry, 2019; Lillo-Martin, 1986; Napoli & Sutton-Spence, 2014). For example, although ASL follows the same default SVO order as English, signers have access to various non-canonical word orders (under certain pragmatic conditions) like sentence-final pronouns (VS), pre-verbal objects (OV), “verb sandwiches” (VOV), as well as options for null arguments (for examples, see Cheng & Mayberry, 2019).

Taken together, this inherent diversity and flexibility in structuring utterances may increase the number of possible continuations relative to a spoken language with fewer of these devices. For example, native deaf signers tend to use more constructed actions and classifiers (which can be idiosyncratic and variable) over lexical signs when retelling narratives (Beal et al., 2021). Thus, prediction of specific signs may be less rampant (and perhaps less useful) during sign language comprehension. Note, this hypothesis does not imply that sign languages are less structured in the ways that these depictive devices are integrated into the language (for examples of the structured ways that these devices are integrated into sign languages, see Ferrara & Hodge, 2018; Goldin-Meadow & Brentari, 2017)—we are simply arguing that there may be less agreement about how

a particular utterance is *likely* to continue within a group of signers relative to a group of hearing individuals, reducing our ability to observe form-based prediction in this population.

To assess this hypothesis, we evaluated the variability in the responses that deaf signers provided in our original cloze task. In this task, participants watched our story stimulus and occasionally guessed which sign would come next. All of our target signs were lexicalized nouns in ASL. If deaf signers expected the story to continue as the narrator intended, then they should also provide a noun (regardless of it being the exact noun from the story). Alternatively, if there is greater variability in how signers expect utterances to continue, we may see responses from a wider range of categories (e.g. verbs, adjectives, constructed actions, classifiers). In addition, we compared signers' responses to a set of different responses from a parallel cloze task conducted in spoken English with hearing adults and children. Specifically, Waite et al. (2023) had hearing adults and five- and six-year-old children watch a cartoon narrative of the same children's story from the present study, and guess a subset of target nouns. The similarity between these two tasks allows us to explore whether there is greater variability in the type of continuations that deaf and hearing individuals consider during comprehension of narratives.

We first selected all incorrect responses across our three populations (i.e. instances when the participants did not correctly guess the intended word). Then, to allow for cleaner comparisons across populations, we grouped participants' incorrect responses into four syntactic categories: Nouns, Adjectives, Verbs, and Other. The Other category contained responses like adverbs, prepositions, determiners, number words, etc. For signers, this category also contained classifier predicates, constructed actions, and fingerspelling. As expected, we found greater variability in the type of responses provided by deaf adults relative to hearing adults and children. Specifically, the proportion of nouns given by deaf adults was lower than the other two groups. Signers appeared

to entertain a greater proportion of verbs (20%), adjectives (10%), and other continuations (e.g. *fingerspelling* = 5%, *classifiers* = 4%). In contrast, hearing adults and children rarely provided verbs (adults = 3%, children = 6%), adjectives (adults = 4%; children = 5%), or other continuations from non-noun syntactic categories (adults = 6%; children = 8%).

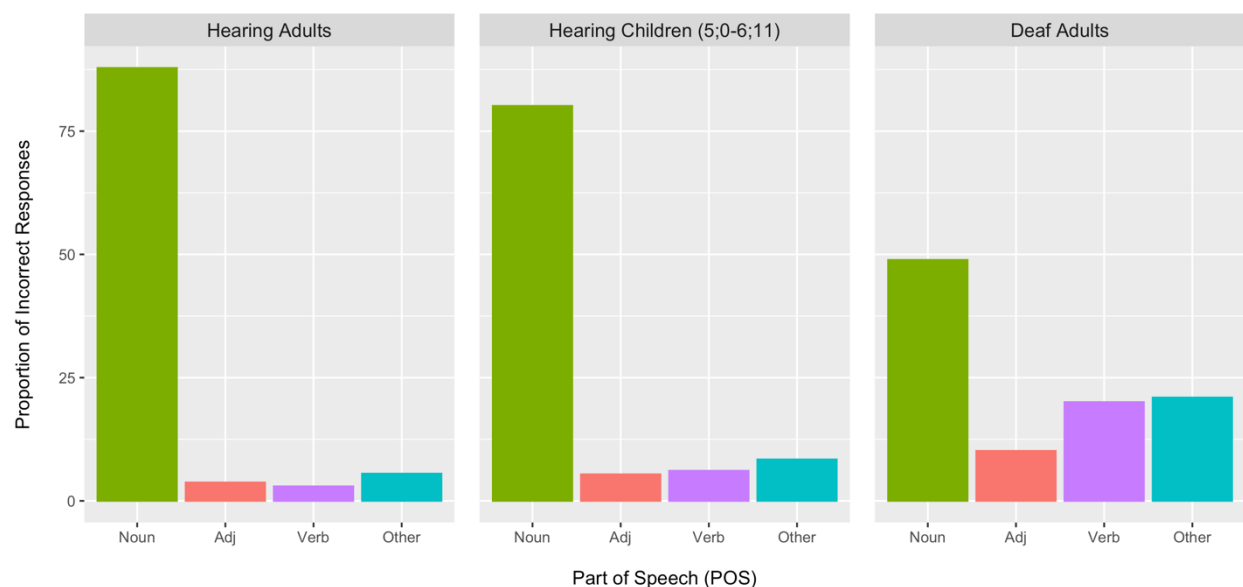


Figure 4.5. Comparing the variability in cloze responses across hearing and signing populations. We compared incorrect responses provided by hearing and deaf participants in a set of parallel cloze tasks. Specifically, we grouped incorrect responses into four syntactic categories: *nouns*, *adjectives*, *verbs*, and *other*. The *other* category contained responses like *adverbs*, *prepositions*, *number words*, etc. For the signing population, the *other* category also contained *classifier predicates*, *constructed actions*, and *fingerspelling*. The three panels represent the different participant populations: hearing adults (left), hearing children (middle), and deaf adults (right). Each bar represents the proportion of all incorrect responses (by population) that were classified as belonging to each syntactic category group. The target word in all cases was a lexicalized noun.

These results tentatively suggest that the baseline predictability of utterances in ASL may be lower due to more variable expectations about how an utterance is likely to continue. For these reasons, we might expect signers to be less likely to engage in form-based prediction during comprehension. At face value, the findings from the present study appear to support this claim. As we mentioned above, however, we are hesitant to accept these conclusions. In the final sections

below, we discuss a few limitations to the present study, which may better explain the lack of evidence for form-based prediction in ASL than the two hypotheses outlined above.

4.4.2. Addressing the moderate predictability of our target signs

One methodological limitation in the present study is the relatively low number of highly predictable target signs in our ASL story. In most psycholinguistic studies, highly predictable target words are classified as those with cloze probabilities greater than ~65% (Block & Baldwin, 2010; Brothers & Kuperberg, 2021; Kutas & Hillyard, 1984; Luke & Christianson, 2016). Using this particular threshold, only 28 of our target signs were highly predictable. The items in our higher cloze group had an average cloze probability of only 58.5% ($SD = 16.6\%$, $Range = 33\text{--}100\%$). To the best of our knowledge, all of the prior studies showing robust form-based prediction used high cloze items with an average cloze probability of at least 80% (DeLong et al., 2021; Ito et al., 2016; Kim & Lai, 2012; Laszlo & Federmeier, 2009). Thus, it is possible that we did not have enough highly predictable target words to evoke form-based predictions in our signers.

This limitation is something that researchers must address when using naturalistic stimuli. Prior work using cloze methods has shown that highly predictable target words are rather rare in natural discourses (Lowder et al., 2018; Luke & Christianson, 2016; Smith & Levy, 2013). To address this limitation, we would have needed to directly manipulate the plotlines and/or sentence structures in our story to create a set of target signs with artificially high levels of predictability. In doing so, we would have sacrificed the natural variability in word choice and syntactic structure, and perhaps made the story harder to follow for native ASL signers. Thus, we opted to use a fully natural stimulus and then select the highest and lowest cloze signs that occurred in our story.

It is also possible that naturally produced signs (i.e. those produced in conversation between two deaf signers) are generally more predictable than those in our translated story. If that is true, then the moderate predictability of our target signs is a unique property of the particular stimulus that we used. On the other hand, this finding could reflect a meaningful cross-linguistic difference in predictability between written, spoken, and sign languages. Future work should further investigate the degree to which signers can make explicit lexical predictions using a wider range of naturalistic stimuli and cloze tasks.

4.4.3. Addressing the variability in handshape similarity across our non-sign manipulations.

The moderate predictability of our target signs may have limited our ability to detect form-based prediction during sign comprehension. We suspect, however, that there is another feature of our stimuli that may be responsible for these findings—namely, the fact that handshapes naturally vary in their perceptual similarity to one another. In a sign language like ASL, there is a limited set of permissible handshapes (for discussion, see Whitworth, 2011). Moreover, these handshapes—much like phonemes in spoken language—can be further categorized using a set of smaller articulatory and perceptual features (Brentari, 2011; Stungis, 1981). For example, the feature COMPACT describes whether any of the middle fingers are extended: the “A” and “E” handshapes are considered to be [+COMPACT], whereas the “3” handshape is [−COMPACT] (e.g. Lane et al., 1976). In this example, it is clear that the “A” and “E” handshapes could be perceived as being more similar to one another than the “A” and “3” handshapes. Thus, it is possible that some of our non-signs were more similar to the intended target sign than others. If this is true, and signers are engaging in form-based prediction, we might expect that non-signs with more

perceptual (or featural) similarity will show smaller N400s than those with less similarity—but only in the higher cloze items.

There is some EEG evidence in the spoken language literature to support this hypothesis. In a study by Liu et al. (2006), a group of native Mandarin speakers listened to a set of sentences with predictable and unpredictable endings. The authors manipulated sentence-final words to be either the intended word from the sentence or one of the following nonwords: a *minimal-onset-mismatch* (a nonword with an onset that mismatched the original by one or two articulatory features); a *maximal-onset-mismatch* (a nonword with an onset that mismatched the original by two or more articulatory features); and a *first-syllable-mismatch* (a nonword with a completely different first syllable than the original word). The authors found N400 effects to all nonwords (regardless of predictability). In predictable sentences, however, there were graded N400 effects that varied as a function of the nonwords' similarity to the original word: *minimal-onset-mismatch* < *maximal-onset-mismatch* < *first-syllable-mismatch*. In unpredictable contexts, there were no graded N400 effects, as all three nonwords elicited similarly sized N400s. These findings provide evidence that listeners are sensitive to these sub-phonemic features, demonstrating graded sensitivity to words' similarities on these dimensions.

To investigate this phenomenon in the present study, we characterize the similarity between our target handshapes and the handshapes used in our non-sign manipulations using a set of similarity values from Stungis (1981). In this study, native deaf ASL signers identified a series of 20 handshapes in visual noise. The author obtained roughly 400 ratings for each handshape and used those ratings to calculate a confusion matrix (i.e. when given a particular handshape in visual noise, how often did the signer identify it as another handshape). For example, signers incorrectly identified the “A” handshape as being the “E” handshape in 61/404 trials. Thus, for the purpose of

this exploratory analysis, “A” and “E” would have a similarity value of 61. In contrast, “A” and “3” were only confused once, giving them a similarity value of 1.

We decided to find the similarities between the target signs and all of the non-signs (regardless of being a handshape-change or an all-change condition). Unfortunately, Stungis (1981) only tested a set of 20 handshapes—thus, we needed to restrict our analyses to the subset of trials in which the target and *both* non-sign handshapes appeared in that original study. Initial characterization of handshape similarity revealed that the handshapes in our *handshape-change* conditions were *less* similar to their target signs (*Mean similarity* = 8.06, *Range* = 0–37) than the handshapes in our *all-change* conditions (*Mean similarity* = 10.3, *Range* = 0–28). To understand the impact of these differences, we used a median split approach to separate items with similarity values greater than 6 into a *Similar* group and all others into a *Dissimilar* group. We also separated items into the higher and lower cloze groups from our prior analyses.

Following the logic above, we might expect that the N400s to non-signs with highly similar handshapes would be smaller than the N400s to non-signs with less similar handshapes in higher but not lower cloze conditions. In Figure 4.6, we demonstrate that when a non-sign’s handshape is more similar to the target handshape, the N400 response is reduced—but only in higher cloze conditions. In lower cloze conditions, there is no difference between the two non-sign manipulations (regardless of the similarity between their handshape and the target handshape).

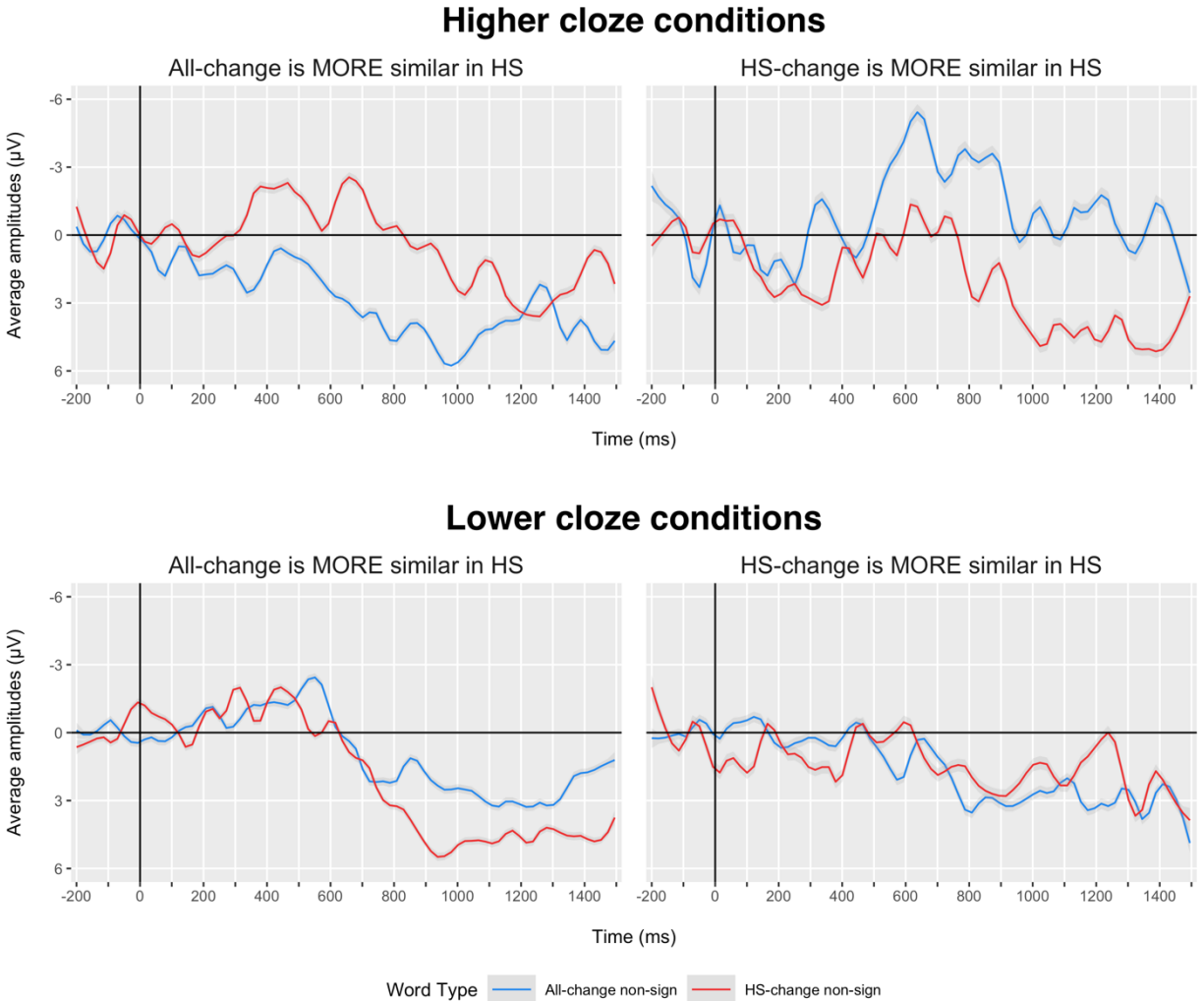


Figure 4.6. Average waveforms for non-signs across handshape similarity and predictability. Grand average waveforms from centroparietal electrodes are plotted for higher cloze (top panels) and lower cloze conditions (bottom panels). The waveforms from trials in which the all-change non-signs had a more similar handshape to the target than the handshape-change non-signs are on the left. The waveforms from trials with the opposite relationship are on the right. The red lines indicate handshape-change non-sign average waveforms, whereas the blue lines indicate all-change non-sign average waveforms. All waveforms were produced in R and then smoothed using local regression (loess) smoothing techniques.

One limitation to this analysis, however, is that there are only 34 trials in which both non-sign and the target handshapes appeared in Stungis (1981)—thus, we would need additional data in order to draw any strong conclusions from these patterns. In an effort to gain more power to observe an effect, we simply grouped non-signs by the similarity of their handshapes to the target words that they were meant to replace (ignoring whether they were originally handshape-changes

or all-changes). By collapsing in this way, we were able to use all of the data from trials with similarity values for at least one non-sign, increasing our number of usable trials to 71. In Figure 4.7, we show the grand average waveforms by predictability for similar and dissimilar handshape manipulations relative to their baseline controls (left panel). Figure 4.7 also presents the mean N400 amplitudes from our previous analysis, but rather than separating them by sign type, we separated these values by similarity group and predictability (right panel). Note, the original sign type for each observation is depicted as a red point (handshape-change non-sign) or a blue point (all-change non-sign).

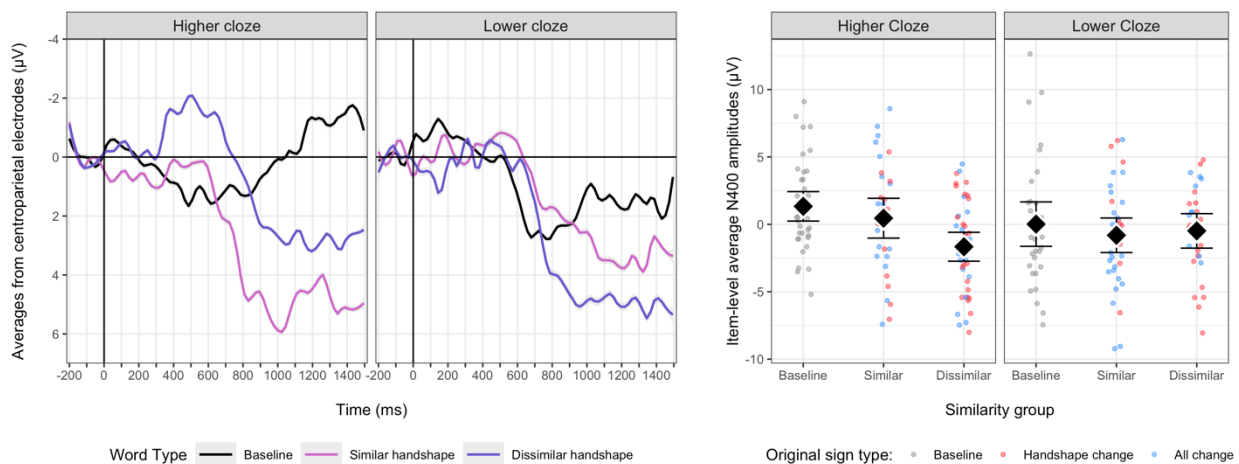


Figure 4.7. *Average waveforms across handshape similarity groups and predictability.* Grand average waveforms from centroparietal electrodes are plotted for higher cloze and lower cloze conditions. These waveforms indicate when the non-signs had a similar handshape to the target (pink lines, left panels) or a dissimilar handshape (purple lines, left panels). Item-level mean N400 amplitudes for baseline signs, similar non-signs, and dissimilar non-signs are plotted in the right panels. Gray points indicate baseline N400s, whereas red and blue points indicate the original non-sign type (e.g. handshape-change or all-change non-signs, respectively). Black diamonds represent the mean amplitude for the group. These waveforms were produced in R and then smoothed using local regression (loess) smoothing techniques.

These exploratory observations provide tentative evidence for form-based prediction in our study. The N400 responses to violations with similar handshapes to their respective target signs were reduced in predictable but not unpredictable contexts. Interestingly, this reduction did not

appear to be sensitive to whether the non-sign matched in location and movement with the target sign. Both handshape-change and all-change violations produced smaller N400s when their handshapes were perceptually similar to (and easily confusable with) the target handshape (Figures 4.6 and 4.7).

These data suggest that signers are particularly sensitive to handshape features during comprehension. Recent evidence from the priming literature provides tentative support for this claim. In a study by Meade et al. (2022), native deaf ASL signers showed equally robust priming effects when the target shared the same handshape with the prime, and when the target shared both the handshape *and* location with the prime. Thus, handshape could be a particularly prominent feature of sign recognition, perhaps similar to consonants in spoken language. Like consonants, handshapes are more perceptually distinct, greater in number, and are less likely to have a grammatical component. For example, altering a handshape is less likely to change the identity of that sign relative to altering its location or movement. Another possibility is that signers only generate expectations about handshapes, and not about location or movement, during comprehension. Handshape features seem to be fully articulated well before location or movement—thus, it may be advantageous to predict early emerging parameters relative to late emerging parameters. It would be interesting to learn in future studies that similarities in handshape alone are enough to explain the modulation of the N400 during sign comprehension.

4.5. Conclusion and future directions

The present study finds clear evidence for predictive processing in ASL signers during a naturalistic comprehension task. The N400 responses to non-manipulated baseline signs were smaller as a function of the signs' predictability in context. In most contemporary theories of the

N400, this evidence is argued to reflect the pre-activation of lexicosemantic features associated with a word (e.g. Federmeier, 2007; Kuperberg, 2007; Kuperberg & Jaeger, 2016, Section 3, p. 39; Kutas & Federmeier, 2011). We also explored the degree to which handshape similarity could have impacted our findings. In a series of exploratory analyses, we found that N400s were reduced to violations that have similar handshapes to the target sign that they were meant to replace in the story. These reductions, however, only seemed to emerge in predictable contexts, presumably when signers were able to pre-activate the handshape of an upcoming sign. Moreover, these reductions seemed to be independent of whether the violation overlapped in location and movement features with the intended sign, generating open questions about which features are actually predicted during naturalistic sign comprehension. In sum, the present study sought to characterize prediction in sign language, and we found promising (but tentative) evidence that prediction does occur in ASL—but confirmation of these conclusions will require further theoretical and statistical testing.

In our future research, we plan to continue investigating predictive processing in ASL. Specifically, we have identified three projects to pursue: First, we plan to replicate the finding that the N400 response to non-manipulated signs has an inverse linear correlation with the predictability of that sign in its context (e.g. DeLong et al., 2005; Kutas & Hillyard, 1984). In the present study, our target items were not evenly distributed across the possible range of cloze probabilities, restricting our ability to assess any continuous relationship between the N400 and the predictability of a sign. We therefore plan to identify and use a stimulus with target items that sufficiently represent the continuous nature of cloze probabilities. Second, we hope to further explore signers' sensitivity to handshape features and to determine whether signers specifically target this parameter in their predictions. To do this, we will need a careful set of stimuli that

balances the perceptual similarity between the three main parameters in ASL. Finally, in the present study, we narrowly focused on how reading experience influenced prediction in one's primary linguistic modality. To do this, we investigated the nature of prediction in signers who are also fluent readers of English. There are, however, interesting questions about whether signers make robust predictions while reading. It is possible that, for highly literate signers, prediction is quite robust in their written language but not in their sign language. A simple approach to this question would be to have signers participate in the prior reading studies that clearly demonstrated form-based prediction with hearing participants.

Chapter 5

[Conclusion]

5.1. Conclusion

The field of psycholinguistics has long known that language comprehension involves prediction. What has changed over the past few decades is that prediction has become a core aspect of our linguistic theories and contemporary research into language processing. Thus, the current challenges faced by prediction researchers are not to simply prove the existence of prediction, but rather to characterize the specificity of our predictions, to assess the benefits and limitations to making them, and to understand the ways in which prediction emerges (and functions) in the world's many languages. In this dissertation, I presented three studies that began to tackle these challenges by assessing prediction in a wide range of populations and linguistic environments. This Conclusion serves as a brief summary of the key findings from each study and the overall successes of the *Storytime* paradigm.

5.2. Summarizing the key findings from Papers 1–3

The primary goal of this dissertation was to understand whether form-based prediction persists in more natural language contexts. Over the course of my undergraduate and graduate training, I have seen comprehenders achieve remarkable feats in our typical psycholinguistic experiments, e.g., they reason about complex syntactic structures and dependencies, they predict

what someone is about to say (and then act upon those predictions), and they even contort language to arrive at novel linguistic interpretations (e.g. the process of *verbing* nouns or creating new, artificial languages). I have also seen how the language that we study in the lab can be rigid, contrived, and overly simplistic at times. It is undeniable that, as a field, we have benefitted immensely from using tightly controlled language and clever experimental designs in our research—but now, we have access to new and exciting techniques that allow us to explore whether the insights gleaned from foundational studies can explain language processing in the wild.

For example, a classic finding from the ERP literature is that the amplitude of the N400 response is inversely correlated with the predictability of a word in its context (e.g. DeLong et al., 2005; Kutas et al., 1984). Specifically, as the predictability of a word increases, the N400 response evoked by that word decreases (or becomes more positive). There are many principled reasons to believe that this finding might be a product of the kinds of paradigms and sentences that we use in the lab. Most experimental designs place predictable words at the ends of sentences and ensure that the unfolding context provides enough support to predict those words. Moreover, researchers often present sentences at slower-than-natural rates, providing additional time for comprehenders to process the input and make inferences about what is likely to come next. In contrast, natural language has been shown to be less predictable on average, rampant with disfluencies and redundancies, and produced at a wider range of presentation rates.

For these reasons, it would be important to understand whether this classic finding extends to more naturalistic settings—and indeed, recent studies have shown that the N400 responses to words in natural, written discourses still correlate with word-level predictability despite being in less predictable and more variable linguistic environments (e.g. Lowder et al., 2018; Luke & Christianson, 2016; Smith & Levy, 2013). For comprehension of natural speech, we also replicated

this effect using the *Storytime* paradigm. In Figure 5.1, I present the data from Papers 1–3, which show that the N400s evoked by the non-manipulated, baseline words decrease (or become more positive) as the predictability of the word increases. This simple replication not only affirms the validity of our theories and the insights from prior work in the lab, but it also provides a proof of concept that language processing can be studied in more naturalistic settings.

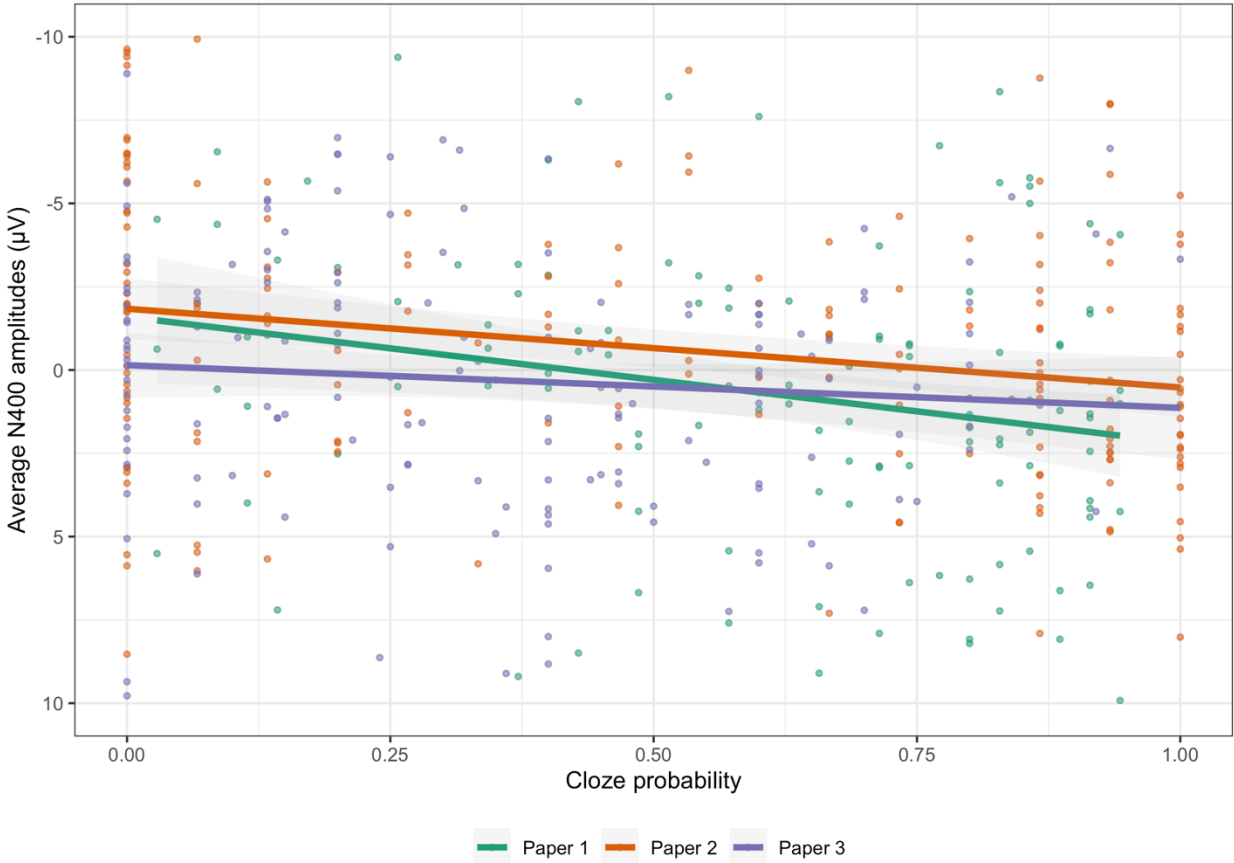


Figure 5.1. Baseline N400 responses across all three papers in this dissertation. This plot demonstrates the relationship between the N400 responses and the predictability of non-manipulated, baseline words in Paper 1 (Spanish-English bilinguals), Paper 2 (English-speaking adults), and Paper 3 (Deaf ASL signers). Each dot represents the by-item average for all baseline words in each paper.

5.2.1. Summary of Paper 1

At the broadest level, Paper 1 explores whether well-known linguistic phenomena can be understood within a general predictive framework of language comprehension. At the narrowest level, we wanted to see if code-switching—i.e., when bilingual speakers shift between their two languages within a single utterance—could provide insight into the nature of comprehenders’ expectations during comprehension. As discussed in Paper 1 (Chapter 2), bilinguals have access to linguistic systems in which a single (lexical) concept can often be expressed by two distinct word-forms (one from each of their languages). As a result, bilingual comprehension is an interesting case study on whether comprehenders make predictions about specific lexical word-forms, or whether their predictions are simply about the meaning of upcoming linguistic material.

Paper 1 addressed these questions by constructing tightly controlled experimental manipulations and splicing them into naturally spoken narratives. Specifically, we manipulated a set of target words within two stories in English (e.g. “...And the wig is so hot and heavy on my...**head**”). For each target word, we either spliced in the identical word (*head*) or replaced it with one of three unexpected words: a Spanish translation of the original word (*cabeza*), an unexpected English word (*cranium*), or the Spanish translation of that unexpected English word (*cráneo*). Using this design, we could investigate the N400 responses to both code-switches and non-switches, as well as to words that either strongly or weakly fit into their story contexts.

Prior work has argued that code-switches, regardless of their contextual fit, incur processing costs on the N400 (e.g. Alvarez et al., 2003; Fernandez et al., 2019; van Hell et al., 2018; van Hell et al., 2015). On this account, there are two effects reflected in the N400 response: the typical N400 effect for unexpected words and the N400 effect for switching into another language. Paper 1 argues that, if there are distinct costs to code-switching on the N400, then there

should be an additive effect for code-switches that weakly fit within their contexts (*cráneo*). Alternatively, if the N400 effects to code-switches in the prior literature simply reflect bilinguals predicting a specific word in a particular language, then we should not see any evidence of an additive effect.

In Paper 1, we found four main effects: First, all of the unexpected words evoked similarly sized N400 responses—i.e., we found no evidence for a unique code-switching cost incurred on the N400. Second, the N400 responses to the original words from the story varied across predictability (see Figure 5.1). Third, as shown in prior research, code-switched words evoked late posterior positivities relative to non-switched words. These positivities demonstrated bilinguals' recognition of the language switches, and further confirmed their sensitivity to the experimental manipulations in our naturalistic task. Finally, the words that only weakly fit within the unfolding story context produced broad, sustained negativities, suggesting that integration of these words into the discourse was perhaps more challenging than that of the words with more congruent meanings. Taken together, these findings show that bilinguals make predictions about specific words during comprehension. Moreover, Paper 1 demonstrated that classic ERP effects can be replicated using our *Storytime* paradigm, paving the way for the use of this technique to study the nature of language processing in a wider range of linguistic environments.

5.2.2. Summary of Paper 2

In light of the successes from Paper 1, we aimed to further characterize the nature of comprehenders' form-based predictions in Paper 2 (Chapter 3). To do this, we asked whether English-speaking adults predict the phonological features of upcoming words during natural language comprehension. As I discussed throughout this dissertation, prior work has shown that

form-based prediction seems to require highly predictable target words and relatively slow rates of presentation (e.g. Ito et al., 2016; but see, DeLong et al., 2021). For these reasons, many theorists have argued that form-based prediction is unlikely to emerge in more naturalistic contexts (Freunberger & Roehm, 2016, 2017; Indefrey & Levelt, 2004; Ito et al., 2016; Pickering & Garrod, 2007).

To address this claim, we again explored form-based prediction using the *Storytime* paradigm. Specifically, we created an entire 30-minute cartoon narration of a children’s storybook, and carefully spliced in a set of experimental manipulations. These manipulations created a typical 2×3 design within our story context, crossing factors of target word predictability (higher cloze, lower cloze) and word type (baseline, form-similar, and form-dissimilar). To create our form-similar violations, we simply changed the initial vowel of our baseline words (*cake* → *ceke*). For less similar violations, we changed the onset of the baseline word to differ in voicing, manner of articulation, and place of articulation (*cake* → *vake*). Following prior work, we anticipated that if comprehenders predict form features during naturalistic comprehension, we should find reduced N400s to form-similar violations (*ceke*) relative to less similar violations (*vake*)—but only in higher cloze environments.

In Paper 2, we found three main effects: First, we replicated the reduced N400s to form-similar (*ceke*) violations in higher but not lower cloze contexts. Second, we also found late posterior positivities to both violations, again confirming comprehenders’ sensitivity to our manipulations throughout the cartoon story. Finally, as in Paper 1, we demonstrated that the N400 responses to our baseline words varied across predictability (see Figure 5.1). Moreover, these data patterns have been self-replicated, as we ran this study with an initial set of 30 pilot participants and found identical results. Taken together, Paper 2 provided clear (and replicable) evidence that

English-speaking adults make predictions about the phonological features of upcoming words while listening to naturally produced speech.

5.2.3. *Summary of Paper 3*

Papers 1 and 2 demonstrated that form-based prediction is common in spoken language comprehension. Given the robustness of these findings, we wanted to assess how widespread this phenomenon is by looking at form-based prediction in the visual-manual modality. Nearly all of the prior work on prediction has focused on written and spoken language. Thus, in Paper 3 (Chapter 4), we characterized prediction in native deaf signers as they watched an hour-long story in ASL. To assess prediction in ASL, we constructed a 2×3 design like the one in Paper 2 by interweaving various recordings and counterbalancing which manipulations were presented to participants. Specifically, we identified both predictable and unpredictable target signs, and then kept those original signs in the story or replaced them with form-similar and form-dissimilar violations. To create form-similar violations, we simply altered the handshape of the original sign. To create form-dissimilar violations, we altered the handshape, location, and movement of the original sign. Following the logic from Paper 2, if signers generate expectations for the visual-manual features of upcoming signs, then there should be reductions in the N400 to form-similar violations (handshape-change) relative to form-dissimilar violations (all-change).

In our primary analysis, we found three main effects: First, all violations produced late posterior positivities, again confirming signers' sensitivity to our experimental manipulations within this naturalistic context. Second, the N400 responses to baseline signs varied across predictability contexts such that higher cloze signs produced smaller N400s than lower cloze signs (see Figure 5.1, although the significant effect in Paper 3 is categorical, not continuous). Third, we

saw large N400 effects to violations in higher cloze conditions, but there was no evidence of reduced N400s for the handshape-change non-signs. These findings initially suggested that form-based prediction did not emerge in Paper 3—however, we realized that both violations mismatched the original target sign in handshape. If signers are particularly sensitive to handshape during comprehension, we might expect that these two violations would produce similar N400 effects.

In an exploratory investigation, we assessed how the similarity between the handshapes in our violations and the handshape of the original sign influenced the N400 response. We found reduced N400s for handshape-*similar* violations relative to violations with more dissimilar handshapes but only in higher cloze environments. These findings tentatively suggest that deaf signers are predicting the handshapes of upcoming signs during comprehension—although, additional studies and statistical analyses are necessary to provide conclusive evidence for this claim.

5.3. Concluding summary

In this dissertation, we debuted the *Storytime* paradigm, a novel technique for studying language processing in more naturalistic contexts. To do this, we took the tightly controlled experimental designs from prior studies and injected them in natural language stimuli. This technique provides the best of both worlds, i.e., the degree of experimental control that is necessary for testing our hypotheses, and the ecological validity of using naturalistic stimuli to study comprehension in the wild. Across these three studies, the results are rather unambiguous: comprehenders predict the form of upcoming words in natural storytelling contexts. In this dissertation, I showed predictive processing in a wide range of language users, spanning bilingual speakers to deaf signers of ASL. To conclude, this body of work highlights the utility of the

Storytime paradigm for understanding the cognitive processes underlying everyday language comprehension—and most importantly, this technique makes language research accessible to communities and populations that have been historically underserved by our current methodologies.

References

- Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language*, 38(4), 419–439. <https://doi.org/10.1006/jmla.1997.2558>
- Altarriba, J., Kroll, J. F., Sholl, A., & Rayner, K. (1996). The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. *Memory & Cognition*, 24(4), 477–492.
- Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, 30(3), 191–238. [https://doi.org/10.1016/0010-0277\(88\)90020-0](https://doi.org/10.1016/0010-0277(88)90020-0)
- Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264. [https://doi.org/10.1016/s0010-0277\(99\)00059-1](https://doi.org/10.1016/s0010-0277(99)00059-1)
- Altmann, G. T. M., & Mirković, J. (2009). Incrementality and Prediction in Human Sentence Processing. *Cognitive Science*, 33(4), 583–609. <https://doi.org/10.1111/j.1551-6709.2009.01022.x>
- Alvarez, R. P., Holcomb, P. J., & Grainger, J. (2003). Accessing word meaning in two languages: An event-related brain potential study of beginning bilinguals. *Brain and Language*, 87, 290–304. [https://doi.org/10.1016/S0093-934X\(03\)00108-1](https://doi.org/10.1016/S0093-934X(03)00108-1)
- Andrew, K. N., Hoshoooley, J., & Joannisse, M. F. (2014). Sign Language Ability in Young Deaf Signers Predicts Comprehension of Written Sentences in English. *PLoS ONE*, 9(2), e89994. <https://doi.org/10.1371/journal.pone.0089994>
- Atkinson, E., Wagers, M. W., Lidz, J., Phillips, C., & Omaki, A. (2018). Developing incrementality in filler-gap dependency processing. *Cognition*, 179, 132–149. <https://doi.org/10.1016/j.cognition.2018.05.022>
- Auer, P. (1988). A conversation analytic approach to code-switching and transfer. *Code-Switching: Anthropological and Linguistic Perspectives*, 187–214.
- Aurnhammer, C., & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, 134, 107198. <https://doi.org/10.1016/j.neuropsychologia.2019.107198>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology*, 17, 364–390. [https://doi.org/10.1016/0010-0285\(85\)90013-1](https://doi.org/10.1016/0010-0285(85)90013-1)

- Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11(7), 280–289. <https://doi.org/10.1016/j.tics.2007.05.005>
- Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. Context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 238–247. <https://doi.org/10.3115/v1/P14-1023>
- Barr, D. (2021). *Learning statistical models through simulation in R: An interactive textbook. Version 1.0.0*. <https://psyteachr.github.io/stat-models-v1>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting Linear Mixed-Effects Models using lme4* (arXiv:1406.5823). arXiv. <http://arxiv.org/abs/1406.5823>
- Beal, J. S., Scott, J. A., & Spell, K. (2021). Goodnight Gorilla: Deaf Student American Sign Language Narrative Renditions After Viewing a Model. *The Journal of Deaf Studies and Deaf Education*, 26(1), 85–98. <https://doi.org/10.1093/deafed/enaa022>
- Bellugi, U., & Fischer, S. (1972). A comparison of sign language and spoken language. *Cognition*, 1, 173–200. [https://doi.org/10.1016/0010-0277\(72\)90018-2](https://doi.org/10.1016/0010-0277(72)90018-2)
- Beres, A. M. (2017). Time is of the Essence: A Review of Electroencephalography (EEG) and Event-Related Brain Potentials (ERPs) in Language Research. *Applied Psychophysiology and Biofeedback*, 42(4), 247–255. <https://doi.org/10.1007/s10484-017-9371-3>
- Berko, J. (1958). The Child's Learning of English Morphology. *WORD*, 14(2–3), 150–177. <https://doi.org/10.1080/00437956.1958.11659661>
- Bhattachali, S., Brennan, J., Luh, W.-M., Franzluebbers, B., & Hale, J. (2020). The Alice Datasets: fMRI & EEG Observations of Natural Language Comprehension. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 120–125. <https://aclanthology.org/2020.lrec-1.15>
- Biersack, S., Kempe, V., & Knapton, L. (2005). Fine-tuning speech registers: A comparison of the prosodic features of child-directed and foreigner-directed speech. *Interspeech 2005*, 2401–2404. <https://doi.org/10.21437/Interspeech.2005-46>
- Blanco-Elorrieta, E., & Pylkkänen, L. (2016). Bilingual Language Control in Perception versus Action: MEG Reveals Comprehension Control Mechanisms in Anterior Cingulate Cortex and Domain-General Control of Production in Dorsolateral Prefrontal Cortex. *The Journal of Neuroscience*, 36(2), 290–301. <https://doi.org/10.1523/JNEUROSCI.2597-15.2016>
- Blanco-Elorrieta, E., & Pylkkänen, L. (2017). Bilingual Language Switching in the Laboratory versus in the Wild: The Spatiotemporal Dynamics of Adaptive Language Control. *The Journal of Neuroscience*, 37(37), 9022–9036. <https://doi.org/10.1523/JNEUROSCI.0553-17.2017>

- Blanco-Elorrieta, E., & Pylkkänen, L. (2018). Ecological Validity in Bilingualism Research and the Bilingual Advantage. *Trends in Cognitive Sciences*, 22(12), 1117–1126. <https://doi.org/10.1016/j.tics.2018.10.001>
- Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods*, 42(3), 665–670. <https://doi.org/10.3758/BRM.42.3.665>
- Bloomfield, L. (1984). *Language*. University of Chicago Press. <https://press.uchicago.edu/ucp/books/book/chicago/L/bo3636364.html>
- Boersma, P., & Weenink, D. (2001). *Praat speech processing software*. <http://www.praat.org>.
- Boland, J. E., Tanenhaus, M. K., Garnsey, S. M., & Carlson, G. N. (1995). Verb argument structure in parsing and interpretation: Evidence from wh-questions. *Journal of Memory and Language*, 34, 774–806. <https://doi.org/10.1006/jmla.1995.1034>
- Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Toward a Neurobiologically Plausible Model of Language-Related, Negative Event-Related Potentials. *Frontiers in Psychology*, 10. <https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00298>
- Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one’s age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology*, 112(4), 417–436. <https://doi.org/10.1016/j.jecp.2012.01.005>
- Borovsky, A., Elman, J. L., & Kutas, M. (2012). Once is Enough: N400 Indexes Semantic Integration of Novel Word Meanings from a Single Exposure in Context. *Language Learning and Development*, 8(3), 278–302. <https://doi.org/10.1080/15475441.2011.614893>
- Boston, M., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2. <https://doi.org/10.16910/jemr.2.1.1>
- Braze, D., Tabor, W., Shankweiler, D. P., & Mencl, W. E. (2007). Speaking Up for Vocabulary: Reading Skill Differences in Young Adults. *Journal of Learning Disabilities*, 40(3), 226–243. <https://doi.org/10.1177/00222194070400030401>
- Brennan, J. (2016). Naturalistic Sentence Comprehension in the Brain. *Language and Linguistics Compass*, 10(7), 299–313. <https://doi.org/10.1111/lnc3.12198>
- Brennan, J. R., & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE*, 14(1), e0207741. <https://doi.org/10.1371/journal.pone.0207741>

- Brennan, J., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157*, 81–94. <https://doi.org/10.1016/j.bandl.2016.04.008>
- Brentari, D. (2011). Handshape in Sign Language Phonology: Underlying Representations. In M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell Companion to Phonology* (pp. 1–28). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781444335262.wbctp0009>
- Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology: CB*, *28*(5), 803-809.e3. <https://doi.org/10.1016/j.cub.2018.01.080>
- Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, *116*, 104174. <https://doi.org/10.1016/j.jml.2020.104174>
- Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition*, *136*, 135–149. <https://doi.org/10.1016/j.cognition.2014.10.017>
- Brothers, T., 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 (Cambridge, Mass.)*, *1*(1), 135–160. https://doi.org/10.1162/nol_a_00006
- Brothers, T., Zeitlin, M., Perrachione, A. C., Choi, C., & Kuperberg, G. (2022). Domain-general conflict monitoring predicts neural and behavioral indices of linguistic error processing during reading comprehension. *Journal of Experimental Psychology: General*, *151*, 1502–1519. <https://doi.org/10.1037/xge0001130>
- Brown, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of Cognitive Neuroscience*, *5*, 34–44. <https://doi.org/10.1162/jocn.1993.5.1.34>
- Brown, R., & Berko, J. (1960). Word association and the acquisition of grammar. *Child Development*, 1–14.
- Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), Article 1. <https://doi.org/10.5334/joc.72>
- Brysbaert, M., & New, B. (2009). Moving beyond Kucera 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. <https://doi.org/10.3758/BRM.41.4.977>

- Bullock, B. E., & Toribio, A. J. (2019). Conceptual and Empirical Arguments for a Language Feature: Evidence from Language Mixing. In D. L. Arteaga (Ed.), *Contributions of Romance Languages to Current Linguistic Theory* (pp. 93–113). Springer International Publishing. https://doi.org/10.1007/978-3-030-11006-2_5
- Bultena, S., Dijkstra, T., & Hell, J. G. V. (2015). Language switch costs in sentence comprehension depend on language dominance: Evidence from self-paced reading. *Bilingualism: Language and Cognition*, *18*(3), 453–469. <https://doi.org/10.1017/S1366728914000145>
- Caffarra, S., Mendoza, M., & Davidson, D. (2019). Is the LAN effect in morphosyntactic processing an ERP artifact? *Brain and Language*, *191*, 9–16. <https://doi.org/10.1016/j.bandl.2019.01.003>
- Capek, C. M., Grossi, G., Newman, A. J., McBurney, S. L., Corina, D., Roeder, B., & Neville, H. J. (2009). Brain systems mediating semantic and syntactic processing in deaf native signers: Biological invariance and modality specificity. *Proceedings of the National Academy of Sciences*, *106*(21), 8784–8789. <https://doi.org/10.1073/pnas.0809609106>
- Caramazza, A. (1997). How Many Levels of Processing Are There in Lexical Access? *Cognitive Neuropsychology*, *14*(1), 177–208. <https://doi.org/10.1080/026432997381664>
- Caramazza, A. (1998). The interpretation of semantic category-specific deficits: What do they reveal about the organization of conceptual knowledge in the brain? *Neurocase*, *4*, 265–272. <https://doi.org/10.1080/13554799808410627>
- Caramazza, A., & Brones, I. (1979). Lexical access in bilinguals. *Bulletin of the Psychonomic Society*, *13*(4), 212–214. <https://doi.org/10.3758/BF03335062>
- Carroll, J. B. (1971). *Defining Language Comprehension: Some Speculations*.
- Cates, D., Gutiérrez, E., Hafer, S., Barrett, R., & Corina, D. (2013). Location, Location, Location. *Sign Language Studies*, *13*(4), 433–461.
- Chamberlain, C., Morford, J. P., & Mayberry, R. I. (1999). *Language Acquisition By Eye*. Psychology Press.
- Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234–272. <https://doi.org/10.1037/0033-295X.113.2.234>
- Chauncey, K., Grainger, J., & Holcomb, P. J. (2008). Code-switching effects in bilingual word recognition: A masked priming study with event-related potentials. *Brain and Language*, *105*(3), 161–174. <https://doi.org/10.1016/j.bandl.2007.11.006>
- Cheng, Q., & Mayberry, R. I. (2019). Acquiring a First Language in Adolescence: The Case of Basic Word Order in American Sign Language. *Journal of Child Language*, *46*(2), 214–240. <https://doi.org/10.1017/S0305000918000417>

- Christoffels, I. K., Firk, C., & Schiller, N. O. (2007). Bilingual language control: An event-related brain potential study. *Brain Research, 1147*, 192–208. <https://doi.org/10.1016/j.brainres.2007.01.137>
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204. <https://doi.org/10.1017/S0140525X12000477>
- Cleland, A. A., & Pickering, M. (2006). Do writing and speaking employ the same syntactic representations? *Journal of Memory and Language, 54*(2), 185–198. <https://doi.org/10.1016/j.jml.2005.10.003>
- Connolly, J. F., & Phillips, N. A. (1994). Event-Related Potential Components Reflect Phonological and Semantic Processing of the Terminal Word of Spoken Sentences. *Journal of Cognitive Neuroscience, 6*(3), 256–266. <https://doi.org/10.1162/jocn.1994.6.3.256>
- Connolly, J. F., Phillips, N. A., & Forbes, K. A. K. (1995). The effects of phonological and semantic features of sentence-ending words on visual event-related brain potentials. *Electroencephalography and Clinical Neurophysiology, 94*(4), 276–287. [https://doi.org/10.1016/0013-4694\(95\)98479-R](https://doi.org/10.1016/0013-4694(95)98479-R)
- Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6*, 84–107. [https://doi.org/10.1016/0010-0285\(74\)90005-X](https://doi.org/10.1016/0010-0285(74)90005-X)
- Cormier, K., Smith, S., & Zwets, M. (2013). Framing constructed action in British Sign Language narratives. *Journal of Pragmatics, 55*, 119–139. <https://doi.org/10.1016/j.pragma.2013.06.002>
- Coulson, S., King, J., & Kutas, M. (1998). Expect the Unexpected: Event-related Brain Response to Morphosyntactic Violations. *Lang. Cogn. Process, 13*, 21–58. <https://doi.org/10.1080/016909698386582>
- Coulson, S., & Kutas, M. (2001). Getting it: Human event-related brain response to jokes in good and poor comprehenders. *Neuroscience Letters, 316*(2), 71–74. [https://doi.org/10.1016/S0304-3940\(01\)02387-4](https://doi.org/10.1016/S0304-3940(01)02387-4)
- Courteau, É., Martignetti, L., Royle, P., & Steinhauer, K. (2019). Eliciting ERP Components for Morphosyntactic Agreement Mismatches in Perfectly Grammatical Sentences. *Frontiers in Psychology, 10*. <https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01152>
- Cuetos, F., Glez-Nosti, M., Barbon, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Frecuencias de las palabras españolas basadas en los subtítulos de las películas. *Psicológica, 32*(2), 133–144.

- Dambacher, M., Dimigen, O., Braun, M., Wille, K., Jacobs, A. M., & Kliegl, R. (2012). Stimulus onset asynchrony and the timeline of word recognition: Event-related potentials during sentence reading. *Neuropsychologia*, *50*(8), 1852–1870. <https://doi.org/10.1016/j.neuropsychologia.2012.04.011>
- de Groot, A. M. B. (1993). Word-type effects in bilingual processing tasks: Support for a mixed-representational system. In *The bilingual lexicon* (pp. 27–51). John Benjamins Publishing Company. <https://doi.org/10.1075/sibil.6.04gro>
- de Groot, A. M., & Nas, G. L. (1991). Lexical representation of cognates and noncognates in compound bilinguals. *Journal of Memory and Language*, *30*, 90–123. [https://doi.org/10.1016/0749-596X\(91\)90012-9](https://doi.org/10.1016/0749-596X(91)90012-9)
- de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390. <https://doi.org/10.3758/BRM.41.2.385>
- Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283–321. <https://doi.org/10.1037/0033-295X.93.3.283>
- Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120394. <https://doi.org/10.1098/rstb.2012.0394>
- DeLong, K. A., Chan, W., & Kutas, M. (2019). Similar time courses for word form and meaning preactivation during sentence comprehension. *Psychophysiology*, *56*(4), e13312. <https://doi.org/10.1111/psyp.13312>
- DeLong, K. A., Chan, W., & Kutas, M. (2021). Testing limits: ERP evidence for word form preactivation during speeded sentence reading. *Psychophysiology*, *58*(2). <https://doi.org/10.1111/psyp.13720>
- DeLong, K. A., Chan, W.-H., & Kutas, M. (2020). Testing limits: ERP evidence for word form preactivation during speeded sentence reading. *Psychophysiology*, *58*(2). <https://doi.org/10.1111/psyp.13720>
- DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150–162. <https://doi.org/10.1016/j.neuropsychologia.2014.06.016>
- DeLong, K. A., Troyer, M., & Kutas, M. (2014). Pre-Processing in Sentence Comprehension: Sensitivity to Likely Upcoming Meaning and Structure. *Language and Linguistics Compass*, *8*(12), 631–645. <https://doi.org/10.1111/lnc3.12093>
- 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. <https://doi.org/10.1038/nn1504>

- DeLong, K. A., Urbach, T. P., & Kutas, M. (2017). Is there a replication crisis? Perhaps. Is this an example? No: a commentary on Ito, Martin, and Nieuwland (2016). *Language, Cognition and Neuroscience*, 32(8), 966–973. <https://doi.org/10.1080/23273798.2017.1279339>
- Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1), 9–21. <https://doi.org/10.1016/j.jneumeth.2003.10.009>
- Deuchar, M., Davies, P., Herring, J. R., Couto, M. C. P., & Carter, D. (2014). 5. Building Bilingual Corpora. In 5. *Building Bilingual Corpora* (pp. 93–110). Multilingual Matters. <https://doi.org/10.21832/9781783091713-008>
- Dijkstra, T. (2005). Bilingual Visual Word Recognition and Lexical Access. In *Handbook of bilingualism: Psycholinguistic approaches* (pp. 179–201). Oxford University Press.
- Dijkstra, T., Grainger, J., & Van Heuven, W. J. B. (1999). Recognition of Cognates and Interlingual Homographs: The Neglected Role of Phonology. *Journal of Memory and Language*, 41(4), 496–518. <https://doi.org/10.1006/jmla.1999.2654>
- Dijkstra, T., Van Hell, J. G., & Brenders, P. (2015). Sentence context effects in bilingual word recognition: Cognate status, sentence language, and semantic constraint. *Bilingualism: Language and Cognition*, 18, 597–613. <https://doi.org/10.1017/S1366728914000388>
- Dijkstra, T., & Van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, 5(3), 175–197. <https://doi.org/10.1017/S1366728902003012>
- Dikker, S., & Pykkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation. *Brain and Language*, 127(1), 55–64. <https://doi.org/10.1016/j.bandl.2012.08.004>
- Duyck, W., Van Assche, E., Drieghe, D., & Hartsuiker, R. J. (2007). Visual word recognition by bilinguals in a sentence context: Evidence for nonselective lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(4), 663–679. <https://doi.org/10.1037/0278-7393.33.4.663>
- Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24, 409–436. <https://doi.org/10.1007/BF02143160>
- Ehri, L. C. (1987). Learning to Read and Spell Words. *Journal of Reading Behavior*, 19(1), 5–31. <https://doi.org/10.1080/10862968709547585>
- Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning & Verbal Behavior*, 20, 641–655. [https://doi.org/10.1016/S0022-5371\(81\)90220-6](https://doi.org/10.1016/S0022-5371(81)90220-6)

- Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, 27(2), 143–165. [https://doi.org/10.1016/0749-596X\(88\)90071-X](https://doi.org/10.1016/0749-596X(88)90071-X)
- Emmorey, K., Borinstein, H., Thompson, R., & Gollan, T. (2008). Bimodal bilingualism. *Bilingualism (Cambridge, England)*, 11(1), 43–61. <https://doi.org/10.1017/S1366728907003203>
- Emmorey, K., & Lane, H. L. (2013). *The Signs of Language Revisited: An Anthology To Honor Ursula Bellugi and Edward Klima*. Psychology Press.
- Favier, S., Meyer, A. S., & Huettig, F. (2021). Literacy can enhance syntactic prediction in spoken language processing. *Journal of Experimental Psychology: General*, 150(10), 2167–2174. <https://doi.org/10.1037/xge0001042>
- Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, 44(4), 491–505. <https://doi.org/10.1111/j.1469-8986.2007.00531.x>
- Federmeier, K. D. (2022). Connecting and considering: Electrophysiology provides insights into comprehension. *Psychophysiology*, 59(1), e13940. <https://doi.org/10.1111/psyp.13940>
- 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. <https://doi.org/10.1006/jmla.1999.2660>
- Federmeier, K. D., McLennan, D. B., De Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology*, 39, 133–146. <https://doi.org/10.1111/1469-8986.3920133>
- Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, 1146, 75–84. <https://doi.org/10.1016/j.brainres.2006.06.101>
- Fernald, A., Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardies, B. de, & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants*. *Journal of Child Language*, 16(3), 477–501. <https://doi.org/10.1017/S0305000900010679>
- Fernandez, C. B., Litcofsky, K. A., & van Hell, J. (2019). Neural correlates of intra-sentential code-switching in the auditory modality. *Journal of Neurolinguistics*, 51, 17–41. <https://doi.org/10.1016/j.jneuroling.2018.10.004>
- Ferrara, L., & Hodge, G. (2018). Language as Description, Indication, and Depiction. *Frontiers in Psychology*, 9. <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00716>

- Ferreira, F., & Henderson, J. M. (1990). Use of verb information in syntactic parsing: Evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 555–568. <https://doi.org/10.1037/0278-7393.16.4.555>
- Fields, E. C. (2017). *Factorial mass univariate ERP toolbox*.
- Fields, E. C. (2019). *Using FMUT [Github Wiki Page]*. <https://github.com/ericcfields/FMUT/wiki>
- Fields, E. C., & Kuperberg, G. R. (2020). Having your cake and eating it too: Flexibility and power with mass univariate statistics for ERP data. *Psychophysiology*, *57*(2), e13468. <https://doi.org/10.1111/psyp.13468>
- Fischler, I., & Bloom, P. A. (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. *Journal of Verbal Learning and Verbal Behavior*, *18*(1), 1–20. [https://doi.org/10.1016/S0022-5371\(79\)90534-6](https://doi.org/10.1016/S0022-5371(79)90534-6)
- FitzPatrick, I., & Indefrey, P. (2014). Head start for target language in bilingual listening. *Brain Research*, *1542*, 111–130.
- Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*(6), 627–635.
- Foucart, A., & Frenck-Mestre, C. (2011). Grammatical gender processing in L2: Electrophysiological evidence of the effect of L1–L2 syntactic similarity. *Bilingualism: Language and Cognition*, *14*(3), 379–399. <https://doi.org/10.1017/S136672891000012X>
- Foucart, A., & Frenck-Mestre, C. (2012). “Can late L2 learners acquire new grammatical features? Evidence from ERPS and eye-tracking”: Corrigendum. *Journal of Memory and Language*, *67*, 238–238. <https://doi.org/10.1016/j.jml.2012.02.009>
- Foucart, A., Martin, C., Moreno, E., & Costa, A. (2014). Can Bilinguals See It Coming? Word Anticipation in L2 Sentence Reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*. <https://doi.org/10.1037/a0036756>
- Freunberger, D., & Roehm, D. (2016). Semantic prediction in language comprehension: Evidence from brain potentials. *Language, Cognition and Neuroscience*, *31*(9), 1193–1205. <https://doi.org/10.1080/23273798.2016.1205202>
- Freunberger, D., & Roehm, D. (2017). The costs of being certain: Brain potential evidence for linguistic preactivation in sentence processing. *Psychophysiology*, *54*(6), 824–832. <https://doi.org/10.1111/psyp.12848>
- Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, *6*(2), 78–84.

- Friederici, A. D. (2005). Neurophysiological markers of early language acquisition: From syllables to sentences. *Trends in Cognitive Sciences*, 9(10), 481–488. <https://doi.org/10.1016/j.tics.2005.08.008>
- Friedrich, M., & Friederici, A. D. (2006). Early N400 development and later language acquisition. *Psychophysiology*, 43(1), 1–12. <https://doi.org/10.1111/j.1469-8986.2006.00381.x>
- Frisson, S., Harvey, D., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language*, 95, 200–214. <https://doi.org/10.1016/j.jml.2017.04.007>
- Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456), 815–836. <https://doi.org/10.1098/rstb.2005.1622>
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), Article 2. <https://doi.org/10.1038/nrn2787>
- Fromont, L. A., Steinhauer, K., & Royle, P. (2020). Verbing nouns and nouning verbs: Using a balanced design provides ERP evidence against “syntax-first” approaches to sentence processing. *PLOS ONE*, 15(3), e0229169. <https://doi.org/10.1371/journal.pone.0229169>
- Gambi, C., Gorrie, F., Pickering, M. J., & Rabagliati, H. (2018). The development of linguistic prediction: Predictions of sound and meaning in 2- to 5-year-olds. *Journal of Experimental Child Psychology*, 173, 351–370. <https://doi.org/10.1016/j.jecp.2018.04.012>
- Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37, 58–93. <https://doi.org/10.1006/jmla.1997.2512>
- Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
- Gibson, E. J., Osser, H., & Pick, A. D. (1963). A Study of the Development of Grapheme-Phoneme Correspondences. *Journal of Verbal Learning and Verbal Behavior*, 2(2), 142–146.
- Goldin-Meadow, S., & Brentari, D. (2017). Gesture, sign and language: The coming of age of sign language and gesture studies. *The Behavioral and Brain Sciences*, 40, e46. <https://doi.org/10.1017/S0140525X15001247>
- Goldin-Meadow, S., & Mayberry, R. I. (2001). How do profoundly deaf children learn to read? *Learning Disabilities Research & Practice*, 16, 222–229. <https://doi.org/10.1111/0938-8982.00022>

- Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-Five Years Using the Intermodal Preferential Looking Paradigm to Study Language Acquisition: What Have We Learned? *Perspectives on Psychological Science*, 8(3), 316–339. <https://doi.org/10.1177/1745691613484936>
- Gollan, T. H., & Acenas, L.-A. R. (2004). What Is a TOT? Cognate and Translation Effects on Tip-of-the-Tongue States in Spanish-English and Tagalog-English Bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 246–269. <https://doi.org/10.1037/0278-7393.30.1.246>
- Goodwin, A., Fein, D., & Naigles, L. R. (2012). Comprehension of Wh-Questions Precedes Their Production in Typical Development and Autism Spectrum Disorders. *Autism Research*, 5(2), 109–123. <https://doi.org/10.1002/aur.1220>
- Grainger, J., & Beauvillain, C. (1987). Language blocking and lexical access in bilinguals. *Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology - QUART J EXP PSYCH A-HUM EXP P*, 39, 295–319. <https://doi.org/10.1080/14640748708401788>
- Grainger, J., & Dijkstra, T. (1992). On the representation and use of language information in bilinguals. In *Cognitive processing in bilinguals* (pp. 207–220). North-Holland. [https://doi.org/10.1016/S0166-4115\(08\)61496-X](https://doi.org/10.1016/S0166-4115(08)61496-X)
- Grainger, J., & Holcomb, P. J. (2009). Watching the word go by: On the time-course of component processes in visual word recognition. *Language and Linguistics Compass*, 3(1), 128–156.
- Grainger, J., Kiyonaga, K., & Holcomb, P. J. (2006). The Time Course of Orthographic and Phonological Code Activation. *Psychological Science*, 17, 1021–1026. <https://doi.org/10.1111/j.1467-9280.2006.01821.x>
- Green, D. W. (1998). Bilingualism and Thought. *Psychologica Belgica*, 38(4), 251–276.
- Grey, S., Schubel, L. C., Mcqueen, J. M., & Van Hell, J. G. (2019). Processing foreign-accented speech in a second language: Evidence from ERPs during sentence comprehension in bilinguals. *Bilingualism: Language and Cognition*, 22(5), 912–929. <https://doi.org/10.1017/S1366728918000937>
- Grey, S., & van Hell, J. G. (2017). Foreign-accented speaker identity affects neural correlates of language comprehension. *Journal of Neurolinguistics*, 42, 93–108. <https://doi.org/10.1016/j.jneuroling.2016.12.001>
- Groppe, D. M., Urbach, T. P., & Kutas, M. (2011a). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, 48(12), 1711–1725. <https://doi.org/10.1111/j.1469-8986.2011.01273.x>

- Groppe, D. M., Urbach, T. P., & Kutas, M. (2011b). Mass univariate analysis of event-related brain potentials/fields II: Simulation studies. *Psychophysiology*, *48*(12), 1726–1737. <https://doi.org/10.1111/j.1469-8986.2011.01272.x>
- Grosjean, F. (2001). *The bilingual's language modes. One mind, two languages: Bilingual language processing*, ed. By Janet Nicol, 31–55.
- Grosvald, M., Gutierrez, E., Hafer, S., & Corina, D. (2012). Dissociating linguistic and non-linguistic gesture processing: Electrophysiological evidence from American Sign Language. *Brain and Language*, *121*(1), 12–24. <https://doi.org/10.1016/j.bandl.2012.01.005>
- Grushkin, D. A. (2017). Writing Signed Languages: What For? What Form? *American Annals of the Deaf*, *161*(5), 509–527. <https://doi.org/10.1353/aad.2017.0001>
- Guajardo, L. F., & Wicha, N. Y. Y. (2014). Morphosyntax can modulate the N400 component: Event related potentials to gender-marked post-nominal adjectives. *NeuroImage*, *91*, 262–272. <https://doi.org/10.1016/j.neuroimage.2013.09.077>
- Gumperz, J. J. (1982). *Discourse Strategies*. Cambridge University Press.
- Gunter, T., Friederici, A., & Schriefers, H. (2000). Syntactic Gender and Semantic Expectancy: ERPs Reveal Early Autonomy and Late Interaction. *Journal of Cognitive Neuroscience*, *v. 12*, 556–568 (2000), *12*. <https://doi.org/10.1162/089892900562336>
- Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun—An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, *47*(4), 930–944. <https://doi.org/10.3758/s13428-014-0529-0>
- Gutierrez, E., Williams, D., Grosvald, M., & Corina, D. (2012). Lexical access in American Sign Language: An ERP investigation of effects of semantics and phonology. *Brain Research*, *1468*, 63–83. <https://doi.org/10.1016/j.brainres.2012.04.029>
- Hagoort, P. (1993). Impairments of Lexical-Semantic Processing in Aphasia: Evidence from the Processing of Lexical Ambiguities. *Brain and Language*, *45*(2), 189–232. <https://doi.org/10.1006/brln.1993.1043>
- Hagoort, P., & Brown, C. M. (1999). Gender Electrified: ERP Evidence on the Syntactic Nature of Gender Processing. *Journal of Psycholinguistic Research*, *28*(6), 715–728. <https://doi.org/10.1023/A:1023277213129>
- Hahne, A., & Friederici, A. D. (1999). Electrophysiological Evidence for Two Steps in Syntactic Analysis: Early Automatic and Late Controlled Processes. *Journal of Cognitive Neuroscience*, *11*(2), 194–205. <https://doi.org/10.1162/089892999563328>
- Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. NAACL 2001. <https://aclanthology.org/N01-1021>

- Hartsuiker, R. J., Pickering, M., & Veltkamp, E. (2004). Is Syntax Separate or Shared Between Languages?: Cross-Linguistic Syntactic Priming in Spanish-English Bilinguals. *Psychological Science, 15*(6), 409–414. <https://doi.org/10.1111/j.0956-7976.2004.00693.x>
- Hasson, U., & Egidi, G. (2015). *Cognitive Neuroscience of Natural Language Use: What are naturalistic comprehension paradigms teaching us about language?* 228–255. <https://doi.org/10.1017/cbo9781107323667.011>
- Hasting, A., & Kotz, S. (2008). Speeding Up Syntax: On the Relative Timing and Automaticity of Local Phrase Structure and Morphosyntactic Processing as Reflected in Event-related Brain Potentials. *Journal of Cognitive Neuroscience, 20*, 1207–1219. <https://doi.org/10.1162/jocn.2008.20083>
- Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences, 119*(32), e2201968119. <https://doi.org/10.1073/pnas.2201968119>
- Heller, M. (2007). *Bilingualism: A social approach*. Springer.
- Henderson, L. M., Baseler, H. A., Clarke, P. J., Watson, S., & Snowling, M. J. (2011). The N400 effect in children: Relationships with comprehension, vocabulary and decoding. *Brain and Language, 117*(2), 88–99. <https://doi.org/10.1016/j.bandl.2010.12.003>
- Heredia, R. R., & Altarriba, J. (2001). Bilingual language mixing: Why do bilinguals code-switch? *Current Directions in Psychological Science, 10*(5), 164–168.
- Hickok, G., Bellugi, U., & Klima, E. S. (2001). SIGN language in the BRAIN. *Scientific American, 284*(6), 58–65.
- Hoffmeister, R. J., & Caldwell-Harris, C. L. (2014). Acquiring English as a second language via print: The task for deaf children. *Cognition, 132*(2), 229–242. <https://doi.org/10.1016/j.cognition.2014.03.014>
- Holcomb, P. J., & Grainger, J. (2006). On the time course of visual word recognition: An event-related potential investigation using masked repetition priming. *Journal of Cognitive Neuroscience, 18*(10), 1631–1643. <https://doi.org/10.1162/jocn.2006.18.10.1631>
- Holcomb, P. J., & Neville, H. J. (1991). Natural speech processing: An analysis using event-related brain potentials. *Psychobiology, 19*(4), 286–300. <https://doi.org/10.3758/BF03332082>
- Holmes, V. M., Stowe, L., & Cupples, L. (1989). Lexical expectations in parsing complement-verb sentences. *Journal of Memory and Language, 28*(6), 668–689. [https://doi.org/10.1016/0749-596X\(89\)90003-X](https://doi.org/10.1016/0749-596X(89)90003-X)

- Hoversten, L. J., & Traxler, M. J. (2020). Zooming in on zooming out: Partial selectivity and dynamic tuning of bilingual language control during reading. *Cognition*, *195*, 104118. <https://doi.org/10.1016/j.cognition.2019.104118>
- Hubbard, R. J., & Federmeier, K. D. (2021). Representational Pattern Similarity of Electrical Brain Activity Reveals Rapid and Specific Prediction during Language Comprehension. *Cerebral Cortex (New York, N.Y.: 1991)*, *31*(9), 4300–4313. <https://doi.org/10.1093/cercor/bhab087>
- Huetting, F., & Altmann, G. T. M. (2004). *The On-Line Processing of Ambiguous and Unambiguous Words in Context: Evidence from Head-Mounted Eyetracking*.
- Huetting, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, *96*(1), B23–B32. <https://doi.org/10.1016/j.cognition.2004.10.003>
- Huetting, F., & Altmann, G. T. M. (2011). Looking at anything that is green when hearing “frog”: How object surface colour and stored object colour knowledge influence language-mediated overt attention. *Quarterly Journal of Experimental Psychology*, *64*(1), 122–145. <https://doi.org/10.1080/17470218.2010.481474>
- Huetting, F., Audring, J., & Jackendoff, R. (2022). A parallel architecture perspective on pre-activation and prediction in language processing. *Cognition*, *224*, 105050. <https://doi.org/10.1016/j.cognition.2022.105050>
- Huetting, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, *31*(1), 19–31. <https://doi.org/10.1080/23273798.2015.1072223>
- Huetting, F., & Pickering, M. (2019). Literacy advantages beyond reading: Prediction of spoken language. *Trends in Cognitive Sciences*, *23*, 464–475. <https://doi.org/10.1016/j.tics.2019.03.008>
- Huetting, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151–171. <https://doi.org/10.1016/j.actpsy.2010.11.003>
- Husband, E. M. (2022). Prediction in the maze: Evidence for probabilistic pre-activation from the English a/an contrast. *Glossa Psycholinguistics*, *1*(1). <https://doi.org/10.5070/G601153>
- Hut, S. C. A., & Leminen, A. (2017). Shaving Bridges and Tuning Kitaraa: The Effect of Language Switching on Semantic Processing. *Frontiers in Psychology*, *8*. <https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01438>
- Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1), 101–144. <https://doi.org/10.1016/j.cognition.2002.06.001>

- Ito, A. (2016). *Prediction during native and non-native language comprehension: The role of mediating factors.*
- Ito, A. (2019). Prediction of orthographic information during listening comprehension: A printed-word visual world study. *Quarterly Journal of Experimental Psychology*, 72(11), 2584–2596. <https://doi.org/10.1177/1747021819851394>
- 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. <https://doi.org/10.1016/j.jml.2015.10.007>
- Ito, A., Gambi, C., Pickering, M. J., Fuellenbach, K., & Husband, E. M. (2020). Prediction of phonological and gender information: An event-related potential study in Italian. *Neuropsychologia*, 136, 107291. <https://doi.org/10.1016/j.neuropsychologia.2019.107291>
- Ito, A., Martin, A. E., & Nieuwland, M. S. (2017a). Why the A/AN prediction effect may be hard to replicate: A rebuttal to DeLong, Urbach & Kutas (2017). *Language, Cognition and Neuroscience*, 32(8), 974–983. <https://doi.org/10.1080/23273798.2017.1323112>
- Ito, A., Martin, A. E., & Nieuwland, M. S. (2017b). On predicting form and meaning in a second language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(4), 635–652. <https://doi.org/10.1037/xlm0000315>
- Ito, A., Martin, A. E., & Nieuwland, M. S. (2017c). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience*, 32(8), 954–965. <https://doi.org/10.1080/23273798.2016.1242761>
- Ito, A., & Pickering, M. (2021). *Automaticity and prediction in non-native language comprehension.* <https://doi.org/10.1075/bpa.12.02ito>
- Ito, A., Pickering, M. J., & Corley, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language*, 98, 1–11. <https://doi.org/10.1016/j.jml.2017.09.002>
- Jordan, T., & Thomas, S. (2002). In Search of Perceptual Influences of Sentence Context on Word Recognition. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 28, 34–45. <https://doi.org/10.1037//0278-7393.28.1.34>
- Juottonen, K., Revonsuo, A., & Lang, H. (1996). Dissimilar age influences on two ERP waveforms (LPC and N400) reflecting semantic context effect. *Cognitive Brain Research*, 4(2), 99–107. [https://doi.org/10.1016/0926-6410\(96\)00022-5](https://doi.org/10.1016/0926-6410(96)00022-5)
- Jyotishi, M., Fein, D., & Naigles, L. (2017). Investigating the Grammatical and Pragmatic Origins of Wh-Questions in Children with Autism Spectrum Disorders. *Frontiers in Psychology*, 8. <https://www.frontiersin.org/articles/10.3389/fpsyg.2017.00319>

- 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. <https://doi.org/10.1080/016909600386084>
- Kaan, E., Kheder, S., Kreidler, A., Tomić, A., & Valdés Kroff, J. R. (2020). Processing Code-Switches in the Presence of Others: An ERP Study. *Frontiers in Psychology, 11*. <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01288>
- Kail, R. (1991). Developmental change in speed of processing during childhood and adolescence. *Psychological Bulletin, 109*, 490–501. <https://doi.org/10.1037/0033-2909.109.3.490>
- Kail, R., & Salthouse, T. A. (1994). Processing speed as a mental capacity. *Acta Psychologica, 86*(2), 199–225. [https://doi.org/10.1016/0001-6918\(94\)90003-5](https://doi.org/10.1016/0001-6918(94)90003-5)
- Kail, R. V., & Ferrer, E. (2007). Processing speed in childhood and adolescence: Longitudinal models for examining developmental change. *Child Development, 78*, 1760–1770. <https://doi.org/10.1111/j.1467-8624.2007.01088.x>
- Kambe, G., Rayner, K., & Duffy, S. A. (2001). Global context effects on processing lexically ambiguous words: Evidence from eye fixations. *Memory & Cognition, 29*, 363–372. <https://doi.org/10.3758/BF03194931>
- Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye-movements. *Journal of Memory and Language, 49*(1), 133–156. [https://doi.org/10.1016/s0749-596x\(03\)00023-8](https://doi.org/10.1016/s0749-596x(03)00023-8)
- Kappenman, E. S., & Luck, S. J. (2011). ERP Components: The Ups and Downs of Brainwave Recordings. In E. S. Kappenman & S. J. Luck (Eds.), *The Oxford Handbook of Event-Related Potential Components* (p. 0). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780195374148.013.0014>
- Kaushanskaya, M., Blumenfeld, H. K., & Marian, V. (2020). The Language Experience and Proficiency Questionnaire (LEAP-Q): Ten years later. *Bilingualism (Cambridge, England), 23*(5), 945–950. <https://doi.org/10.1017/s1366728919000038>
- Keane, J., & Brentari, D. (2015). Fingerspelling*: Beyond Handshape Sequences. In M. Marschark & P. E. Spencer (Eds.), *The Oxford Handbook of Deaf Studies in Language* (p. 0). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190241414.013.10>
- Khamis-Dakwar, R., & Froud, K. (2007). Lexical processing in two language varieties: An event-related brain potential study of Arabic native speakers. In M. A. Mughazy (Ed.), *Perspectives on Arabic Linguistics: Papers from the annual symposium on Arabic linguistics. Volume XX: Kalamazoo, Michigan, March 2006* (pp. 153–166). John Benjamins Publishing Company. <https://doi.org/10.1075/cilt.290.13kha>

- Kidd, C., White, K. S., & Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science*, *14*(4), 925–934. <https://doi.org/10.1111/j.1467-7687.2011.01049.x>
- Kim, A., & Lai, V. (2012). Rapid Interactions between Lexical Semantic and Word Form Analysis during Word Recognition in Context: Evidence from ERPs. *Journal of Cognitive Neuroscience*, *24*(5), 1104–1112. https://doi.org/10.1162/jocn_a_00148
- Kim, A., & Sikos, L. (2011). Conflict and surrender during sentence processing: An ERP study of syntax-semantics interaction. *Brain and Language*, *118*(1), 15–22. <https://doi.org/10.1016/j.bandl.2011.03.002>
- King, J. W., & Kutas, M. (1995). Who did what and when? Using word-and clause-level ERPs to monitor working memory usage in reading. *Journal of Cognitive Neuroscience*, *7*(3), 376–395.
- Kluender, R., & Kutas, M. (1993). Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience*, *5*(2), 196–214.
- Kochari, A. R., & Flecken, M. (2019). Lexical prediction in language comprehension: A replication study of grammatical gender effects in Dutch. *Language, Cognition and Neuroscience*, *34*(2), 239–253. <https://doi.org/10.1080/23273798.2018.1524500>
- Kolk, H., & Chwilla, D. (2007). Late positivities in unusual situations. *Brain and Language*, *100*(3), 257–261. <https://doi.org/10.1016/j.bandl.2006.07.006>
- Kotz, S. A., & Elston-Güttler, K. (2004). The role of proficiency on processing categorical and associative information in the L2 as revealed by reaction times and event-related brain potentials. *Journal of Neurolinguistics*, *17*(2), 215–235. [https://doi.org/10.1016/S0911-6044\(03\)00058-7](https://doi.org/10.1016/S0911-6044(03)00058-7)
- Kroll, J. F., Dussias, P. E., Bogulski, C. A., & Kroff, J. R. V. (2012). Chapter Seven - Juggling Two Languages in One Mind: What Bilinguals Tell Us About Language Processing and its Consequences for Cognition. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 56, pp. 229–262). Academic Press. <https://doi.org/10.1016/B978-0-12-394393-4.00007-8>
- Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*(6), 2528–2543. <https://doi.org/10.3758/s13428-021-01546-0>
- Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23–49. <https://doi.org/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. <https://doi.org/10.1080/23273798.2015.1130233>

- Kuperberg, G. R. (2021). Tea With Milk? A Hierarchical Generative Framework of Sequential Event Comprehension. *Topics in Cognitive Science*, 13(1), 256–298. <https://doi.org/10.1111/tops.12518>
- Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A Tale of Two Positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, 32(1), 12–35. https://doi.org/10.1162/jocn_a_01465
- Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension. *Language, Cognition and Neuroscience*, 31(1), 32–59. <https://doi.org/10.1080/23273798.2015.1102299>
- Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and Cognitive Processes*, 8(4), 533–572. <https://doi.org/10.1080/01690969308407587>
- Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A Look around at What Lies Ahead: Prediction and Predictability in Language Processing. In *Predictions in the Brain*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195395518.003.0065>
- Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, 4(12), 463–470. [https://doi.org/10.1016/S1364-6613\(00\)01560-6](https://doi.org/10.1016/S1364-6613(00)01560-6)
- Kutas, M., & Federmeier, K. D. (2009). N400. *Scholarpedia*, 4(10), 7790.
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62, 621–647. <https://doi.org/10.1146/annurev.psych.093008.131123>
- Kutas, M., Federmeier, K. D., & Urbach, T. P. (2014). The “negatives” and “positives” of prediction in language. In *The cognitive neurosciences, 5th ed* (pp. 649–656). MIT Press.
- Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology*, 11(2), 99–116. [https://doi.org/10.1016/0301-0511\(80\)90046-0](https://doi.org/10.1016/0301-0511(80)90046-0)
- Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947), 161–163. <https://doi.org/10.1038/307161a0>
- Kutas, M., Lindamood, T. E., & Hillyard, S. A. (1984). Word Expectancy and Event-Related Brain Potentials During Sentence Processing. *Preparatory States & Processes*, 217–237. <https://doi.org/10.4324/9781315792385-11>
- Kutas, M., Moreno, E., & Wicha, N. Y. Y. (2009). Code-switching and the brain. In *B.E. Bullock, A.J. Toribio (Eds.), The Cambridge handbook of linguistic code-switching*. Cambridge University Press.

- Kutas, M., Neville, H. J., & Holcomb, P. J. (1987). A preliminary comparison of the N400 response to semantic anomalies during reading, listening and signing. *Electroencephalography and Clinical Neurophysiology. Supplement*, 39, 325–330.
- Kutas, M., & Van Petten, C. K. (1994). Psycholinguistics electrified: Event-related brain potential investigations. In *Handbook of psycholinguistics* (pp. 83–143). Academic Press.
- Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Psycholinguistics Electrified II (1994–2005). In *Handbook of Psycholinguistics* (pp. 659–724). Elsevier. <https://doi.org/10.1016/B978-012369374-7/50018-3>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **ImerTest** Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13). <https://doi.org/10.18637/jss.v082.i13>
- Lane, H., Boyes-Braem, P., & Bellugi, U. (1976). Preliminaries to a distinctive feature analysis of handshapes in American Sign Language. *Cognitive Psychology*, 8(2), 263–289. [https://doi.org/10.1016/0010-0285\(76\)90027-X](https://doi.org/10.1016/0010-0285(76)90027-X)
- Laszlo, S., & Federmeier, K. D. (2009). A Beautiful Day in the Neighborhood: An Event-Related Potential Study of Lexical Relationships and Prediction in Context. *Journal of Memory and Language*, 61(3), 326–338. <https://doi.org/10.1016/j.jml.2009.06.004>
- Lau, E., Almeida, D., Hines, P. C., & Poeppel, D. (2009). A lexical basis for N400 context effects: Evidence from MEG. *Brain and Language*, 111(3), 161–172. <https://doi.org/10.1016/j.bandl.2009.08.007>
- Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). DISSOCIATING N400 EFFECTS OF PREDICTION FROM ASSOCIATION IN SINGLE WORD CONTEXTS. *Journal of Cognitive Neuroscience*, 25(3), 484–502. https://doi.org/10.1162/jocn_a_00328
- Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience*, 9(12), Article 12. <https://doi.org/10.1038/nrn2532>
- Lau, E., Stroud, C., Plesch, S., & Phillips, C. (2006). The role of structural prediction in rapid syntactic analysis. *Brain and Language*, 98, 74–88. <https://doi.org/10.1016/j.bandl.2006.02.003>
- Lederberg, A., Schick, B., & Spencer, P. (2012). Language and Literacy Development of Deaf and Hard-of-Hearing Children: Successes and Challenges. *Developmental Psychology*, 49. <https://doi.org/10.1037/a0029558>
- Lee, C., & Federmeier, K. D. (2006). To mind the mind: An event-related potential study of word class and semantic ambiguity. *Brain Research*, 1081(1), 191–202. <https://doi.org/10.1016/j.brainres.2006.01.058>

- Lee, C., & Federmeier, K. D. (2009). Wave-ering: An ERP study of syntactic and semantic context effects on ambiguity resolution for noun/verb homographs. *Journal of Memory and Language*, *61*(4), 538–555. <https://doi.org/10.1016/j.jml.2009.08.003>
- Lee, C., & Federmeier, K. D. (2012). Ambiguity's aftermath: How age differences in resolving lexical ambiguity affect subsequent comprehension. *Neuropsychologia*, *50*(5), 869–879. <https://doi.org/10.1016/j.neuropsychologia.2012.01.027>
- Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2021). *Emmeans: Estimated marginal means, aka least-squares means. R Package Version 1 (2018)*.
- Levari, T., & Snedeker, J. (2018). *Children's N400 is sensitive to both predictability and frequency: Evidence from natural listening*. [Conference Presentation]. Forty-second annual Boston University Conference on Language Development, Boston, MA, United States. <https://www.bu.edu/buclcd/conference-info/browse-abstracts-2018/2018-friday-session-b-0900/>
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.
- Lewendon, J., Mortimore, L., & Egan, C. (2020). The Phonological Mapping (Mismatch) Negativity: History, Inconsistency, and Future Direction. *Frontiers in Psychology*, *11*. <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01967>
- Lew-Williams, C., & Fernald, A. (2007). Young Children Learning Spanish Make Rapid Use of Grammatical Gender in Spoken Word Recognition. *Psychological Science*, *18*(3), 193–198. <https://doi.org/10.1111/j.1467-9280.2007.01871.x>
- Li, J., Bhattasali, S., Zhang, S., Franzluebbers, B., Luh, W., Spreng, R., Brennan, J., Yang, Y., Pallier, C., & Hale, J. (2021). *Le Petit Prince: A multilingual fMRI corpus using ecological stimuli*. <https://doi.org/10.1101/2021.10.02.462875>
- Li, P. (1996). Spoken Word Recognition of Code-Switched Words by Chinese–English Bilinguals. *Journal of Memory and Language*, *35*(6), 757–774. <https://doi.org/10.1006/jmla.1996.0039>
- Li, X., Li, X., & Qu, Q. (2022). Predicting phonology in language comprehension: Evidence from the visual world eye-tracking task in Mandarin Chinese. *Journal of Experimental Psychology. Human Perception and Performance*, *48*(5), 531–547. <https://doi.org/10.1037/xhp0000999>
- Liao, C.-H., & Chan, S.-H. (2016). Direction matters: Event-related brain potentials reflect extra processing costs in switching from the dominant to the less dominant language. *Journal of Neurolinguistics*, *40*, 79–97. <https://doi.org/10.1016/j.jneuroling.2016.06.004>
- Lillo-Martin, D. (1986). Two Kinds of Null Arguments in American Sign Language. *Natural Language & Linguistic Theory*, *4*(4), 415–444.

- Litcofsky, K. A., & van Hell, J. (2017). Switching direction affects switching costs: Behavioral, ERP and time-frequency analyses of intra-sentential codeswitching. *Neuropsychologia*, *97*, 112–139. <https://doi.org/10.1016/j.neuropsychologia.2017.02.002>
- Liu, Y., Shu, H., & Wei, J. (2006). Spoken word recognition in context: Evidence from Chinese ERP analyses. *Brain and Language*, *96*(1), 37–48. <https://doi.org/10.1016/j.bandl.2005.08.007>
- Loebell, H., & Bock, K. (2003). Structural priming across languages. *Linguistics*, *41*, 791–824. <https://doi.org/10.1515/ling.2003.026>
- 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. <https://doi.org/10.3389/fnhum.2014.00213>
- Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science*, *42*, 1166–1183. <https://doi.org/10.1111/cogs.12597>
- Luck, S. (2004). Ten simple rules for designing and interpreting erp experiments.[in:] event-related potentials: A methods handbook. *Handy, TC (Ed.)*.
- Luck, S. (2014). *An Introduction to the Event-Related Potential Technique, second edition*. MIT Press.
- Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157. <https://doi.org/10.1111/psyp.12639>
- Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60. <https://doi.org/10.1016/j.cogpsych.2016.06.002>
- MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676–703. <https://doi.org/10.1037/0033-295X.101.4.676>
- Macnamara, J., & Kushnir, S. L. (1971). Linguistic independence of bilinguals: The input switch. *Journal of Verbal Learning and Verbal Behavior*, *10*(5), 480–487. [https://doi.org/10.1016/S0022-5371\(71\)80018-X](https://doi.org/10.1016/S0022-5371(71)80018-X)
- Mani, N., & Huettig, F. (2012). Prediction During Language Processing is a Piece of Cake—But Only for Skilled Producers. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 843–847. <https://doi.org/10.1037/a0029284>
- Mani, N., & Huettig, F. (2014). Word reading skill predicts anticipation of upcoming spoken language input: A study of children developing proficiency in reading. *Journal of Experimental Child Psychology*, *126*, 264–279. <https://doi.org/10.1016/j.jecp.2014.05.004>

- Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and Multilinguals. *Journal of Speech, Language, and Hearing Research, 50*(4), 940–967. [https://doi.org/10.1044/1092-4388\(2007/067\)](https://doi.org/10.1044/1092-4388(2007/067))
- Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition, 6*(2), 97–115. <https://doi.org/10.1017/S1366728903001068>
- Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177–190. <https://doi.org/10.1016/j.jneumeth.2007.03.024>
- Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*(1), 71–102. [https://doi.org/10.1016/0010-0277\(87\)90005-9](https://doi.org/10.1016/0010-0277(87)90005-9)
- Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing Spoken Words: The Importance of Word Onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 576–585. <https://doi.org/10.1037/0096-1523.15.3.576>
- Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is Production: The missing link between language production and comprehension. *Scientific Reports, 8*(1), Article 1. <https://doi.org/10.1038/s41598-018-19499-4>
- Martin, C. D., Thierry, G., Kuipers, J.-R., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language, 69*(4), 574–588. <https://doi.org/10.1016/j.jml.2013.08.001>
- McRae, K., & Matsuki, K. (2009). People Use their Knowledge of Common Events to Understand Language, and Do So as Quickly as Possible: Event Knowledge & Language Comprehension. *Language and Linguistics Compass, 3*(6), 1417–1429. <https://doi.org/10.1111/j.1749-818X.2009.00174.x>
- Meade, G., Lee, B., Massa, N., Holcomb, P. J., Midgley, K. J., & Emmorey, K. (2022). Are form priming effects phonological or perceptual? Electrophysiological evidence from American Sign Language. *Cognition, 220*, 104979. <https://doi.org/10.1016/j.cognition.2021.104979>
- Meade, G., Lee, B., Midgley, K. J., Holcomb, P. J., & Emmorey, K. (2018). Phonological and semantic priming in American Sign Language: N300 and N400 effects. *Language, Cognition and Neuroscience, 33*(9), 1092–1106. <https://doi.org/10.1080/23273798.2018.1446543>
- Meyer, A. S., Huettig, F., & Levelt, W. J. M. (2016). Same, different, or closely related: What is the relationship between language production and comprehension? *Journal of Memory and Language, 89*, 1–7. <https://doi.org/10.1016/j.jml.2016.03.002>

- Midgley, K. J., Holcomb, P. J., & Grainger, J. (2009). Masked repetition and translation priming in second language learners: A window on the time-course of form and meaning activation using ERPs. *Psychophysiology*, *46*(3), 551–565.
- Milburn, E., Warren, T., & Dickey, M. W. (2016). World knowledge affects prediction as quickly as selectional restrictions: Evidence from the visual world paradigm. *Language, Cognition and Neuroscience*, *31*(4), 536–548.
<https://doi.org/10.1080/23273798.2015.1117117>
- Milroy, L., & Gordon, M. (2008). *Sociolinguistics: Method and interpretation*. John Wiley & Sons.
- Mishra, R. K., Singh, N., Pandey, A., & Huettig, F. (2012). Spoken language-mediated anticipatory eye-movements are modulated by reading ability—Evidence from Indian low and high literates. *Journal of Eye Movement Research*, *5*(1).
<https://doi.org/10.16910/jemr.5.1.3>
- Molinaro, N., Barber, H. A., Caffarra, S., & Carreiras, M. (2015). On the left anterior negativity (LAN): The case of morphosyntactic agreement: A Reply to Tanner et al. *Cortex*, *66*, 156–159. <https://doi.org/10.1016/j.cortex.2014.06.009>
- Molinaro, N., Barber, H. A., & Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex*, *47*(8), 908–930.
<https://doi.org/10.1016/j.cortex.2011.02.019>
- Monzalvo, K., & Dehaene-Lambertz, G. (2013). How reading acquisition changes children’s spoken language network. *Brain and Language*, *127*(3), 356–365.
<https://doi.org/10.1016/j.bandl.2013.10.009>
- Moreno, E. M., Federmeier, K. D., & Kutas, M. (2002). Switching languages, switching palabras (words): An electrophysiological study of code switching. *Brain and Language*, *80*(2), 188–207.
- Moreno, E. M., Rodríguez-Fornells, A., & Laine, M. (2008). Event-related potentials (ERPs) in the study of bilingual language processing. *Journal of Neurolinguistics*, *21*(6), 477–508.
<https://doi.org/10.1016/j.jneuroling.2008.01.003>
- Morgan-Short, K., & Tanner, D. (2013). Event-related potentials (ERPs). In *Research Methods in Second Language Psycholinguistics* (pp. 127–152). Routledge.
<https://doi.org/10.4324/9780203123430>
- Musselman, C. (2000). How Do Children Who Can’t Hear Learn to Read an Alphabetic Script? A Review of the Literature on Reading and Deafness. *Journal of Deaf Studies and Deaf Education*, *5*(1), 9–31. <https://doi.org/10.1093/deafed/5.1.9>
- Napoli, D. J., & Sutton-Spence, R. (2014). Order of the major constituents in sign languages: Implications for all language. *Frontiers in Psychology*, *5*.
<https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00376>

- Nation, K., & Snowling, M. J. (2004). Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of Research in Reading, 27*(4), 342–356. <https://doi.org/10.1111/j.1467-9817.2004.00238.x>
- Neville, H. J., Coffey, S. A., Lawson, D. S., Fischer, A., Emmorey, K., & Bellugi, U. (1997). Neural Systems Mediating American Sign Language: Effects of Sensory Experience and Age of Acquisition. *Brain and Language, 57*(3), 285–308. <https://doi.org/10.1006/brln.1997.1739>
- Neville, H., Nicol, J. L., Barss, A., Forster, K. I., & Garrett, M. F. (1991). Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of Cognitive Neuroscience, 3*(2), 151–165.
- Nevins, A., Dillon, B., Malhotra, S., & Phillips, C. (2007). The role of feature-number and feature-type in processing Hindi verb agreement violations. *Brain Research, 1164*, 81–94. <https://doi.org/10.1016/j.brainres.2007.05.058>
- Ng, S., Gonzalez, C., & Wicha, N. Y. (2014). The fox and the cabra: An ERP analysis of reading code switched nouns and verbs in bilingual short stories. *Brain Research, 1557*, 127–140.
- Ng, S., Payne, B. R., Stine-Morrow, E. A. L., & Federmeier, K. D. (2018). How struggling adult readers use contextual information when comprehending speech: Evidence from event-related potentials. *International Journal of Psychophysiology, 125*, 1–9. <https://doi.org/10.1016/j.ijpsycho.2018.01.013>
- Nicenboim, B., Vasisht, S., & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia, 142*, 107427. <https://doi.org/10.1016/j.neuropsychologia.2020.107427>
- Nieuwland, M. S. (2019). Do ‘early’ brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience & Biobehavioral Reviews, 96*, 367–400. <https://doi.org/10.1016/j.neubiorev.2018.11.019>
- Nieuwland, M. S., Martin, A. E., & Carreiras, M. (2013). Event-related brain potential evidence for animacy processing asymmetries during sentence comprehension. *Brain and Language, 126*(2), 151–158. <https://doi.org/10.1016/j.bandl.2013.04.005>
- Nieuwland, M. S., Otten, M., & Van Berkum, J. J. A. (2007). Who are You Talking About? Tracking Discourse-level Referential Processing with Event-related Brain Potentials. *Journal of Cognitive Neuroscience, 19*(2), 228–236. <https://doi.org/10.1162/jocn.2007.19.2.228>
- Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsturn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., ... Huettig, F. (2018). Large-scale replication study reveals a limit on

- probabilistic prediction in language comprehension. *ELife*, 7, e33468.
<https://doi.org/10.7554/eLife.33468>
- Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When Peanuts Fall in Love: N400 Evidence for the Power of Discourse. *Journal of Cognitive Neuroscience*, 18(7), 1098–1111.
<https://doi.org/10.1162/jocn.2006.18.7.1098>
- Nour Eddine, S., Brothers, T., & Kuperberg, G. R. (2022). The N400 in silico: A review of computational models. In *Psychology of Learning and Motivation* (Vol. 76, pp. 123–206). Elsevier. <https://doi.org/10.1016/bs.plm.2022.03.005>
- Ochshorn, R. M., & Hawkins, M. (2017). *Gentle: A robust yet lenient forced aligner built on Kaldi*. <https://lowerquality.com/gentle/>
- Olson, D. R. (1993). How writing represents speech. *Language & Communication*, 13(1), 1–17.
[https://doi.org/10.1016/0271-5309\(93\)90017-H](https://doi.org/10.1016/0271-5309(93)90017-H)
- O'Rourke, P. L., & Van Petten, C. (2011). Morphological agreement at a distance: Dissociation between early and late components of the event-related brain potential. *Brain Research*, 1392, 62–79. <https://doi.org/10.1016/j.brainres.2011.03.071>
- Osterhout, L. (1997). On the brain response to syntactic anomalies: Manipulations of word position and word class reveal individual differences. *Brain and Language*, 59, 494–522.
<https://doi.org/10.1006/brln.1997.1793>
- Osterhout, L., Allen, M. D., Mclaughlin, J., & Inoue, K. (2002). Brain potentials elicited by prose-embedded linguistic anomalies. *Memory & Cognition*, 30(8), 1304–1312.
<https://doi.org/10.3758/BF03213412>
- Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31, 785–806. [https://doi.org/10.1016/0749-596X\(92\)90039-Z](https://doi.org/10.1016/0749-596X(92)90039-Z)
- Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 786–803.
<https://doi.org/10.1037/0278-7393.20.4.786>
- Osterhout, L., Mclaughlin, J., Kim, A., Greenwald, R., & Introduction. (2004). Sentences in the Brain: Event-Related Potentials as Real-Time Reflections of Sentence Comprehension and Language Learning. *The Online Study of Sentence Comprehension : Eyetracking, ERP, and Beyond*.
- Osterhout, L., & Mobley, L. A. (1995). Event-Related Brain Potentials Elicited by Failure to Agree. *Journal of Memory and Language*, 34(6), 739–773.
<https://doi.org/10.1006/jmla.1995.1033>

- Otten, M., Nieuwland, M. S., & Van Berkum, J. J. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience*, 8(1), 89. <https://doi.org/10.1186/1471-2202-8-89>
- Otten, M., & Van Berkum, J. J. A. (2008). Discourse-Based Word Anticipation During Language Processing: Prediction or Priming? *Discourse Processes*, 45(6), 464–496. <https://doi.org/10.1080/01638530802356463>
- Otten, M., & Van Berkum, J. J. A. (2009). Does working memory capacity affect the ability to predict upcoming words in discourse? *Brain Research*, 1291, 92–101. <https://doi.org/10.1016/j.brainres.2009.07.042>
- Padden, C. A. (1998). The ASL lexicon. *Sign Language & Linguistics*, 1(1), 39–60. <https://doi.org/10.1075/sll.1.1.04pad>
- 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. <https://doi.org/10.1111/psyp.12515>
- Payne, B. R., Ng, S., Shantz, K., & Federmeier, K. D. (2020). Chapter Four - Event-related brain potentials in multilingual language processing: The N's and P's. In K. D. Federmeier & H.-W. Huang (Eds.), *Psychology of Learning and Motivation* (Vol. 72, pp. 75–118). Academic Press. <https://doi.org/10.1016/bs.plm.2020.03.003>
- Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time Course of Word Identification and Semantic Integration in Spoken Language. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 25, 394–417. <https://doi.org/10.1037/0278-7393.25.2.394>
- Pickering, M., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, 144(10), 1002–1044. <https://doi.org/10.1037/bul0000158>
- Pickering, M., & Garrod, S. (2013). How tightly are production and comprehension interwoven? *Frontiers in Psychology*, 4. <https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00238>
- Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, 144(10), 1002–1044. <https://doi.org/10.1037/bul0000158>
- Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11(3), 105–110. <https://doi.org/10.1016/j.tics.2006.12.002>

- Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4), 329–347. <https://doi.org/10.1017/S0140525X12001495>
- Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, 198, 181–197. <https://doi.org/10.1016/j.neuroimage.2019.05.026>
- Poplack, P. (1980). Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPANOL: Toward a typology of Code-switching. *Linguistics*, 18, 581–618.
- Poplack, S. (2018). When the quest for symmetry meets inherent variability: Categories of grammar and categories of speech. In N. Shin & D. Erker (Eds.), *Questioning Theoretical Primitives in Linguistic Inquiry: Papers in honor of Ricardo Otheguy* (pp. 7–34). John Benjamins Publishing Company. <https://doi.org/10.1075/sfsl.76.02pop>
- Poullisse, N., & Bongaerts, T. (1994). First Language Use in Second Language Production. *Applied Linguistics*, 15(1), 36–57.
- Poulton, V. R., & Nieuwland, M. S. (2022). Can You Hear What's Coming? Failure to Replicate ERP Evidence for Phonological Prediction. *Neurobiology of Language*, 3(4), 556–574. https://doi.org/10.1162/nol_a_00078
- Proverbio, A., Cok, B., & Zani, A. (2002). Electrophysiological Measures of Language Processing in Bilinguals. *Journal of Cognitive Neuroscience*, 14, 994–1017. <https://doi.org/10.1162/089892902320474463>
- Proverbio, A., Leoni, G., & Zani, A. (2004). Language switching mechanisms in simultaneous interpreters: An ERP study. *Neuropsychologia*, 42(12), 1636–1656.
- R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rabagliati, H., Gambi, C., & Pickering, M. J. (2016). Learning to predict or predicting to learn? *Language, Cognition and Neuroscience*, 31(1), 94–105. <https://doi.org/10.1080/23273798.2015.1077979>
- Ratner, N. (2013). Why Talk with Children Matters: Clinical Implications of Infant- and Child-Directed Speech Research. *Seminars in Speech and Language*, 34(04), 203–214. <https://doi.org/10.1055/s-0033-1353449>
- Rayner, K., Binder, K. S., Ashby, J., & Pollatsek, A. (2001). Eye movement control in reading: Word predictability has little influence on initial landing positions in words. *Vision Research*, 41, 943–954. [https://doi.org/10.1016/S0042-6989\(00\)00310-2](https://doi.org/10.1016/S0042-6989(00)00310-2)
- Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How Psychological Science Informs the Teaching of Reading. *Psychological Science in the Public Interest*, 2(2), 31–74. <https://doi.org/10.1111/1529-1006.00004>

- Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, 3(4), 504–509. <https://doi.org/10.3758/BF03214555>
- Roelofs, A., Meyer, A. S., & Levelt, W. J. (1998). A case for the lemma/lexeme distinction in models of speaking: Comment on Caramazza and Miozzo (1997). *Cognition*, 69(2), 219–230. [https://doi.org/10.1016/s0010-0277\(98\)00056-0](https://doi.org/10.1016/s0010-0277(98)00056-0)
- 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. <https://doi.org/10.1075/ml.8.2.05roy>
- Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high-and low-frequency words. *Memory & Cognition*, 18(4), 367–379. <https://doi.org/10.3758/BF03197126>
- Ruigendijk, E., Hentschel, G., & Zeller, J. P. (2016). How L2-learners' brains react to code-switches: An ERP study with Russian learners of German. *Second Language Research*, 32(2), 197–223.
- Rumelhart, D., & McClelland, J. (1982). An interactive activation model of context effects in letter perception: Pt. 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, 89, 60–94. *Psychological Review*, 89, 60–94.
- 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. <https://doi.org/10.1016/j.neuropsychologia.2021.107855>
- Scarborough, D. L., Gerard, L., & Cortese, C. (1984). Independence of lexical access in bilingual word recognition. *Journal of Verbal Learning and Verbal Behavior*, 23(1), 84–99. [https://doi.org/10.1016/S0022-5371\(84\)90519-X](https://doi.org/10.1016/S0022-5371(84)90519-X)
- Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 344–354. <https://doi.org/10.1037/0278-7393.14.2.344>
- Sebba, M., Mahootian, S., & Jonsson, C. (2012). *Language mixing and code-switching in writing: Approaches to mixed-language written discourse*. (M. Sebba, S. Mahootian, & C. Jonsson, Eds.). Routledge. <https://eprints.lancs.ac.uk/id/eprint/54152/>
- Sehr, Z. S., Caselli, N., Cohen-Goldberg, A. M., & Emmorey, K. (2021). The ASL-LEX 2.0 Project: A Database of Lexical and Phonological Properties for 2,723 Signs in American Sign Language. *The Journal of Deaf Studies and Deaf Education*, 26(2), 263–277. <https://doi.org/10.1093/deafed/ena038>

- Seidenberg, M. (2017). *Language at the Speed of Sight: How We Read, Why So Many Can't, and What Can Be Done About It*. Basic Books.
- Seidl, A., Hollich, G., & Jusczyk, P. W. (2003). Early Understanding of Subject and Object Wh-Questions. *Infancy*, 4(3), 423–436. https://doi.org/10.1207/S15327078IN0403_06
- Severens, E., Jansma, B. M., & Hartsuiker, R. J. (2008). Morphophonological influences on the comprehension of subject–verb agreement: An ERP study. *Brain Research*, 1228, 135–144. <https://doi.org/10.1016/j.brainres.2008.05.092>
- Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a Measure of Intersentential Comprehension. *Reading Research Quarterly*, 17(2), 229. <https://doi.org/10.2307/747485>
- Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. (2023). *Afex: Analysis of Factorial Experiments. R package version 1.2-1*. <https://CRAN.R-project.org/package=afex>
- Skarakis-Doyle, E., & Dempsey, L. (2008). Assessing story comprehension in preschool children. *Topics in Language Disorders*, 28(2), 131–148.
- Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302–319. <https://doi.org/10.1016/j.cognition.2013.02.013>
- Soares, C., & Grosjean, F. (1984). Bilinguals in a monolingual and a bilingual speech mode: The effect on lexical access. *Memory & Cognition*, 12, 380–386. <https://doi.org/10.3758/BF03198298>
- Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 32(25), 8443–8453. <https://doi.org/10.1523/JNEUROSCI.5069-11.2012>
- Speer, N. K., Zacks, J. M., & Reynolds, J. R. (2007). Human Brain Activity Time-Locked to Narrative Event Boundaries. *Psychological Science*, 18(5), 449–455. <https://doi.org/10.1111/j.1467-9280.2007.01920.x>
- Spivey, M. J., & Marian, V. (1999). Cross Talk Between Native and Second Languages: Partial Activation of an Irrelevant Lexicon. *Psychological Science*, 10(3), 281–284. <https://doi.org/10.1111/1467-9280.00151>
- Staub, A. (2015). The Effect of Lexical Predictability on Eye Movements in Reading: Critical Review and Theoretical Interpretation. *Language and Linguistics Compass*, 9(8), 311–327. <https://doi.org/10.1111/lnc3.12151>
- Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, 82, 1–17. <https://doi.org/10.1016/j.jml.2015.02.004>

- Steinhauer, K., & Drury, J. E. (2012). On the early left-anterior negativity (ELAN) in syntax studies. *Brain and Language*, *120*(2), 135–162. <https://doi.org/10.1016/j.bandl.2011.07.001>
- Stokoe, W. C. (1980). Sign language structure. *Annual Review of Anthropology*, *9*, 365–470. <https://doi.org/10.1146/annurev.an.09.100180.002053>
- Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology*, *38*(6), 934–947.
- Strong, M., & Prinz, P. M. (1997). A Study of the Relationship Between American Sign Language and English Literacy. *Journal of Deaf Studies and Deaf Education*, *2*(1), 37–46. <https://doi.org/10.1093/oxfordjournals.deafed.a014308>
- Stungis, J. (1981). Identification and discrimination of handshape in American Sign Language. *Perception & Psychophysics*, *29*(3), 261–276. <https://doi.org/10.3758/BF03207293>
- Sussman, R. S., & Sedivy, J. (2003). The time-course of processing syntactic dependencies: Evidence from eye movements. *Language and Cognitive Processes*, *18*(2), 143–163. <https://doi.org/10.1080/01690960143000498>
- Swaab, T. Y., Ledoux, K., Camblin, C. C., & Boudewyn, M. A. (2012). Language-related ERP components. In *The Oxford handbook of event-related potential components* (pp. 397–439). Oxford University Press.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634. <https://doi.org/10.1126/science.7777863>
- Tanner, D. (2015). On the left anterior negativity (LAN) in electrophysiological studies of morphosyntactic agreement: A Commentary on “Grammatical agreement processing in reading: ERP findings and future directions” by Molinaro et al., 2014. *Cortex*, *66*, 149–155. <https://doi.org/10.1016/j.cortex.2014.04.007>
- Tanner, D., Grey, S., & van Hell, J. G. (2017). Dissociating retrieval interference and reanalysis in the P600 during sentence comprehension. *Psychophysiology*, *54*(2), 248–259. <https://doi.org/10.1111/psyp.12788>
- Tanner, D., & Van Hell, J. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia*, *56*, 289–301. <https://doi.org/10.1016/j.neuropsychologia.2014.02.002>
- Tauroza, S., & Allison, D. (1990). Speech Rates in British English. *Applied Linguistics*, *11*(1), 90–105. <https://doi.org/10.1093/applin/11.1.90>
- Taylor, W. L. (1953). “Cloze procedure”: A new tool for measuring readability. *Journalism Quarterly*, *30*, 415–433.

- Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology*, *83*, 382–392. <https://doi.org/10.1016/j.ijpsycho.2011.12.007>
- Tokowicz, N., & MacWhinney, B. (2005). Implicit and Explicit Measures of Sensitivity to Violations in Second Language Grammar: An Event-Related Potential Investigation. *Studies in Second Language Acquisition*, *27*, 173–204. <https://doi.org/10.1017/S0272263105050102>
- Treiman, R., & Hirsh-Pasek, K. (1983). Silent reading: Insights from second-generation deaf readers. *Cognitive Psychology*, *15*(1), 39–65. [https://doi.org/10.1016/0010-0285\(83\)90003-8](https://doi.org/10.1016/0010-0285(83)90003-8)
- Trueswell, J. C., & Tanenhaus, M. K. (1994). Toward a lexicalist framework of constraint-based syntactic ambiguity resolution. In *Perspectives on sentence processing* (pp. 155–179). Lawrence Erlbaum Associates, Inc.
- Urbach, T. P., DeLong, K. A., Chan, W.-H., & Kutas, M. (2020). An exploratory data analysis of word form prediction during word-by-word reading. *Proceedings of the National Academy of Sciences*, *117*(34), 20483–20494. <https://doi.org/10.1073/pnas.1922028117>
- van Berkum, J. (2004). Sentence comprehension in a wider discourse: Can we use ERPs to keep track of things? *Journal of Neurology - J NEUROL*, 229–270.
- Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*(3), 443–467. <https://doi.org/10.1037/0278-7393.31.3.443>
- 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. <https://doi.org/10.1162/089892999563724>
- Van De Meerendonk, N., Kolk, H. H. J., Chwilla, D. J., & Vissers, C. Th. W. M. (2009). Monitoring in Language Perception. *Language and Linguistics Compass*, *3*(5), 1211–1224. <https://doi.org/10.1111/j.1749-818X.2009.00163.x>
- van de Meerendonk, N., Kolk, H. H. J., Vissers, C. Th. W. M., & Chwilla, D. J. (2010). Monitoring in Language Perception: Mild and Strong Conflicts Elicit Different ERP Patterns. *Journal of Cognitive Neuroscience*, *22*(1), 67–82. <https://doi.org/10.1162/jocn.2008.21170>
- van den Brink, D., Brown, C. M., & Hagoort, P. (2001). Electrophysiological Evidence for Early Contextual Influences during Spoken-Word Recognition: N200 Versus N400 Effects. *Journal of Cognitive Neuroscience*, *13*(7), 967–985. <https://doi.org/10.1162/089892901753165872>

- Van Der Meij, M., Cuetos, F., Carreiras, M., & Barber, H. A. (2011). Electrophysiological correlates of language switching in second language learners. *Psychophysiology*, *48*(1), 44–54.
- van Hell, J., & de Groot, A. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition*, *1*(3), 193–211. <https://doi.org/10.1017/S1366728998000352>
- van Hell, J., & de Groot, A. M. B. (2008). Sentence context modulates visual word recognition and translation in bilinguals. *Acta Psychologica*, *128*(3), 431–451. <https://doi.org/10.1016/j.actpsy.2008.03.010>
- van Hell, J., Fernandez, C. B., Kootstra, G. J., Litcofsky, K. A., & Ting, C. Y. (2018). Electrophysiological and experimental-behavioral approaches to the study of intra-sentential code-switching. *Linguistic Approaches to Bilingualism*, *8*(1), 134–161.
- van Hell, J., Litcofsky, K., & Ting, C. Y. (2015). Intra-sentential code-switching: Cognitive and neural approaches. *The Cambridge Handbook of Bilingual Processing*, 459–482.
- van Hell, J., & Tanner, D. (2012). Second Language Proficiency and Cross-Language Lexical Activation. *Language Learning*, *62*(s2), 148–171. <https://doi.org/10.1111/j.1467-9922.2012.00710.x>
- van Hell, J., & Witteman, M. J. (2009). 3. The neurocognition of switching between languages: A review of electrophysiological studies. In L. Isurin, D. Winford, & K. de Bot (Eds.), *Multidisciplinary Approaches to Code Switching* (pp. 53–84). John Benjamins Publishing Company. <https://doi.org/10.1075/sibil.41.06hel>
- van Herten, M., Kolk, H. H. J., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research*, *22*(2), 241–255. <https://doi.org/10.1016/j.cogbrainres.2004.09.002>
- Van Petten, C. (1993). A comparison of lexical and sentence-level context effects in event-related potentials. *Language and Cognitive Processes*, *8*(4), 485–531. <https://doi.org/10.1080/01690969308407586>
- Van Petten, C. (1995). Words and sentences: Event-related brain potential measures. *Psychophysiology*, *32*(6), 511–525. <https://doi.org/10.1111/j.1469-8986.1995.tb01228.x>
- Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 394–417. <https://doi.org/10.1037/0278-7393.25.2.394>
- Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition*, *18*, 380–393. <https://doi.org/10.3758/BF03197127>

- Van Petten, C., & Kutas, M. (1991). Electrophysiological evidence for the flexibility of lexical processing. In *Understanding word and sentence* (pp. 129–174). North-Holland.
[https://doi.org/10.1016/S0166-4115\(08\)61532-0](https://doi.org/10.1016/S0166-4115(08)61532-0)
- Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, 83(2), 176–190.
<https://doi.org/10.1016/j.ijpsycho.2011.09.015>
- Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes*, 56(3), 229–255. <https://doi.org/10.1080/0163853X.2018.1448677>
- Vissers, C. Th. W. M., Chwilla, D. J., & Kolk, H. H. J. (2006). Monitoring in language perception: The effect of misspellings of words in highly constrained sentences. *Brain Research*, 1106(1), 150–163. <https://doi.org/10.1016/j.brainres.2006.05.012>
- Waite, B., Yacovone, A., & Snedeker, J. (2023, March 10). *Children make robust lexical predictions in a naturalistic context*. 36th Annual Conference on Human Sentence Processing, Pittsburgh, PA.
- Wang, L., Kuperberg, G.R., & Jensen, O. (2018). Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. *ELife*, 7, e39061.
<https://doi.org/10.7554/eLife.39061>
- Wang, L., Wlotko, E. W., Alexander, E. J., Schoot, L., Kim, M., Warnke, L., & Kuperberg, G. R. (2020). Neural Evidence for the Prediction of Animacy Features during Language Comprehension: Evidence from MEG and EEG Representational Similarity Analysis. *The Journal of Neuroscience*, 40(16), 3278–3291. <https://doi.org/10.1523/jneurosci.1733-19.2020>
- Wang, L., Brothers, T., Jensen, O., Kuperberg, G.R. (under revision). Dissociating the pre-activation of word meaning and form during sentence comprehension: Evidence from EEG Representational Similarity Analysis.
- Weber-Fox, C. M., & Neville, H. J. (1996). Maturation Constraints on Functional Specializations for Language Processing: ERP and Behavioral Evidence in Bilingual Speakers. *Journal of Cognitive Neuroscience*, 8(3), 231–256.
<https://doi.org/10.1162/jocn.1996.8.3.231>
- Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. *PLOS ONE*, 9(11), e112575.
<https://doi.org/10.1371/journal.pone.0112575>
- Whitworth, C. (2011). Features and natural classes in ASL handshapes. *Sign Language Studies*, 12, 46–71. <https://doi.org/10.1353/sls.2011.0014>

- Wicha, N. Y. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003). Potato not Pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters*, *346*(3), 165–168. [https://doi.org/10.1016/S0304-3940\(03\)00599-8](https://doi.org/10.1016/S0304-3940(03)00599-8)
- Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2003). Expecting gender: an event related brain potential study on the role of grammatical gender in comprehending a line drawing within a written sentence in Spanish. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *39*(3), 483. [https://doi.org/10.1016/s0010-9452\(08\)70260-0](https://doi.org/10.1016/s0010-9452(08)70260-0)
- Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating Words and Their Gender: An Event-related Brain Potential Study of Semantic Integration, Gender Expectancy, and Gender Agreement in Spanish Sentence Reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288. <https://doi.org/10.1162/0898929041920487>
- Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction During Natural Language Comprehension. *Cerebral Cortex*, *26*(6), 2506–2516. <https://doi.org/10.1093/cercor/bhv075>
- Wlotko, E. W., & Federmeier, K. D. (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex*, *68*, 20–32. <https://doi.org/10.1016/j.cortex.2015.03.014>
- Xu, J., Kemeny, S., Park, G., Frattali, C., & Braun, A. (2005). Language in context: Emergent features of word, sentence, and narrative comprehension. *NeuroImage*, *25*(3), 1002–1015. <https://doi.org/10.1016/j.neuroimage.2004.12.013>
- Yacovone, A., Moya, E., & Snedeker, J. (2021). Unexpected words or unexpected languages? Two ERP effects of code-switching in naturalistic discourse. *Cognition*, *215*, 104814. <https://doi.org/10.1016/j.cognition.2021.104814>
- Yan, S., Kuperberg, G. R., & Jaeger, T. F. (2017). *Prediction (or not) during language processing. A commentary on Nieuwland et al. (2017) and DeLong et al. (2005)* [Preprint]. Neuroscience. <https://doi.org/10.1101/143750>
- Yarkoni, T., Speer, N. K., & Zacks, J. M. (2008). Neural substrates of narrative comprehension and memory. *NeuroImage*, *41*(4), 1408–1425. <https://doi.org/10.1016/j.neuroimage.2008.03.062>
- Yuan, S., Fisher, C., Kandhadai, P., & Fernald, A. (2011). You can stipe the pig and nerk the fork: Learning to use verbs to predict nouns. *Proceedings of the 35th Annual Boston University Conference on Language Development*, 665–677.
- Zeller, J. P. (2020). Code-switching does not equal code-switching. An event-related potentials study on switching from L2 German to L1 Russian at prepositions and nouns. *Frontiers in Psychology*, *11*, 1387.

Zeller, J. P., Hentschel, G., & Ruigendijk, E. (2016). Psycholinguistic aspects of Belarusian-Russian language contact. An ERP study on code-switching between closely related languages. *Slavic Languages in Psycholinguistics. Chances and Challenges for Empirical and Experimental Research*, 257–278.