



Uncovering the effects of semantic context on the cortical processing of continuous speech using computational models of language

Michael P. Broderick, B.A.I.

Under the supervision of Dr. Edmund C. Lalor

A thesis presented to the
University of Dublin, Trinity College
In fulfilment of the requirements for the degree of
Doctor of Philosophy

January, 2021



Department of Electronic and Electrical Engineering
University of Dublin, Trinity College

Declaration

I, Michael Broderick, confirm that this thesis has not been submitted as an exercise for a degree at this or any other university and is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed,



Michael P. Broderick

September th, 2020

Summary

The semantic context in which spoken words appear will greatly shape how they are understood by the human brain. This understanding is underpinned by a hierarchical system that processes increasingly abstract features of words at successive stages. Discerning how semantic context influences processing at each of these hierarchical stages has been major objective for the cognitive neurosciences. Modern neuroimaging has provided researchers with much insight in this regard, with electrophysiology measuring speech-related neural processes with precise temporal resolution. However, the cortical mechanisms underlying contextual language processing remain unclear. Furthermore, the impact of context on the neural processing of more continuous stretches of naturalistic speech materials is poorly understood. In this thesis, we introduce methodological frameworks for addressing these issues. We utilize modern computational language models, popular in the field natural language processing, in conjunction with recently developed system identification approaches to investigate the role of context in everyday language processing.

The first half of this thesis introduces a novel approach to derive electrophysiological correlates for the semantic processing of continuous speech. In the first study (chapter 3), subjects were presented with narrative speech excerpts while their non-invasive EEG signal was recorded. Semantic features of the stimulus were estimated using word embedding models. We found that the function mapping semantic features to neural responses shared traits with classic N400 component. This index of semantic processing was sensitive to attention and speech intelligibility when tested using different EEG datasets. Our results indicate that the derived neural measures were robustly sensitive to speech comprehension.

Chapter 4 further tests the sensitivity of the measure as a marker of the semantic processing of words in context and attempts to dissociate it from lower levels of processing such as word identification or recognition. This was done by systematically scrambling the order of words in a speech stimulus that was presented to subjects while their EEG was recorded. We found that, when higher levels of word scrambling were introduced, subjects' ability to comprehend the speech decreased. This coincided decrease in the neural marker of semantic processing. These results further support the notion that the newly derived measure reflects the semantic processing of speech.

The second half of this thesis investigates the precise mechanisms underlying contextual language processing and whether the brain uses context to predictively preactivate features of upcoming words. In chapter 5, the method is applied to datasets of younger and older subjects listening to natural speech. Difference in these populations use context-based predictions, particularly at the

level of semantic representation, have been observed in previous studies. This chapter builds on the previous 2 studies by using additional language models that index the relationship between words and context at different linguistic levels. We observe a dissociation between the neural correlates of different language models for older adults. Our results suggest that, while younger and older subjects both employ context-based lexical predictions, older subjects are significantly less likely to pre-activate the semantic features relating to upcoming words.

Chapter 6 investigates how bottom-up sensory inputs combine with top-down contextual prior knowledge to subserve perception. As with previous chapters, we used computational language models to quantify how strongly words relate to their preceding context. We then measure whether this information impacts the acoustic and phonetic encoding of words using a 2-stage regression approach. Our results support the top-down influence of semantic context on the early auditory encoding of words. In addition, we find that the different language model measures independently affect the encoding of auditory information that their effects are dissociable in time. We interpret these results through the lens of predictive coding, where it is believed that distinct neuronal populations exist at each cortical level, encoding prediction and error.

Acknowledgements

I am extremely grateful for the people in my life who have provided friendship and support over these past 4 years. I would like to thank:

Ed, for your brilliant mentorship and constant encouragement. Thanks for making these past 4 years both interesting and fun. You have taught me a lot about the brain and introduced me to frisbee golf, which I will probably never master but will remain forever enthusiastic about.

Richard, for being my surrogate supervisor here at Trinity and for all the advice and support.

Giovanni, for all your help in the first year of my PhD. You were always someone I could turn to if I ever had any questions, no matter how small. You have a fantastic way of getting people excited about science. Thanks for all the enjoyable conversations.

Andy, for the many helpful chats over Skype and your invaluable insight into language modelling

Lalor Lab members past and present, Mick, James, Denis, Aisling, Adam, Nate, Emily, Andy, Aaron, Kevin, Emily, Farhin, Shyanthony, Lauren, for all the helpful feedback and discussion

Aisling, Adam and Emily, for letting me crash with you in Rochester and for being great craic at conferences.

Reilly Lab members, for all the fun boardgame nights, Christmas parties and chats over cake and in the pub

Morgan and Rory, for being incredible house mates in the early days of the PhD and for being an amazing support over the years.

Barry and Daisy, for friendship and the many fun times we've shared together

Mary and Coleman, for being great parents-in-law and for graciously welcoming me into the Parker-O'Toole clan.

My brothers, Matthew, Daniel, Ben and Joe and their partners, Fee, Martha, Phoebe and Millie, for always being a strong support system for me.

Mum, for the wisdom that you continually gift me with and your unconditional love and support

Dad, for your inquisitive mind that has passed down a generation. Thanks for constantly inspiring me. I hope our conversations about Aristotle, Leonard Cohen and the Rorschach continue for many years to come

My life partner Róisín, for being so incredibly supportive and loving to me over the years. You've encouraged me in the toughest times and I couldn't have done this without you. I'm also really impressed at how well you can describe my research to other people. I'm excited to see what life has in store for us

This research was kindly supported by the Irish Research Council (IRC) through a Government of Ireland Postgraduate Scholarship

MICHAEL BRODERICK

Trinity College Dublin

September 2020

Table of Contents

Summary	iii
Acknowledgements	v
Publications Arising from this Thesis	ix
Glossary of Acronyms	xi
List of Figures and Tables.....	xii
Chapter 1. Introduction	1
1.1 Background	2
1.2 Research Goals.....	4
1.3 Thesis Outline	5
Chapter 2. Literature Review	7
2.1 Cortical Processing of Speech.....	7
2.1.1 Hierarchical organisation of speech perception	7
2.1.2 Semantic Processing of Speech.....	10
2.1.3 The Role of Context in Language Comprehension.....	13
2.2 Measuring Cortical Responses to Spoken Language	17
2.2.1 Electroencephalography and Event Related Potentials.....	18
2.2.2 The N400	21
2.2.3 The Temporal Response Function	24
2.2.4 Applying the TRF: System Identification in Action	27
2.3 Uncovering Language Processing in the Brain – Future Directions	29
2.3.1 Language Models and Contextualized Word Embeddings.....	29
2.3.2 Computational Neurolinguistics	32
Chapter 3. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of continuous narrative speech.....	34
3.1 Introduction.....	34
3.2 Methods.....	35
3.2.1 Subjects	35
3.2.2 Stimuli and Experimental Procedure	36
3.2.3 Data Acquisition and Preprocessing	38
3.2.4 Semantic Dissimilarity Calculation	38
3.2.5 TRF Estimation and Evaluation.....	40
3.2.6 Statistical Analysis.....	41
3.3 Results.....	41
3.3.1 A neural correlate of semantic dissimilarity in natural speech	41
3.3.2 Neural signatures of semantic dissimilarity depend on intelligibility.....	43

3.3.3 No evidence of contextual semantic processing for unattended speech	44
3.3.4 On the relationship of the semantic dissimilarity TRF negativity to the N400 component.....	47
3.4 Discussion	49
Chapter 4. Neural Markers of Comprehension for Different Levels of Speech Scrambling	52
4.1 Introduction.....	52
4.2 Methods.....	53
4.2.1 Subjects	53
4.2.2 Stimuli and Experimental Procedure.....	53
4.2.3 Data Acquisition and Preprocessing	54
4.2.4 Stimulus Characterisation	55
4.2.5 TRF Estimation and Evaluation	55
4.2.6 Statistical Analysis.....	56
4.3 Results.....	57
4.3.1 Scrambling window impacts comprehension despite stability of attention	57
4.3.2 Semantic TRF weights pattern with comprehension.....	58
4.3.3 Envelope tracking is unaffected by speech understanding.....	60
4.3.4 Semantic Dissimilarity TRF predicts comprehension scores.....	62
4.4 Discussion	63
Chapter 5. Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy for younger and older adults.....	66
5.1 Introduction.....	66
5.2 Methods.....	70
5.2.1 Subjects	70
5.2.2 Stimuli and Experimental Procedure.....	70
5.2.3 Data Acquisition and Preprocessing	71
5.2.4 Speech Representations.....	71
5.2.5 TRF Estimation and Evaluation	72
5.2.6 Statistical Analysis.....	73
5.3 Results.....	73
5.3.1 Age-related dissociable neural correlates of lexical surprisal and semantic dissimilarity.....	73
5.3.2 Stronger envelope tracking for older subjects despite weakened semantic responses	76
5.3.3 Semantic dissimilarity TRF prediction accuracy in correlated with verbal fluency score	77
5.4 Discussion	78
Chapter 6. Semantic Context Enhances the Early Auditory Encoding of Natural Speech	82
6.1 Introduction.....	82

6.2 Methods.....	84
6.2.1 Data Preprocessing.....	84
6.2.2 Stimulus Characterisation	84
6.2.3 Modelling of the Stimulus-Response Relationship.....	86
6.2.4 Modelling the relationship between semantic information and speech encoding.....	88
6.2.5 Statistical Analysis.....	91
6.3 Results.....	91
6.3.1 Early Auditory Encoding of words is influenced by their semantic information	91
6.3.2 Semantic Information Interacts with the Encoding of Speech Features at Different Hierarchical Levels	94
6.3.3 Impact of Semantic Similarity on Theta Band Speech Tracking	96
6.3.4 Dissociable effects of lexico-semantic measures on auditory encoding of speech....	97
6.4 Discussion	102
Chapter 7. General Discussion.....	106
7.1 A Novel Approach to Study Natural Speech Comprehension	106
7.2 Predictive Language Processing	108
7.3 Modelling the N400	111
7.4 Future Work	113
7.4.1 A framework for modelling wider contextual discourse.....	113
7.4.2 Refining the models	114
7.5 Summary and Conclusions.....	115
Bibliography	116
Appendix I. Comparing Embedding Models for Neural Correlates of Semantic Dissimilarity	137

Publications Arising from this Thesis

Journal Publications

- **Michael P. Broderick**, Andrew J. Anderson, Giovanni M. Di Liberto, Michael J. Crosse, Edmund C. Lalor, “Electrophysiological correlates of semantic dissimilarity reflect the comprehension of continuous, narrative speech”, *Current Biology*, **2018**, 28, 1-7, <https://doi.org/10.1016/j.cub.2018.01.080>
- **Michael P. Broderick**, Andrew J. Anderson, Edmund C. Lalor, “Semantic Context Enhances the Early Auditory Encoding of Natural Speech”, *Journal of Neuroscience*, **2019**, 39(38) 7564-7575, <https://doi.org/10.1523/JNEUROSCI.0584-19.2019>
- **Michael P. Broderick**, Giovanni M. Di Liberto, Andrew J. Anderson, Adriá Rofes, Edmund C. Lalor, “Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing”, *in review*
- **Michael P. Broderick**, Edmund C. Lalor, “Co-existence of representational and error signals in electrophysiological responses to natural speech”, *in prep*

Conference Publications

- **Michael P. Broderick**, Andrew J. Anderson, Giovanni M. Di Liberto, Michael J. Crosse, Edmund C. Lalor, “Electrophysiological correlates of semantic dissimilarity reflect the comprehension of continuous, narrative speech (one page abstract)”. *Irish Conference on Artificial Intelligence and Cognitive Science*, 2018

Other Related Journal Publications

- Andrew J. Anderson, **Michael P. Broderick**, Edmund C. Lalor, “Neuroscience: Great Expectations at the Speech-Language Interface”, *Current Biology*, **2019**, 28 (24), R1396-R1398, <https://doi.org/10.1016/j.cub.2018.10.063>

Poster and Oral presentations

- **Oral Presentation**, ExpORL: Research Group Experimental Otorhinolaryngology, Leuven, Belgium, September 2019, “Modelling higher levels of the linguistic processing with EEG and distributional semantics”.
- **Poster Presentation**, Cognitive Neuroscience Society, San Francisco, USA, March 2019, “Semantic Context Enhances the Early Auditory Encoding of Natural Speech”.

- **Oral Presentation**, Irish Conference for Artificial Intelligence and Cognitive Science (AICS), Dublin, Ireland, December 2018, “Electrophysiological correlates of semantic dissimilarity reflect the comprehension of continuous, narrative speech”.
- **Poster Presentation**, Auditory EEG Signal Processing (AESOP) symposium, Leuven, Belgium, May 2018, “Top-down influences of semantic similarity on low-level speech encoding”.
- **Poster Presentation**, Speech in Noise (SPiN) Workshop, Glasgow, Scotland, January, 2018, “Decoding Attention at Different Hierarchical Levels of Speech Processing”
- **Poster Presentation**, Society for Neuroscience (SfN), Washington D.C, USA, November 2017, “Electrophysiological correlates of semantic dissimilarity reflect the comprehension of continuous, narrative speech”
- **Poster Presentation**, Society for Neurobiology of Language (SNL), Baltimore M.A., USA, November 2017, “Electrophysiological correlates of semantic dissimilarity reflect the comprehension of continuous, narrative speech”

Glossary of Acronyms

A	Audio-only
AI	Artificial Intelligence
ANN	Artificial Neural Network
AV	Audio-Visual
BCI	Brain Computer Interface
CBOW	Continuous Bag of Words
CCA	Canonical Component Analysis
DSM	Distributional Semantic Models
ECoG	Electrocorticography
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
ERP	Event Related Potential
FDR	False Discovery Rate
fMRI	Functional Magnetic Resonance Imaging
HAL	Hyperspace Analogue to Language
HSD	Honestly Significant Difference
ICA	Independent Component Analysis
IFG	Inferior Frontal Gyrus
LME	Linear Mixed Effects
LSA	Latent Semantic Analysis
LTI	Linear Time-Invariant
MCCA	Multiway Canonical Component Analysis
MEG	Magnetoencephalography
MOCA	Montreal Cognitive Assessment
MTG	Middle Temporal Gyrus
mTRF	Multivariate Temporal Response Function
MVPA	Multivoxel Pattern Analysis
NLP	Natural Language Processing
RMS	Root Mean Squared
RSA	Representational Similarity Analysis
SD	Standard Deviation
SI	System Identification
SNR	Signal to Noise Ratio
STG	Superior Temporal Gyrus
STRF	Spectro-Temporal Response Function
STS	Superior Temporal Sulcus
TRF	Temporal Response Function
VF	Verbal Fluency
VLPFC	Ventrolateral Prefrontal Cortex

List of Figures and Tables

Figure 2.1 Neuroanatomy of speech perception.	8
Figure 2.2 Neuroanatomy of semantic processing.	12
Figure 2.3 Models of information flow in speech perception.	14
Figure 2.4 Hierarchical predictive coding.	15
Figure 2.5 Sharpening and Predictive Coding models in language perception.	17
Figure 2.6 Electroencephalography.	19
Figure 2.7 ERP Responses.	21
Figure 2.8 The N400.	22
Figure 2.9 Distributional Embedding Space.	31
Figure 2.10 Computational Neurolinguistics.	33
Figure 3.1 Regularized Regression Analysis for Estimating an Electrophysiological Correlate of Semantic Dissimilarity to Natural Speech.	40
Figure 3.2 Temporal response functions for natural and time-reversed speech.	42
Figure 3.3 The late negativity in the TRF is linearly related to semantic dissimilarity.	43
Figure 3.4 Assessing the effect of comprehension on the electrophysiological index of semantic dissimilarity.	46
Figure 3.5 Comparison of semantic dissimilarity TRF and the classic N400 event-related potential component.	48
Figure 4.1 Behavioural effects for each scrambling condition.	58
Figure 4.2 Semantic dissimilarity TRFs for speech at each scrambling window	59
Figure 4.3 EEG prediction accuracy for semantic dissimilarity.	60
Figure 4.4 Envelope TRFs for speech at each scrambling window.	62
Figure 5.1 Computational models of predictive processing at lexical and semantic level.	69
Figure 5.2 Lexical surprisal and semantic dissimilarity values of individual words.	72
Figure 5.3 TRF weights and prediction accuracies.	75
Figure 5.4 Envelope in older and younger subjects.	77
Figure 5.5 Correlation of TRF measures with verbal fluency.	78
Figure 6.1 Analysis framework.	89
Figure 6.2 Backward Modelling Approach.	95
Figure 6.3 Forward Modelling Approach	98
Figure 6.4 Effects of Semantic Similarity on Tracking in Different Frequency Bands.	99
Figure 6.5 Lexical-semantic β coefficients for 2 nd stage model.	100
Figure 6.6 Lexical-semantic β coefficients for 2 nd stage model based on phonetic feature TRF prediction accuracies.	101
Figure I.1 TRF weights and EEG prediction accuracies for semantic dissimilarity derived using GloVe and word2vec	138
Table 4.1 Subjects LME Model.	63
Table 4.2 Trial LME model.	63
Table 6.1 Variable Correlations.	90
Table 6.2 Variable Correlations (Analysis 2)	91
Table 6.3 Model weights for 2 nd stage regression	92
Table 6.4 Model weights for 2 nd stage model with both lexical-semantic features.	97

Chapter 1. Introduction

Oscar Wilde's *The Importance of Being Earnest* depicts two men's attempts, through lies and deception, to assume the false alias of a man named Earnest. There are many humorous moments throughout the play where the word "earnest" – an adjective meaning sincerity – is juxtaposed against the insincerity of the characters pretending to be named Earnest. This literary device, used by Wilde, Shakespeare, Dickens, and many of the great writers, highlights the fact that a word's meaning – be it the false alias of an insincere character or the quality of sincerity itself – will very much depend on the context in which the word appears. This example is one of many that underscores the complexity of spoken language. Words often have multiple meanings, are heard in noisy environments, are pronounced by unfamiliar accents, and are masked by interfering sounds or speakers. Despite this complexity, the human brain is highly adept at understanding speech. It utilises the information available in context to perform a series of computations that extract meaning from the time-varying sound patterns generated by a speaker's voice. How our brains achieve this remarkable feat has been the focus of intense scrutiny for decades.

Humans possess the unique ability to interact with the world around them through language. Whereas animals and primates can understand and recognise familiar objects, it is our ability to verbally communicate information about such objects that set us apart as a species. Communication is vital for humans. It allows us to express our thoughts and emotions; to relay our past and plan our future. Speech comprehension is foundational for all communication and understanding its neural underpinnings is a vital part of understanding the brain as a whole. Cognitive neuroscience and linguistics have provided much insight into the neural mechanisms supporting language comprehension, including how a word's semantic context can influence how it is perceived by the brain. However, the precise role of semantic context for the linguistic processing of speech is not fully understood and remains a controversial topic.

Advanced modern neuroimaging methods have provided promising directions for uncovering language processing in the brain and have produced interpretable neural correlates reflecting the higher-level linguistic processing of speech and how they are affected by semantic context.

However, methodological constraints have limited researchers to using experimental stimuli consisting of individual words or carefully crafted sentences. Such stimuli are extrapolated somewhat from the type of speech we experience every day. Thus, doubts remain as to whether effects of semantic context are as pervasive in everyday speech processing. The research outlined in this thesis focuses on identifying interpretable neural correlates of higher-level linguistic processing using more naturalistic stimuli. Throughout this body of work, we investigate the role of semantic context and how it impacts online speech comprehension.

1.1 Background

The context in which words are heard will greatly shape how they are processed by the brain. For example, the sentence *I take my coffee with milk and...* will facilitate our ability to recognize and comprehend the word *sugar* (Miller et al., 1951; Fischler and Bloom, 1979; Ehrlich and Rayner, 1981; Staub, 2015). Contexts are rich with information. Non-linguistically, they provide us with the visual information of a speaker's face; information about who is speaking (a scientist or a poet, for example); and information about our listening environment. They are also rich with acoustic and linguistic cues: different intonation or stresses are placed on particular words, lexical associations exist between words, and language follows syntactic rules. It is believed that the brain uses some or all of what has been previously experienced in context to facilitate the processing of new information. Semantic context is considered an important factor in this regard. It represents the overall meaning of what has been previously conveyed to a listener. Semantic information is stored in the brain and routinely accessed during speech comprehension. However, it is unclear how contextual knowledge at the level of semantic representation will influence the semantic processing of incoming information. Furthermore, whether this knowledge penetrates the mechanism of perception at lower processing levels from the “top down” has been widely debated.

There is growing support for the idea that semantic context is used predictively. Prominent theories posit our brains as “prediction machines” (Clark, 2013), constantly predicting the upcoming sensory input based on what has been previously experienced. Within this framework features of upcoming words are pre-activated across multiple linguistic levels before they are encountered from “bottom-up” (Kuperberg and Jaeger, 2016). However, the neural underpinnings for the facilitatory effects of context-based predictions are not fully understood and, although it is widely accepted in other sensory modalities, prediction in language has received considerable push-back (Jackendoff, 2003).

Electrophysiology has provided tools for which we can measure the impact of context on the neural processing of spoken language. Much has been learned from the N400 component of the

event related potential. This is a neural response to words typically measured using electroencephalography (EEG). The N400 largely reflects how the brain processes word meaning and has provided strong evidence for the importance of semantic context for the processing of new information. It has also been used to support theories of prediction in language (Federmeier, 2007). The N400 is typically elicited in response to individual (often anomalous) words within constructed sentences, for example, *I take my coffee with cream and socks* (Kutas and Hillyard, 1980). Although such stimuli have been crucial in isolating the neural processes relating to language, they are rarely heard in everyday speech and lack ecological validity. This raises the question as to whether such mechanisms underlying the N400 are as pervasive when a listener experiences a stretch of more natural speech. Furthermore, evidence in favour of prediction in language has been questioned as to whether such mechanisms exist outside the laboratory environment when individuals have no explicit tasks or listening instructions (Huettig and Mani, 2016). Investigating neural indices for the processing of more naturalistic speech has been an important new direction in recent years.

Recent work has looked at electrophysiological indices of natural, continuous speech processing using advanced techniques such as system identification (Crosse et al., 2016). This work has mainly focused on how the brain processes the low-level acoustic features of continuous speech. However, it is unclear whether these existing indices capture neural processes that reflect the full nature of speech understanding or how meaning is processed by the brain. For example, one can process the acoustic features of a foreign language speaker and, in some cases, recognise familiar phonemes or syllables in the speech signal without comprehending what a speaker is saying. Furthermore, one can hear lists of words in one's own native language and recognise each individual word without understanding the overall meaning (if any) that they represent. It is the process of grasping the meaning of words within their context that truly reflects our understanding of the message conveyed by a speaker. Therefore, measures that capture the processing of word meaning in context, like the N400, could potentially serve as more accurate indices of speech understanding. Such measures have so far been absent in research investigating electrophysiological indices of continuous speech processing.

Deriving measures that reflect the online comprehension of speech that surpass the need for controlled experimental stimuli could have many practical implications for a range of clinical and non-clinical applications. Such measures could be used to test language development in infants; provide an objective assessment of language learning and could be used to test awareness levels for patients in reduced states of consciousness. Language processing in older adults has been a topic of intense focus for years and carries contrasting views from different literatures (Payne and Silcox, 2019). Online measures of contextual language processing that directly model the mechanisms underlying comprehension could serve to reconcile such views. Furthermore, neural

correlates of semantic processing in older groups that are derived from simple tasks like listening to audio books could offer an attractive prospect for monitoring brain health on a continual basis. At a clinical level, combining such methods with the portability and relatively low cost of EEG could serve as diagnostic tools for a range of debilitating language disorders.

Natural language processing (NLP), an area of research that has been amassing much excitement in recent years, could provide the key to deriving objective measures for the higher-level linguistic processing of speech. Each year, computational models of language are achieving increasingly human-like performance at a range of linguistic tasks (Radford et al., 2019). These models, trained on large text corpora, embody the statistical structure of language, and are often grounded in theories surrounding the neurobiology of language. Neurolinguists have begun adopting such models when studying the brain. Advances in NLP could greatly support the development of a methodological framework that derives neural correlates for the semantic processing of words in context during natural speech processing. Major advances in deep learning, such as back-propagation, were the result of a combined collaborative effort between neuroscientists and computer scientists (Rumelhart et al., 1986). Hence, not only does this research stand to benefit from advances in AI and NLP, it could also potentially reciprocate such advances by providing a better understanding of how the brain uses context during language processing.

1.2 Research Goals

The overarching goal of this thesis is to investigate impact of context on cortical processing of continuous, natural speech. Our primary aim is to develop a novel methodological framework to derive interpretable neural correlates for the semantic processing of words in context from electrophysiological responses to natural, continuous speech. This approach will utilise modern computational models that embody the statistical structure of language and are grounded in theories surrounding the neurobiology of language. We aim to test our methodology by applying it to a range of experiments where the speech understanding of participating subjects' is manipulated. A motivation behind this goal is to provide the initial steps towards developing indices that more accurately capture the comprehension levels of individuals, which can be assessed by non-invasive, portable neuroimaging technology.

This research also seeks to better understand the precise neural mechanisms underlying contextual language processing. Specifically, we wish to test whether semantic context is used to predict upcoming words during natural everyday language comprehension and to investigate which linguistic features of upcoming words are predictively preactivated through this process. Investigating how context use changes as a function of age could provide key insights in this area

while also revealing whether older adults utilise context differently during natural speech processing. In addition, the impact of top-down contextual knowledge on bottom up sensory processing is investigated. We will address questions of whether and how context-based predictions impact the processing of speech at lower levels.

1.3 Thesis Outline

In **chapter 2**, the current understanding of the neural basis for speech processing is reviewed. We briefly outline the functional neuroanatomy of speech perception with a particular focus on how the meaning of words is stored in and accessed from the brain's internal store of semantic knowledge. The role context in speech processing and theories surrounding prediction are also discussed. The methods and technology that allow us to investigate spoken language processing in the brain are then reviewed. The N400, a neural marker of language processing, is discussed in detail, given its relevance to this research. Finally, this chapter reviews some of the advances in neurolinguistics that have benefitted from modern computational models of language. We review some of the popular models that are used for a variety of NLP tasks.

Chapter 3 describes the methodological framework used to derive neural correlates for the semantic processing of continuous speech. This method estimates the *semantic dissimilarity* of words compared to their preceding context using word embedding models. This framework is then tested in **Chapters 3 and 4** using a range of different experimental datasets where speech understanding is manipulated. This work reveals an electrophysiological correlate for the semantic processing of speech that is robustly sensitive to whether listeners could understand the speech they were hearing. In addition, **Chapter 4** reveals the measures specificity to the semantic processing of word's following a coherent context by using speech stimuli that are systematically scrambled.

The next two chapters focus on neural mechanisms underlying prediction in language. Specifically, **chapter 5** investigates whether semantic context predictively preactivates linguistic features of upcoming words at distinct levels of representation. Here, we introduce a second computational measure, *lexical surprisal*, that estimates the relationship between words and their preceding context at the level of lexical form. This model and semantic dissimilarity are tested on an EEG dataset of older and younger adults listening continuous natural speech. Results reveal an age-related dissociation between the different language models that suggest the prevalence of context-based predictions at multiple linguistic levels and a shift in comprehension strategy that occurs with age.

Chapter 6 further utilizes the different language models to investigate how top down lexical-semantic context influences the bottom-up auditory processing of words in the early stages of

their utterance. This was done using a novel 2-stage regression framework, which was tested on a natural speech EEG dataset. We reveal the impact of context on the early encoding on acoustic and phonetic features of speech, implicating mechanisms of prediction and supporting theories of predictive coding.

Finally, **Chapter 7** provides a general discussion of the main findings of the research carried out in this thesis and outlines future works that could build on the foundations of this work.

Chapter 2. Literature Review

This chapter provides a review of the literature relevant to this thesis and is divided into 3 areas of focus. Section 1 provides an overview of the neurobiological systems involved in mapping sound to meaning during speech comprehension. Section 2 reviews some of the popular methods currently employed in studying how the brain process language, as well as neural markers arising from such methods that have been attributed to how the brain processes the meaning of words in context. Furthermore, this section covers a relatively new analysis framework that allows for the study of more naturalistic stimulus processing. Finally, section 3 reviews modern computational language models and their application to neurolinguistics.

2.1 Cortical Processing of Speech

To understand how context impacts the neural processes underlying speech comprehension, it is first necessary to understand the functional neuroanatomy of these processes. This knowledge is important in developing the models and methodological frameworks to investigate the role of semantic context in everyday speech. There has been much work characterising the perceptual and conceptual processes that support spoken language comprehension. This section begins with an overview of this work. It gives a brief description of the functional neuroanatomy of hierarchical speech perception. It then looks at the how the brain stores, represents and interacts with the conceptual knowledge necessary to understand word meaning. Finally, this section looks at the top-down effects of context on language perception and reviews theories surrounding prediction in language.

2.1.1 Hierarchical organisation of speech perception

Convergent evidence in the literature indicates that speech perception is underpinned by a hierarchical system, where increasingly abstract features are processed at higher linguistic levels. Neuroimaging studies point to hierarchical levels of specialisation in the brain whereby

anatomically distinct regions mediate between the incoming speech waveform and its computed meaning (Binder, 2000; Davis and Johnsrude, 2003; Scott and Johnsrude, 2003; Rauschecker and Scott, 2009; Okada et al., 2010; DeWitt and Rauschecker, 2012). Vibrations in the air generated by the sound wave are first converted into electrical signals by the cochlea, located in the peripheral auditory system, which travel to the cortex via the central auditory system (Schnupp et al., 2019).

Auditory cortex, located bilaterally in the superior portion of the temporal lobe, is the area of the cortex most relevant to the early stages of the perceptual process (Hackett et al., 2001). Primary auditory areas, including Heschl's gyrus are non-specific to speech and activated for all sounds (Patterson et al., 2002). Like the cochlea, this core area is tonotopically organised and frequency selective (Baumann et al., 2013; Saenz and Langers, 2014). Regions emanating from primary auditory areas show increased sensitivity to speech or speech-like sounds (Figure 2.1). The superior temporal gyrus (STG) is believed to act as the functional and anatomical interface between lower-level acoustic and higher-level linguistic stages of processing (Yi et al., 2019). This region is consistently activated for speech or for sounds that match the spectro-temporal patterns of speech (Zatorre et al., 1992). A meta-analysis of fMRI studies observed a functional subdivision in this area and in regions of the superior temporal sulcus (STS) in the processing of higher-level linguistic units such as phonemes, words, and phrases (DeWitt and Rauschecker, 2012).

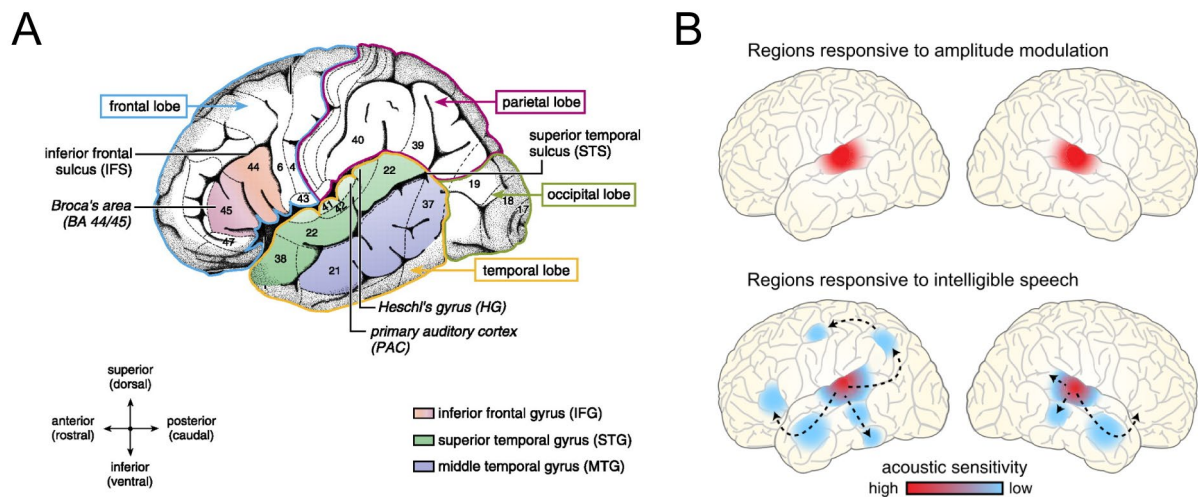


Figure 2.1 Neuroanatomy of speech perception. (A) Figure adapted from Friederici, 2011. The four lobes of the brain (frontal, temporal, occipital, parietal) are marked by coloured borders. Numbers indicate language relevant Brodmann Areas. Brain regions that form the language network are colour coded, this includes superior temporal gyrus (STG), superior temporal sulcus (STS) middle temporal gyrus (MTG) and inferior frontal gyrus (IFG). (B) Regions emanating from primary auditory are believed to show increased sensitivity to speech. Figure adapted from Peelle and Davis, 2012.

An influential model of speech processing, known as the dual stream model, proposes that beyond the STS separate pathways diverge (Hickok and Poeppel, 2007). This consists of a dorsal stream that maps speech to frontal articulatory networks involved in speech production and a ventral stream for auditory-conceptual processing (Hickok and Poeppel, 2004, 2007; Poeppel et al., 2008). Whereas speech processing has been historically linked with left hemisphere lateralisation (Wernicke et al., 1874), the dual stream model proposes that speech comprehension is achieved bilaterally in the ventral stream, with the dorsal stream lateralised to the left hemisphere. Such models have given strong insight into the areas of the brain involved in speech processing. However, from these anatomical models it is unclear how our brains actually transform acoustic signals into meaningful representations like words or sentences. Recent work has given a more mechanistic perspective of how this might be achieved.

During speech perception, listeners extract words from a continuous acoustic stream. However, acoustic properties of any given word are highly variable and will depend on speaker characteristics as well as the acoustic setting in which they are heard (Davis and Johnsrupe, 2007). Listeners are thus faced with the challenge of recognising the same words regardless of the different ways in which they are articulated while also having to detect finer details in the speech signal to distinguish speakers. Although often overlooked because of its innate automaticity in humans, a system capable of addressing this challenge must exhibit combined precision (i.e. ability to detect correct features) and invariance (i.e. show robust recognition in the face of variability). This challenge can be addressed using a hierarchical model. Within this framework, populations of neurons at lower levels are narrowly tuned to detect fine-grained spectro-temporal patterns in the input, whereas neuronal populations at higher levels are largely invariant to these small details and instead respond to more object-centred representations which are generalised and abstracted (DeWitt and Rauschecker, 2012). In other words, it is believed that the perception of spoken words involves some intermediary stage of categorization, whereby continuous acoustic signals are mapped on to a discrete set of linguistic units (Liberman et al., 1967; Mesgarani et al., 2008; Overath et al., 2015). Phonemes have been proposed as the set of intermediary categories in many models of speech perception (McClelland and Elman, 1986; Mesgarani et al., 2014; Di Liberto et al., 2015) as they represent the minimal contrastive units of speech. Although this viewpoint is not unanimous, and the role of the phoneme as an intermediate stage of processing is often questioned (Lotto and Holt, 2000; Daube et al., 2019), the idea that mapping sounds to words via abstract sub-lexical units like phonemes is an attractive one due to its computational efficiency (Scharenborg et al., 2005).

Studies using implanted electrodes have provided further evidence of STG's role as an interface between acoustic and phonetic levels (Chang et al., 2010; Mesgarani et al., 2014). Different regions of STG are sensitive to specific spectro-temporal properties relating to phonological

groupings. It is here that phonetic encoding is believed to be facilitated by feature detectors that integrate specific spectrotemporal cues relative to speech (Chang et al., 2010). The STS is believed to be a downstream area that tracks sequential order of detected phonemes as they appear and binds these smaller units into larger lexical and sub-lexical units (Dehaene et al., 2015). However, there is ongoing debate about whether this is done in anatomically distinct downstream areas, or whether the STG is also responsible for this process, using recurrent connections (Yi et al., 2019).

The ultimate goal of a listener, however, is not to merely perceive and identify speech units like words and phonemes, but to comprehend the message that they represent. This involves making contact with the brain's lexicon, or its internal store of knowledge, to relate formed lexical representations to their underlying meaning. The following section will look at higher levels of lexical and semantic processing, where this is performed.

2.1.2 Semantic Processing of Speech.

Throughout our lifetime, we acquire vast amounts of knowledge about the world around us. Our brains can store this knowledge conceptually, meaning that information can be abstracted and generalized across different events (Binder and Desai, 2011). Semantic processing refers to cognitive act of accessing conceptual knowledge (Binder et al., 2009). It involves the selection and retrieval of internally stored representations supported by a widely distributed network of brain regions. This section reviews popular theories on how conceptual knowledge is stored and accessed during language comprehension.

How conceptual knowledge is stored and represented in semantic memory has been the topic of debate for centuries and can be traced back to Plato and Aristotle. Prevailing theories tend to fall on a continuum between theories of embodiment and disembodiment. Theories of disembodiment propose that all conceptual knowledge is sufficiently represented as symbols in the brain (Fodor, 1983). This implies a rough one-to-one correspondence between a concept and its' symbolic representation. A modification of this theory argues against the idea of discrete symbols and states that all knowledge can be represented as a statistical distribution of words co-occurrences (Landauer and Dumais, 1997). For example, the concept *dog* is represented sufficiently by the words with which it co-occurs in natural language. The most common criticism of disembodiment theories is that they suffer from the grounding problem: how can you understand the meaning of symbol if it is only defined through other symbols (Searle, 1980; Harnad, 1990).

In contrast, embodied cognition theories propose that conceptual knowledge is grounded in sensory and motor experience. In this view, understanding or retrieving a concept involves its simulation or re-enactment in modality specific regions, previously captured by the brain during

perception or action (Barsalou, 1999). For example, when perceiving a dog, internal brain states in visual, auditory and somatosensory modalities are captured and then reactivated later, when the concept *dog* is recalled or imagined. Support for embodiment theories is seen in experiments where patients with neurological diseases affecting the motor cortex, such as Parkinson's or motor neuron disease, show selective impairment for comprehending action words (Bak and Hodges, 2004; Boulenger et al., 2008; Grossman et al., 2008).

Although much empirical evidence suggests that modality specific areas of the brain are indeed *necessary* for conceptual representation, doubts remain as to whether these areas are fully *sufficient* in representing concepts. Functional neuroimaging studies have provided evidence for the existence of supramodal regions, located in the posterior temporal cortex, that are activated consistently across a wide range of semantic tasks (Figure 2.2A; Rodd et al., 2005; Binder et al., 2009; Meteyard et al., 2012). Further support also comes from individuals suffering from semantic dementia. Large atrophy of the temporal lobe is consistently observed in these patients, corresponding to deficits in semantic tasks such as naming objects or sentence comprehension (Snowden, 1989). These findings align with the notion of supramodal convergence zones (Damasio, 1989), where information from different modalities is combined to form generalized concepts (Figure 2.2B; Binder and Desai, 2011; Ralph et al., 2016). This indicates that the organisation of semantic memory falls between extremes of embodiment and disembodiment theories and that conceptual knowledge involves both linguistic (symbolic) and simulated representation (Meteyard et al., 2012). What remains unclear is the extent to which these representations are activated and the depth at which simulation occurs during everyday language comprehension. However, it is believed that the depth of semantic processing, or the extent to which modal features are simulated will much depend on the context, frequency of word use, the goals of the listener (Barsalou et al., 2012).

Once a word is recognised, it interacts with the stored knowledge relating to its meaning in semantic memory. However, not all information about a word is always relevant and will depend very much on the context in which the word appears. For example, different knowledge about a piano is relevant when one is required to listen to it versus when one is required to lift it. In addition, words often have multiple meanings, so that (*parking*) *fine* and (*feeling*) *fine* mean very different things. Therefore, a system capable of dealing with the ambiguity of natural language should require mechanisms of control and selection. Prefrontal cortex is another area of the brain that shows reliable activation during semantic tasks (Rodd et al., 2005). A key study by Badre and colleagues indicated that this region is functionally segregated (Badre et al., 2005). Left anterior ventrolateral prefrontal cortex (VLPFC) was shown to mediate the top-down retrieval of semantic information from posterior temporal cortex, consistent with studies advocating for mechanisms of controlled retrieval in left pre-frontal cortex (Wagner et al., 2001). In contrast,

mid VLPFC supported the selection of relevant semantic knowledge amongst retrieved competitors (Thompson-Schill et al., 1997, 1999). This gives evidence for a mechanism where selection processes operate on active representation that emerge through processes of retrieval.

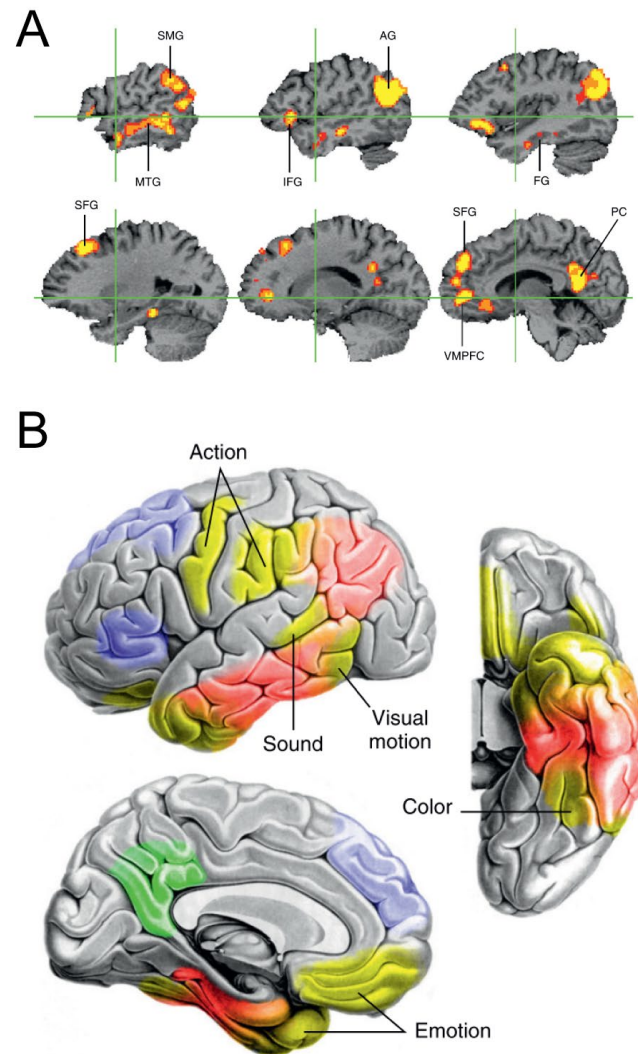


Figure 2.2 Neuroanatomy of semantic processing. (A) Meta-analysis of functional imaging studies of semantic processing. This figure, adapted from Binder et al. 2009, 2011, displays brain regions reliably activated by general semantic processes, based on reported activation peaks from 120 independent functional imaging studies. Selected sagittal sections in the left hemisphere are shown. AG = angular gyrus, FG = fusiform gyrus, IFG = inferior frontal gyrus, MTG = middle temporal gyrus, PC = posterior cingulate gyrus, SFG = superior frontal gyrus, SMG = supramarginal gyrus, VMPFC = ventromedial prefrontal cortex. (B) A neuroanatomical model of semantic processing. Figure adapted from Binder et al. 2011. Modality-specific sensory, action, and emotion systems (yellow regions) provide experiential input to high-level temporal and inferior parietal convergence zones (red regions) that store increasingly abstract representations of entity and event knowledge. Dorsomedial and inferior prefrontal cortices (blue regions) control the goal-directed activation and selection of the information stored in temporoparietal cortices. A similar, somewhat less extensive semantic network exists in the right hemisphere, although the functional and anatomical differences between left and right brain semantic systems are still unclear.

2.1.3 The Role of Context in Language Comprehension

Speech is inherently noisy and often heard under suboptimal listening conditions where the speakers voice is masked by different acoustic and informational content. How do our brains make sense of such a noisy signal? During speech comprehension, our brains undergo the incremental process of building an internal representation of the current event, discourse, or context (Marslen-Wilson, 1973, 1975; Eberhard et al., 1995; Tanenhaus et al., 1995). It is well established that this representation of context will influence the processing of subsequent incoming information (Marslen-Wilson, 1973; Obleser, 2014) and will aid speech understanding in challenging listening conditions (Miller et al., 1951; Warren, 1970). Precisely how context influences the processing of new information is a topic that has drawn considerable controversy over the years. Theories of integration – whereby a word is recognised and subsequently linked with previous input and background knowledge – have historically dominated the thinking in this regard (Kintsch and van Dijk, 1978; Gernsbacher, 1991). Recently, however, the tides appear to be changing in favour of different account: prediction (Kutas et al., 2011; Ferreira and Chantavarin, 2018). The notion for the role of prediction in perception has made considerable gains in popularity over past decade. It stems from theories that posit the brain as a prediction machine (Clark, 2013); constantly estimating sensory information ahead of its subsequent processing. It has also gained empirical support from studies involving anticipatory eye movements (Eberhard et al., 1995; Altmann and Kamide, 1999; Kamide et al., 2003), self-paced reading times (Ehrlich and Rayner, 1981; Monsalve et al., 2012; Smith and Levy, 2013), electrophysiology (reviewed in section 2.2.2) (Federmeier and Kutas, 1999; Van Petten et al., 1999a; DeLong et al., 2005; Federmeier, 2007) and in aspects of spoken communication with turn taking (Sacks et al., 1974).

Central to debates of prediction and integration are questions about how information flows in the brain. As discussed in section 2.1.1, there is a general consensus that the language system is hierarchically organised, where increasingly abstract linguistic features are processed at higher representational levels (Davis and Johnsruide, 2003; Okada et al., 2010). However, there remains some considerable debate about the nature of the connections and the flow of information between the different hierarchical levels. Two opposing views emerge in this debate: the autonomous view argues that information flow proceeds strictly in a feedforward manner (Norris et al., 2000, 2016; Norris and McQueen, 2008), whereas the interactive view argues that feedback connections exist between higher and lower levels and that information flow is bidirectional (McClelland and Elman, 1986; McClelland et al., 2006). Influential models arising from both views have sought to explain the finding that lexical information or context facilitates perception (Helmholtz, 1860.; Miller et al., 1951; Ganong, 1980). In the autonomous view, facilitation occurs at post-perceptual decision stage, whereas the interactive view argues that lexical

information penetrates mechanisms of perception, reshaping the mental representation of a sound at pre-lexical levels. Adjudicating between these opposing views has been difficult because much of the behavioural data can be well explained by both sides. Insights have been gained from neuroimaging studies which provide a more direct measure of the timings and origins of neural processes relating to perception that behavioural tasks fail to detect. One study using combined MEG and EEG revealed an effect of prior knowledge in areas relating to more abstract linguistic processing before areas relating to lower level processes (Sohoglu et al., 2012). However, more work is necessary, especially in the context of natural speech, to compare these competing hypotheses.

Interactive and Autonomous Models

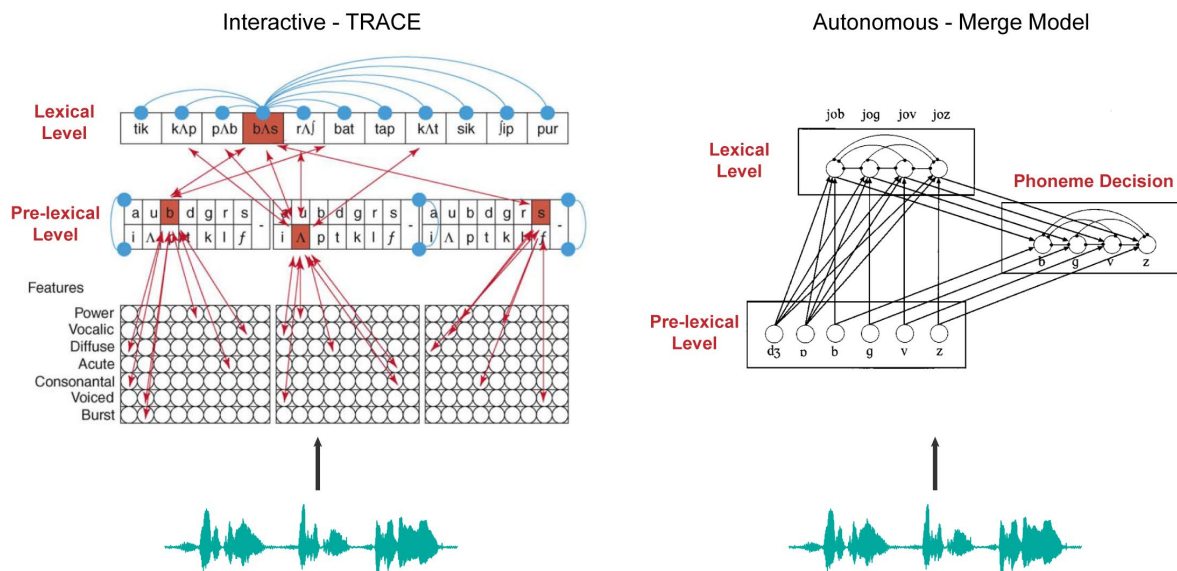


Figure 2.3 Models of information flow in speech perception. Two popular models of speech perception, TRACE (McClelland and Elman, 1986) and Merge model (Norris et al., 2000), are shown. Figure adapted from (McClelland et al., 2006). The models represent opposing views of information flow in the brain. The interactive model (left) proposes bi-directional excitatory connections (red connections) between processing levels with phoneme-level responses produced at the pre-lexical processing level. Units within each layer compete through mutually inhibitory connections (blue connections). The autonomous model (right) posits strictly feedforward excitatory connections from pre-lexical processing to word processing and a separate phoneme decision layer that integrates inputs from phoneme and word processing layers. Units in the word processing and phoneme identification layers compete through mutually inhibitory connections.

The interactive view provides the neural architecture in which predictive mechanisms can exist. However, the precise mechanisms by which predictions are implemented are unclear. Popular theories propose that we interpret our sensory environment through hierarchical Bayesian inference. The brain infers the causes of its sensory inputs by constructing an internal model of

the world and generates predictions which interact with incoming sensory input (Knill and Pouget, 2004; Aitchison and Lengyel, 2017). A central question relating to theories of Bayesian inference is how predicted sensory signals are represented in the brain (Mumford, 1992). Sharpening theories argue that sensory responses that are inconsistent with the current expectation are suppressed by feedback connections from higher levels (Lee and Mumford, 2003), thus an enhanced or ‘sharpened’ version of the expected sensory event is represented in the neural signal through feedforward and feedback connections. Predictive coding offers an alternative mechanistic account whereby feedforward connections carry only the discrepancy (or error) between the predicted signal (embodied in feedback connections) and incoming sensory evidence (Rao and Ballard, 1999; Friston, 2005; Clark, 2013). The goal of perception is to minimise this error by continually updating the brain’s internal model of its sensory environment and thus settling on an accurate representation of the external world. A core tenet of this theory assumes that distinct neuronal subpopulations encoding the prediction (representational units) and the error (error units) exist at each cortical level (Figure 2.5).

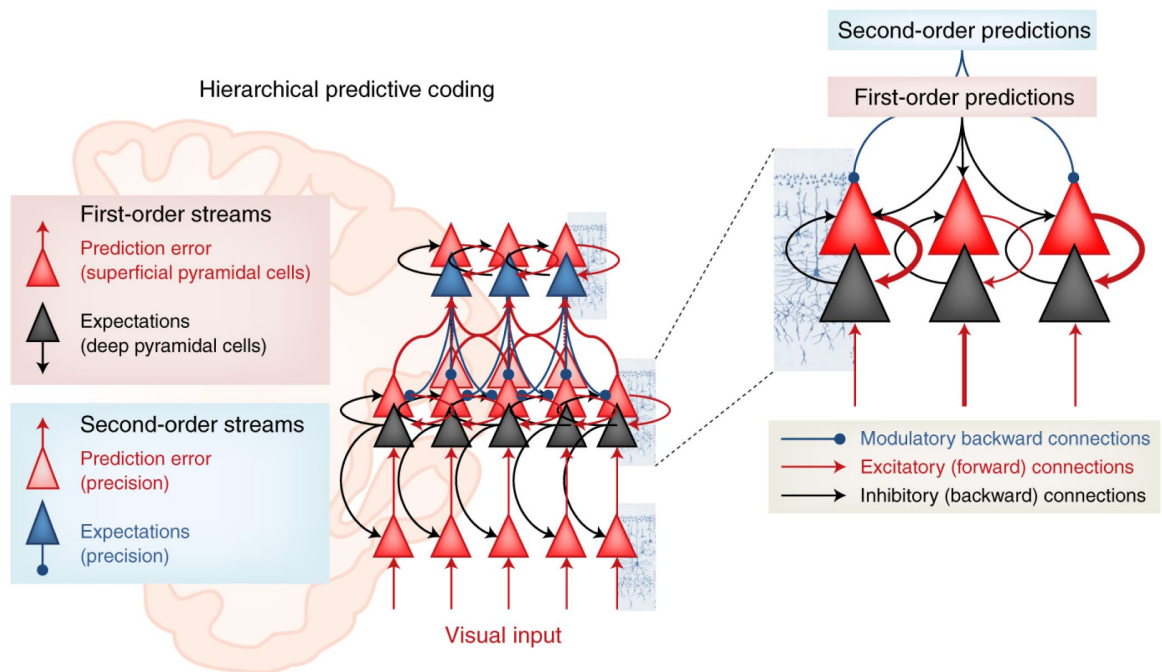


Figure 2.4 Hierarchical predictive coding. Adapted from Friston, 2018. A schematic of how hierarchical message passing operates within predictive coding models. Sensory input is conveyed to sensory cortex via ascending prediction errors. Posterior expectations, encoded by the activity of deep pyramidal cells, are driven by ascending prediction errors (red arrows). These cells then provide descending predictions (black arrows) that inform prediction errors at the lower level. At the same time, they are subject to lateral interactions that mediate (empirical) priors. Subpopulations encoding the prediction (representational units) and the error (error units) exist at each cortical level

Adjudicating between these different accounts has been challenging using traditional contrastive designs. Experiments investigating responses to expected and unexpected stimuli typically reveal a reduction in activity for expected stimuli (Wacongne et al., 2011; Todorovic and de Lange, 2012; Summerfield and De Lange, 2014). However, this can be explained as either a decrease in error signal (predictive coding) or a decrease in unexpected noise (sharpening) (Kersten et al., 2004). Recent pattern analysis techniques (Haxby, 2001; Kriegeskorte, 2008) have contributed to this discussion by testing the represented pattern in the neural signal rather than the amplitude of response. However there has been contrasting results in this regard with some studies revealing a sharpening of expected signals (Kok et al., 2012; Bell et al., 2016; Holdgraf et al., 2016; Leonard et al., 2016) and other studies revealing a decrease in representation, consistent with a suppression of what is expected (Blank and Davis, 2016; Kumar et al., 2017; Blank et al., 2018; Sohoglu and Davis, 2020). These seemingly contrasting results could be due to the use of different stimuli or, alternatively, could be evidence in favour of predictive coding, revealing signals relating to both representation and error. Very few neuroimaging studies, however, have provided empirical support for the simultaneous computation of prediction and prediction error (Egner et al., 2010; Egner and Summerfield, 2013). Recently, a two-process model has been proposed whereby perception is first biased towards prior knowledge and later upweights surprising events (Press et al., 2020). This account captures the dual nature of sufficiently representing what is both true and informative about a sensory event and is consistent with predictive coding accounts.

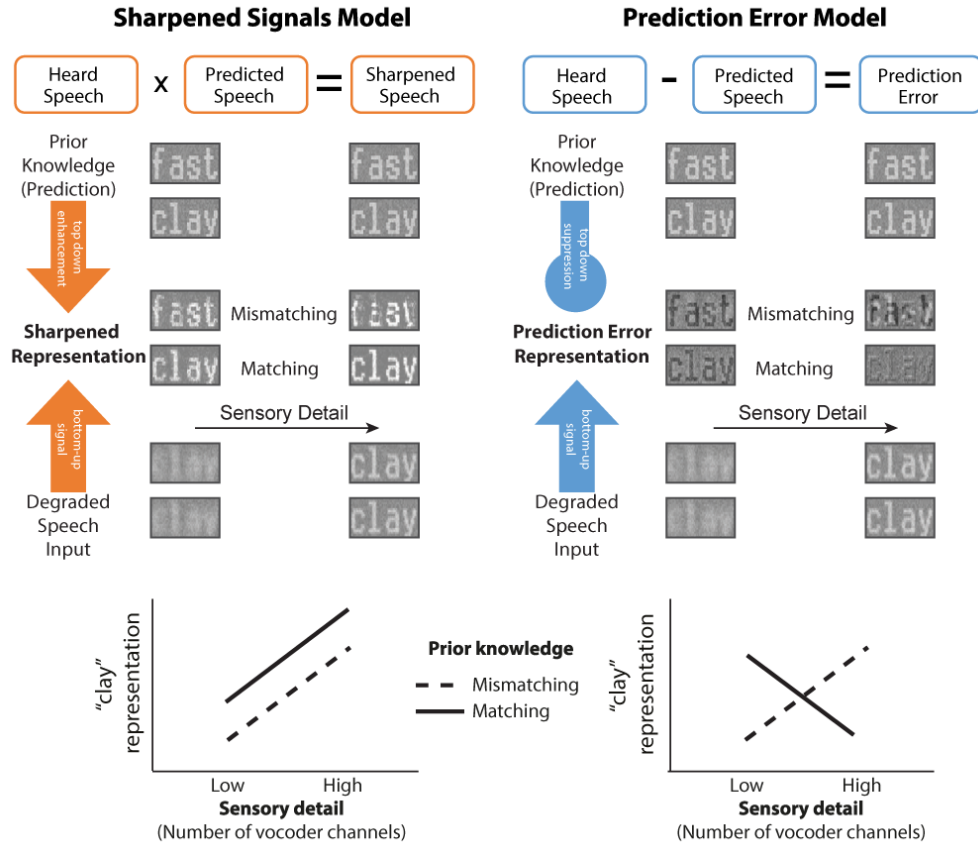


Figure 2.5 Sharpening and Predictive Coding models in language perception. Adapted from Sohoglu and Davis 2020. Schematic illustrating two theories of how top down prior knowledge and bottom-up sensory information are combined to subserve language perception. Schematic depicts an experiment where prior knowledge in the form of a written word (matching or mismatching) is shown to a listener before hearing a degraded version (4 channel or 12 channel) of the spoken word. Degraded speech is depicted as visual degraded text. Under a sharpening scheme, matching prior knowledge enhances the neural representation of the degraded sensory signals. Thus, increasing the sensory information (4ch to 12ch) and prior expectation (mismatching to matching) will result in the increased neural fidelity of the predicted signal. Under the predictive coding scheme, neural representations of the predicted sound are subtracted from bottom up sensory signals. If a word that is expected is confirmed by the bottom up input, the neural representation of this word will be suppressed. This results in an interaction effect between sensory detail and prior knowledge.

2.2 Measuring Cortical Responses to Spoken Language

Before the advent of modern neuroimaging, much of the understanding about language processing in the brain was derived from behavioural experiments (Davis and Johnsrude, 2007) and through studying language impairment in patients with brain lesions or damage to areas of the cortex (Wernicke et al., 1874). While this work is highly valuable and still bears much relevance today it represented an indirect way for measuring the precise neural mechanisms involved in spoken language comprehension. Neuroimaging techniques provided a complementary approach to study the brain, by giving a snapshot of activity at the exact moment

that a stimulus is processed. This section reviews neuroimaging techniques and analysis frameworks most relevant to the thesis. It begins with an overview of electroencephalography and the event related potential. It then covers arguably the most important electrophysiological components relating to language; the N400. This is followed by a look at advances that have been made using system identification methods.

2.2.1 Electroencephalography and Event Related Potentials

The average human brain contains approximately 86 billion neurons (Herculano-Houzel, 2009), which are interconnected forming complex networks. Neurons are excitable cells with intrinsic electrical properties and their activity elicits electric potentials that can be measured from the scalp using electroencephalography (EEG). Specifically, EEG measures the summed, synchronous post-synaptic activity of populations of neurons called pyramidal cells; a type of cell that accounts for approximately 80 percent of all cortical cells, mostly orientated perpendicularly to the cortical surface (Bok, 1960). The activation of these neuron ensembles generates electrical fields whose potentials can be measured from the scalp by placing electrodes on its surface (Figure 2.6).

EEG has several limitations, foremost its low spatial resolution. The number of surface electrodes varies between EEG systems, however even high-density systems with 512 electrodes are limited in their spatial resolution. As a result, EEG electrodes measure only macroscopic population effects of potentials generated by approximately 10^7 to 10^9 neurons. Furthermore, the electrical fields elicited by cortical neurons are separated from recording electrodes by several anatomical layers of tissue and bone which act spatial filters, smearing the signal across the scalp; an effect known as volume conduction (Freeman et al., 2003). In addition, these layers act as low-pass filters which attenuate important neural information above $\sim 30\text{Hz}$.

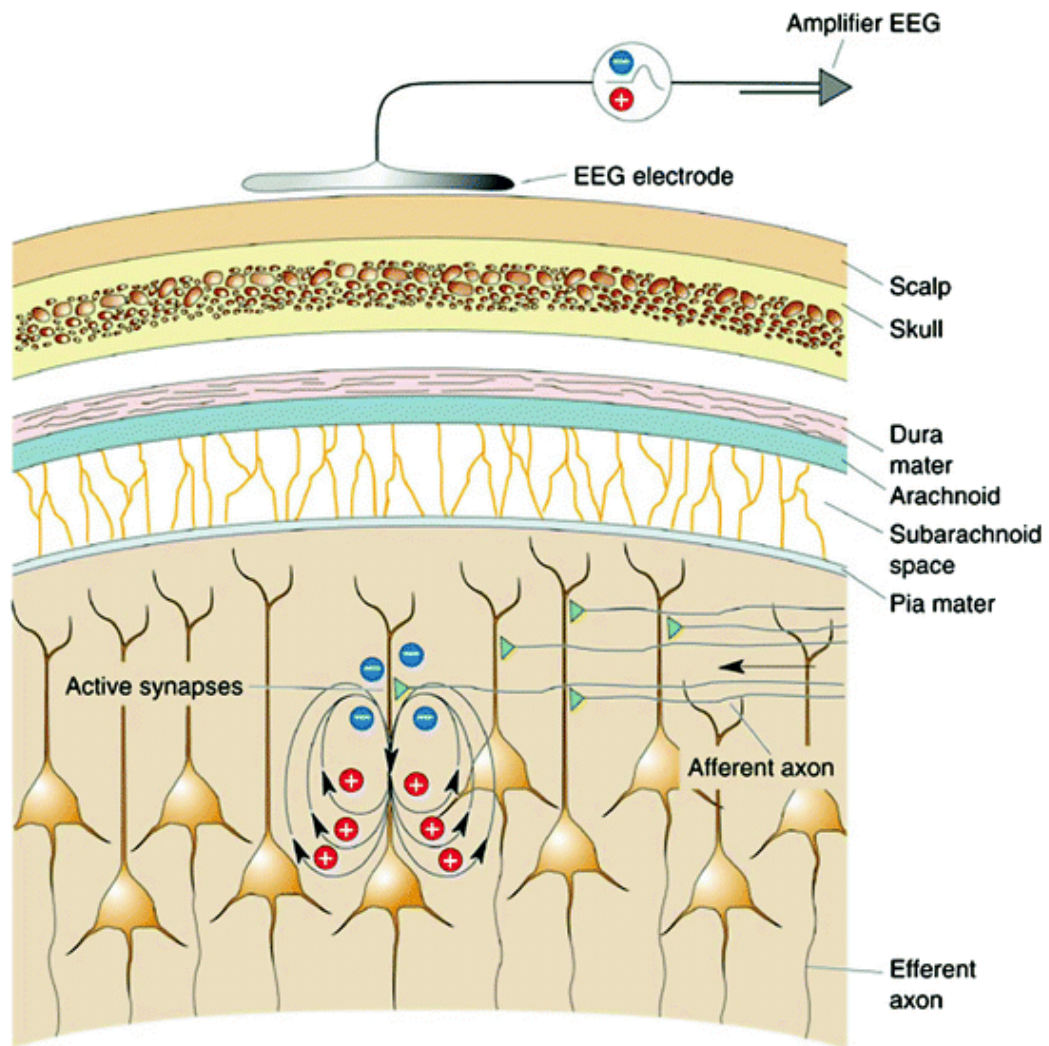


Figure 2.6 Electroencephalography. Adapted from Siuly et al., 2016. Illustration of generation of very small electrical fields by synaptic currents in pyramidal cells. The EEG electrode measures the signal through the thick layers of tissues. It measures the summed, synchronous post-synaptic activity of populations of neurons.

Electrical potentials arising from brain activity are small in magnitude, typically within the range of $20\text{-}40\mu\text{V}$, and therefore must be amplified for detection. However, the skull is also surrounded by muscle tissue which generated interfering electromyographic (EMG) signals, typically orders of magnitude larger than cortical signals and often observed for muscle activity associated with eyeblinks. In addition, eye movement resulting in electrooculogram (EOG) activity can provide further inference (Corby and Kopell, 1972). These signals are referred to as EEG artifacts and can mask recorded neural activity. Researchers typically attempt to remove these artifacts during an EEG preprocessing stage and wide range of techniques are available that help reduce the signal to noise ratio (Hyvärinen, 1999; de Cheveigné, 2010).

Despite its limitations, non-invasive EEG offers some advantages over other neuroimaging techniques. Functional magnetic resonance imaging (fMRI) measures small changes in blood flow in the brain, and thus is thought to reflect neuronal activation. This technique can capture

activity with high spatial resolution, at the level of millimetres. However, fMRI has very poor temporal resolution (sampling rates of $\sim 1\text{Hz}$) which hampers its ability to capture the fast dynamics of speech. It also is unsuitable for capturing the precise timings of neural processes, which is important to understand the mechanisms underlying speech perception. Magnetoencephalography (MEG), like EEG, provides much better temporal resolution in the order of milliseconds. MEG measures the magnetic counterpart to the electric fields elicited by neurons and is non-invasive. In addition, it has higher spatial resolution than EEG. However, MEG is very expensive, unportable and highly susceptible to noise. Although this may not be an issue when its sole purpose is understanding the brain, its potential for use in practical applications is tenuous. Finally, invasive brain recordings, such as electrocorticography (ECoG), which record electrical activity from the brain's surface, provide both high spatial and temporal resolution. However, implanting electrodes is a highly invasive procedure and is typically reserved for patients with severe epilepsy undergoing surgery. Furthermore, ECoG can only sample restricted areas of the cortex, meaning that measurement of the entire language network cannot be captured. In contrast to these listed neuroimaging techniques, EEG is non-invasive, captures that fast dynamics of speech, is portable and cheap, and has the potential to be used in real-world applications.

EEG is primarily used to study how the brain responds to sensory stimuli which typically come in the form of discrete events. EEG activity is analysed within a time window of several hundred milliseconds after stimulus onset; this is known as the event related potential (ERP). The ERP waveform represents the sum of several relatively independent neural components, which can originate from different sources and reflect different cognitive subprocesses (Luck, 2014). As noted above, EEG is highly susceptible to interfering noise and has a low signal to noise ratio (SNR). Therefore, individual responses to stimuli are not readily identified from the raw neural signal. This is overcome by calculating the average of multiple responses to repeated stimuli across time. By assuming that noise is a stochastic signal with zero mean, averaging multiple responses results in its cancellation leaving only the time-locked stimulus related neural response. ERP components are studied in terms of their morphology, scalp topography and timing and are typically denoted based on their polarity and latency. For example, auditory evoked potentials are characterised by positive and negative components at certain time points which are labelled P1, N1, P2, N2 (Picton, 2013; Figure 2.7A). In language processing, it is thought that later components reflect higher level computations (Salmelin, 2007; Figure 2.7B). One such component, known the N400 is commonly related to semantic processing and has been studied in great detail (Kutas and Federmeier, 2011). The following section provides an overview of this component.

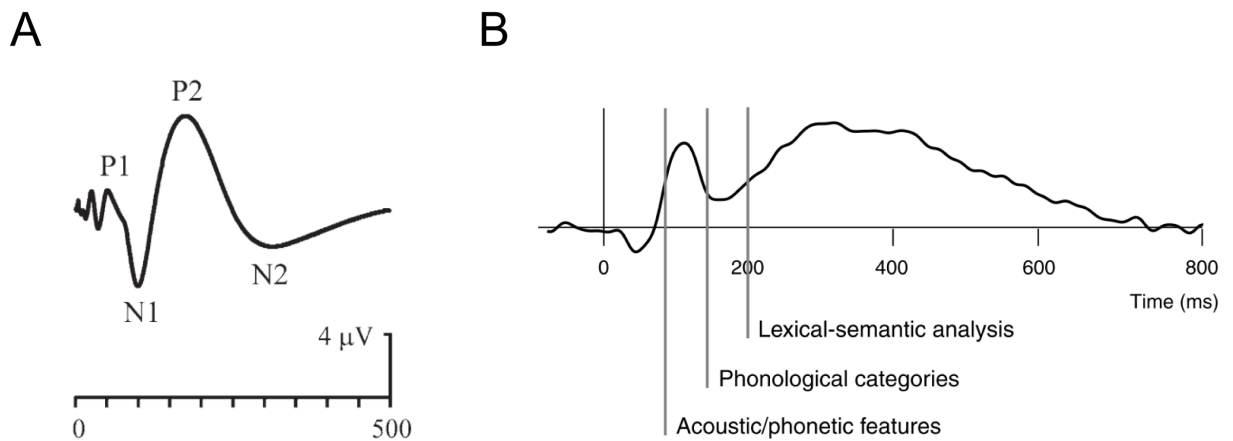


Figure 2.7 ERP Responses. (A) Auditory evoked potential averaged over multiple repetition of some discrete auditory stimulus. Adapted from Picton, 2013. (B) The time course of speech perception, indicated using the ERP. Adapted from Salmelin, 2007.

2.2.2 The N400

The N400 component of the event related potential was first discovered by Kutas and Hillyard in response to semantically anomalous words at the end of visually presented sentences, for example *I take my coffee with milk and socks*. Such endings elicited a late negative response that peaked around 400ms over midline parietal electrodes (Kutas and Hillyard, 1980b). They labelled this response the *N400 component* owing to its peak negative deflection (N) and its robust timing. The *N400 effect* is characterised as the difference in ERPs elicited by congruous and incongruous endings. Further investigation of this response found that it was not only elicited by anomalous words but rather reflected the processing of all meaningful or potentially meaningful stimuli. The strength of the N400 reflects how the processing of a word was facilitated by its context. Its amplitude is linearly related to a word's cloze probability, which is estimated in offline tests as the proportion participants who would continue a sentence fragment with that word (Taylor, 1953; Kutas and Hillyard, 1984). The N400 can also be evoked by pseudowords (*blick*) (Rugg, 1984, 1987) and is sensitive to factors like the position of a word in a sentence, word frequency and word repetition (Kutas, 1993). The response was insensitive to unexpected physical properties of the stimulus ("*Coffee with cream and SUGAR*") (Kutas and Hillyard, 1980a) or grammatical errors ("*Dogs have four leg*") (Kutas and Hillyard, 1983). Furthermore, the N400 isn't limited to stimuli presented in the visual modality but can be evoked by spoken words (Connolly et al., 1990; Holcomb and Neville, 1990; Bentin et al., 1993), pictures (Ganis et al., 1996), sign-language (Kutas et al., 1987), visual-scenes (Sitnikova et al., 2008) and even smells (Grigor et al., 1999). Taken together, this range of findings supports the notion of N400 as a response to semantic meaning that is amodal in nature.

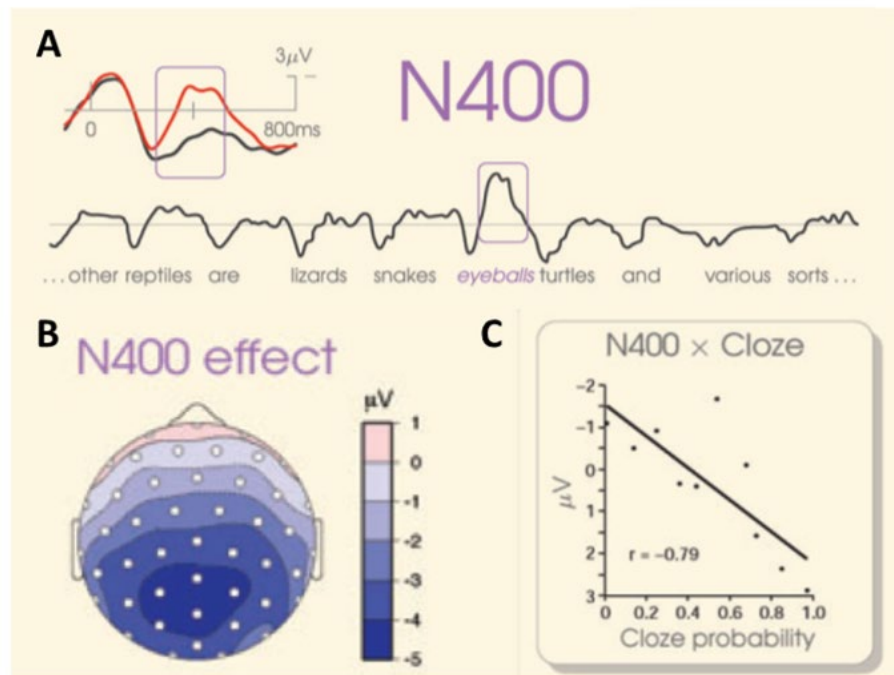


Figure 2.8 The N400. Adapted from Kutas and Federmeier 2011. The N400 component of the event related potential is typically evoked for linguistic stimuli such as spoken or written words. **(A)** It is characterized by a negative deflection, elicited 200-600ms after word onset, peaking at 400ms. Its peak negativity is strongest for words that are incongruent with their previous context (e.g., “I take my coffee with cream and salt”). **(B)** The component is strongest over centro-parietal electrodes. **(C)** N400 amplitude is linearly related to a word’s cloze probability.

The N400 provided key insights into predictive processing in the brain. Several studies have provided powerful and compelling evidence that language comprehension involves the interaction between top-down and bottom-up mechanisms and that context shapes word processing from its earliest stages or even in advance. In one study subjects read highly constraining sentences (*for example: It was a sunny day, so the boy went out to fly....*) where the article preceding the target word (i.e. *a* or *an*) matched or mismatched with the most likely completion (*kite*). Reduced N400 responses were seen for preceding articles that matched the expected completion (*a kite*) compared to articles that matched an unexpected completion (*an airplane*) (DeLong et al., 2005). This finding was supportive of anticipatory processing during language comprehension following the logic that both indefinite articles constitute equally good fits with the preceding context. It is only the anticipated noun that modulates responses. Similar effects were observed when function words matched or mismatched the gender of a predicted target for languages like Spanish (Wicha et al., 2003; Wicha et al., 2003; Wicha et al., 2004).

Another line of work hypothesized about the predictive pre-activation of words and features during language comprehension. Sentences like *I take my coffee with cream and ...* were presented to subjects where the target word was expected (sugar); unexpected but from the same semantic category as the expected item (salt); or unexpected and semantically unrelated (socks).

Both unexpected items showed an increased N400 negativity, relative to the N400 for expected items. Importantly, however, this negativity differed as a function of a word's semantic association with the expected items, with categorically related items showing reduced negativity compared unrelated items (Federmeier and Kutas, 1999). This pattern of responses was also stronger with higher sentential constraint i.e. *I have a sweet tooth and I take my coffee with cream and...* This work provided evidence that the brain anticipates upcoming words ahead of perceiving them by pre-activating their related semantic features. It established that organisation of semantic memory had an inevitable impact on how the brain makes sense of language in real-time (Kutas and Federmeier, 2000). Finally, Van Petten and colleagues observed reductions in N400 responses to unexpected spoken words that shared the same initial phonemes as expected completion, for example *In New York, I bought a pair of jeans for 20 dollars/dolphins/scholars* (Van Petten et al., 1999b). This study indicated the impact of top-down prior expectation on the processing of sub-lexical information at lower hierarchical level. Overall, these studies gave credence to the idea of predictive language processing in the brain and suggested that the N400 component was reliable marker of such processes.

Over the years, several prominent theories have emerged as to the functional specificity of the N400. For the most part they were based on the classical view of language processing (Chomsky, 1957) and placed the N400 somewhere along the pathway of cognitive operations that convert sound to meaning. The “late” or *post-lexical* view of the N400 is that it reflects the process of integrating a word into the current context which is constructed from linguistic and non-linguistic cues to meaning (Hagoort et al., 2009). The more dissimilar a word is with its context, the more effort required in integrating it and thus the larger neural signal. This theory is supported in part by the multi-modal nature of the N400 and can readily explain why N400 responses show a sensitivity to world knowledge (Hagoort et al., 2004), information about the target speaker (Van Berkum et al., 2008) and pragmatic or discourse level representations (Nieuwland and van Berkum, 2006). Problems with this view arise from the fact that pseudowords or illegal word strings that do not map onto lexical items can produce N400 responses (Rugg, 1984, 1987). In addition, the response is sensitive to pre-lexical and lexical information like word frequency and orthographic neighbourhood size (Laszlo and Federmeier, 2011), which is likely processed before integration mechanisms. An alternative view is that the N400 reflects the stage of lexical access, when words are retrieved from the long-term or semantic memory (Kutas and Federmeier, 2000). This *pre-lexical* view aligns with the idea of a predictive brain in that reductions in the N400 are due to facilitated access to words or semantic features that have been pre-activated in advance (Kutas and Federmeier, 2007).

It has often proven difficult to distinguish between these different accounts, as contextual factors that benefit integration could also be viewed as facilitating access. Furthermore, key results in

favour of prediction has been recently challenged by a multi-lab study (Nieuwland et al., 2018) (but see Yan et al., 2017 for an interesting reanalysis of this dataset). Studies have claimed that locating the anatomical origin of the N400 would inform on what cognitive process it reflects (Lau et al., 2008), whereas others have argued that neither of these prominent views of the N400 (nor any to date) have been able to properly capture all the relevant data relating to the ERP component (Kutas and Federmeier, 2011; Rabovsky et al., 2018). One idea that has the potential to unify several of these competing theories is that the N400 reflects the stimulus induced change in a multimodal neural network, wherein an implicit and probabilistic representation of sentence meaning is held (Kutas and Federmeier, 2011). Context serves to constrain the formation of this internal representation as a sentence unfolds. Importantly, the state of this internal network can be shaped by predictions, such that information can be partially or fully activated before the arrival of bottom-up input. Recent work has produced a model embodying this hypothesis which has successfully simulated several different phenomena observed from the N400 effect in previous literature (Rabovsky et al., 2018). Regardless of the debate around its precise neural underpinnings, the N400 has been a very powerful tool for studying the processing of meaning in the brain and how it may be impacted by context. However, issues arise over the nature of how it is estimated.

2.2.3 The Temporal Response Function

ERPs have been used extensively to investigate how the brain processes speech and language. These methods have proven to be very effective when studying short, discrete, and repeated stimuli such as tones, phonemes, syllables, or words. However, in everyday spoken language, such linguistic units are rarely heard in isolation but as part of a continuous signal that represents the wider discourse of connected speech. The necessity of averaging responses to repeated stimuli limits the ERP approach in this regard and often constrains the breath of linguistic stimuli that can be investigated to a subset of categories configured into matched experimental pairs. Thus, the degree of ecological validity of results generated from bespoke ERP setups is unclear because such experimental conditions are rarely experienced in everyday life. For example, the N400 effect is obtained in many studies using incongruent sentences that are rarely spoken or heard. Furthermore, previous studies reveal that evoked responses to syllables differ depending on the context in which they are presented (Bonte et al., 2006), raising concerns of whether the brain responds differently to isolated linguistic units. Recent efforts have therefore begun to focus on using continuous and more naturalistic stimuli in experiments (Hamilton and Huth, 2018).

The ERP waveform can be alternatively viewed as the result of a convolution between a temporally discrete stimulus, characterised as a transient impulse, and the system's (i.e. the brain's) response function (Lalor et al., 2009; Lalor and Foxe, 2010). For continuous stimuli, this

response function can be approximated using system identification (SI) methods (Marmarelis, 2004), which provide a way to mathematically model this mapping between stimulus and response. One class of models used with this technique are known as linear models (Wu et al., 2006). A key underlying assumption for these models is that the brain is a linear time-invariant (LTI) system (Bialek et al., 1991). While this assumption may be very strong (although see Boynton et al., 1996), it affords the estimation of the system through linear regression. The model or mapping function goes by many names but is often referred to as the Temporal Response Function (TRF).

One important consideration when modelling neural responses with the TRF is the initial choice of feature space(s) to represent the stimulus. Speech utterances contain information at different representational levels and prior research and knowledge of what is encoded in the brain should guide this initial feature extraction step. For example, continuous speech can be represented in terms of the fluctuations in acoustic energy, known as the speech envelope, and it has been shown that low frequency neural activity encodes this information (Ahissar et al., 2001; Luo and Poeppel, 2007; Aiken and Picton, 2008). Therefore, the speech envelope is a very common stimulus feature used as input to the TRF.

TRFs can be estimated using a forward modelling approach, which learns the mapping of stimulus features to neural responses, or using a backward modelling approach, which learns the mapping of neural responses back to the eliciting stimulus. For the forward modelling approach, this mapping is described mathematically below. For N recorded channels, the instantaneous neural response $r(t, n)$, sampled at times $t = 1 \dots T$, and channel n , consists of a convolution of the stimulus feature, $s(t)$, with a channel-specific TRF, $w(\tau, n)$. The response is described as follows:

$$r(t, n) = \sum_{\tau} w(\tau, n) s(t - \tau) + \varepsilon(t, n) \quad (2.1)$$

Where $\varepsilon(t, n)$ is the residual response at each channel not described by the model. The TRF, $w(\tau, n)$ linearly transforms the ongoing stimulus feature $s(t - \tau)$ to the ongoing neural response $r(t)$ at each channel n , over a series of specified time lags τ relative the instantaneous occurrence of the stimulus feature $s(t)$. Time-lags typically range over a few hundred milliseconds. This range can be specified to match with the length of an ERP epoch in order to compare responses (Lalor et al., 2009).

The weights of the TRF are estimated by minimizing the mean-square error (MSE) between the actual neural response $r(t, n)$ and its prediction $\hat{r}(t, n)$ which is estimated from the convolution:

$$\min(\varepsilon(t, n)) = \sum_t [r(t, n) - \hat{r}(t, n)]^2 \quad (2.2)$$

This minimization problem is solved using reverse correlation, written in matrix format as follows (de Boer and Kuiper, 1968).

$$w = (S^T S)^{-1} S^T r \quad (2.3)$$

Here, stimulus lag matrix S is a time delayed version of the stimulus feature, with τ_{max} and τ_{min} representing the maximum and minimum time lags.

$$S = \begin{bmatrix} s(1 - \tau_{min}) & (-\tau_{min}) & \dots & s(1) & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & s(1) & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & s(1) \\ s(T) & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 0 & s(T) & \dots & \vdots & \vdots & \dots & \vdots \\ \vdots & 0 & \dots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & s(T) & s(T-1) & \dots & s(T - \tau_{max}) \end{bmatrix} \quad (2.4)$$

One last consideration when calculating the TRF is regularisation. The $S^T S$ term in equation (2.3) refers to the autocovariance matrix of the stimulus. When this matrix is inverted, it is particularly prone to numerical instability when solved with finite precision. Introducing a bias term λ or ‘smoothing factor’ to the equation in (2.3) solves this problem

$$w = (S^T S + \lambda I)^{-1} S^T r \quad (2.5)$$

Where I is the identity matrix. The addition of the bias term penalises very large TRF weights and prevents the model from overfitting to the data on which it is trained, helping it to generalise to new unseen data. This regularization procedure and regression approach are known as Tikhonov regularization and ridge regression, respectively.

The abovementioned model assumes a univariate input signal or single feature dimension. However, this can be easily adapted for the simultaneous testing of multiple feature dimensions. This is relevant for testing features spaces with multiple dimensions, for example the spectro-temporal properties of speech represented by a spectrogram. For an input stimulus representation with multiple features F the stimulus lag matrix is extended so that every column is replaced with F columns. This extended model is known as the multivariate TRF (mTRF).

As mentioned above, the stimulus-response mapping can also be performed in the opposite direction, known as backwards modelling. Both approaches offer complementary advantages. Whereas the forward modelling benefits from producing channel specific TRFs and can more easily incorporate different multivariate speech representations, it suffers from the fact that it treats each EEG channel as a univariate feature. The backward modelling approach, on the other hand, takes into account the signal from all recorded electrodes simultaneously and thus can do a

better job at separating the speech-related signal from the background EEG. Backwards models, or decoders, can be modelled similarly to TRFs. The linear mapping between the neural response, $r(t, n)$, and stimulus, $s(t)$, can be expressed as

$$\hat{s}(t) = \sum_n \sum_\tau r(t + \tau, n)g(\tau, n) \quad (2.6)$$

where $\hat{s}(t)$ is the reconstructed stimulus feature; n indicates the recorded neural channels, and τ is a specified number of time lags. Here, the decoder, g , is estimated using the following matrix operations with regularization:

$$g = (R^T R + \lambda I)^{-1} R^T s \quad (2.7)$$

A central feature of these models is their ability to predict unseen neural data (with forward modelling) or reconstruct stimulus features (with backward modelling). Prediction or reconstruction accuracy can be used as quantitative measure to assess how well particular speech features is encoded in the brain. Model performance is estimated using a cross-validation procedure. A subset of data (training set; e.g. 90% of the data or all but one trial) is used to train the model which then predicts responses or reconstructs stimulus features for the remaining, held-out data (validation set). Model accuracy is then calculated by comparing the predicted/reconstructed data with the original using Pearson's correlation. Model weights can also be studied in terms of their morphology, peak latency, and topographical distribution in a similar way to ERPs.

2.2.4 Applying the TRF: System Identification in Action

System identification methods, encompassing the TRF, offer a major methodological advance for studying how the brain processes natural, everyday speech (Holdgraf et al., 2017). Below are some examples of how the TRF has been used to study speech processing in more naturalistic settings.

Humans are highly adept at focusing their attention on a single speaker and ignoring all other background noise; a phenomenon known as the cocktail party effect (Cherry, 1953; Broadbent, 1958b; Deutsch and Deutsch, 1963; Treisman, 1964; Bronkhorst, 2000). ERPs have been used to study neural underpinnings of this effect but are hampered by the difficulty in resolving responses to more than one simultaneously presented discrete event. Despite efforts to mitigate this problem (Hink and Hillyard, 1976; Woods et al., 1984; Coch et al., 2005; Nager et al., 2008) issues persist in the form of temporal smearing which make it difficult to determine the temporal loci of attentional effects. The TRF approach offers an alternative means to study attentional effects without these issues. Separate TRFs for each speaker can be fit to recorded neural responses to

the multi-speaker stimulus. Using this method, a late locus of attention ($\sim 200\text{ms}$) was observed in TRF weights over temporal EEG electrodes (Power et al., 2012) and MEG sensors (Ding and Simon, 2012b). Backwards TRFs have also been used to successfully decode listeners' attention by comparing the reconstruction accuracies of attended vs unattended decoders (Ding and Simon, 2012a; Mesgarani and Chang, 2012; O'Sullivan et al., 2015). This has prompted a new line of research aimed at decoding auditory attention using an individual's neural signal for the potential use in real-world applications (Biesmans et al., 2017; Wong et al., 2018; Ciccarelli et al., 2019).

The TRF has also been used to study neural responses to speech under various levels of signal degradation. It has been shown that shifts in the tuning properties of the TRF and increases in model prediction accuracy coincide with an increase in the subjective perceptual clarity of a stimulus (Ding et al., 2014; Holdgraf et al., 2016). Increases in perceptual clarity can arise from increasing the SNR of the signal or by providing cues to degraded speech such as prior knowledge (Di Liberto et al., 2018) or visual information about a speaker (Crosse, Di Liberto and Lalor, 2016), which suggests that neural responses indexed by the TRF are impacted by some top-down process. Irrespective of the neural source of perceptual enhancement, the TRF has been shown as reliable measure of speech intelligibility (Vanthornhout et al., 2018; Decruy et al., 2019). Researchers have explored this idea further by attempting to index the comprehension levels of individuals listening to natural speech using the TRF (Keitel et al., 2018; Octave and Reichenbach, 2019; Verschueren et al., 2019; Zou et al., 2019).

Nevertheless, there are some shortcomings with current approaches employing the TRF to assess spoken language processing in the brain. As mentioned previously, the speech envelope is often used as the input stimulus representation to the TRF. The speech envelope is an important cue for intelligibility (Drullman et al., 1994b, 1994a; Shannon et al., 1995) and its clear from previous research that envelope tracking represents more than basic acoustic processing and is likely influenced by some speech-specific component (Pelle et al., 2013). However, the speech envelope is a fundamentally low-level representation of speech that measures amplitude fluctuations in acoustic energy. Although dominant frequencies in this signal correspond to the rate of important linguistic units on multiple time-scales (Ding et al., 2017), the represented linguistic features are conflated into a single univariate measure. Thus, it is unclear which features of speech are being tracked by the brain using the envelope-based TRF and dissociating the neural processes (for example acoustic processing, speech perception or speech comprehension) that underly envelope tracking has been a major challenge (Ding and Simon, 2014). In addition, reliable envelope tracking has been shown for sounds such as reversed speech (Howard and Poeppel, 2010), music (Doelling and Poeppel, 2015), foreign language (Peña and Melloni, 2012) and amplitude modulated tones (Lalor et al., 2009; Drennan and Lalor, 2019), that do not

necessitate either speech perception or comprehension. This has raised questions about the suitability of envelope-based TRFs as a reliable index of speech comprehension.

These issues arise from inherent properties of the TRF. Unlike the ERP approach, the TRF only reflects cortical responses to specific stimulus features defined by the experimenter. If the specified input representation is the speech envelope, this will constrain the model to reflect only the encoding of that feature. Recent efforts have attempted remedy these issues by looking at higher level feature spaces and ‘pushing the envelope’. Most notable are studies that represent speech in terms of phonemes and phonetic features. This work indicates that brain encodes the phonetic features of speech and this can be indexed using EEG (Di Liberto et al., 2015). However, speech is comprised of much higher-level features and, to date, methods attempting to capture higher level speech features of continuous speech with electrophysiology have been vastly unexplored. As alluded to in the introduction, indexing speech at these higher levels of linguistic representation will be central to this thesis. The final section of this literature review looks at the emerging field of computational neurolinguistics as a source of how these higher-level representations might be computationally realised.

2.3 Uncovering Language Processing in the Brain – Future Directions

Natural language processing (NLP) has changed the landscape of science and technology in the past decade. This field has seen the rise of increasingly more accurate models, trained using artificial neural networks (ANNs) (Devlin et al., 2019; Radford et al., 2019), which can be applied to a wide range of language tasks like translating documents; recognising speech; detecting emotion and sentiment; and question answering. This final section introduces some well-known models in NLP and provides an overview of how neuroscientists and linguists have begun adopting these models to study how the brain processes language.

2.3.1 Language Models and Contextualized Word Embeddings

Language has underlying statistics that reflect how it is used in the real world. Humans acquire these statistics during language learning, constructing an internal model which they draw on during language comprehension. Language models attempt to estimate this internal model by training on large collections of written text, known as corpora. These models can be used to compute word probabilities given their preceding word string and have been highly successful in solving a range of difficult language-related tasks. Below, a couple of prominent models are discussed.

A popular model for estimating word sequence probabilities is the n-gram. N-grams calculate the probability of encountering the n^{th} term in a string of words, given the preceding sequence of n-

1 words. The output measure is expressed as surprisal, the negative log probability of a word's occurrence, which indicates how unexpected a word is. N-grams are trained on large text corpora to accurately capture the conditional probabilities of word sequences. Despite being a lot simpler than language models using a deep neural network architecture, n-grams perform surprisingly well in predicting upcoming words (Buck et al., 2014). Previous, influential work has related surprisal to Bayesian belief updating, expressed as the Kullback-Liebler divergence between prior and posterior probability distributions (Levy, 2008).

One caveat with n-gram models is their inability to generalise. Despite, the amount of training data (in many cases trillions of words), there is often a lack of training examples of plausible word strings to correctly inform the model. Optimal models are typically constrained to at most 5-word sequences (5-grams) because the statistics of longer sequences will be poorly represented in a corpus. This is due to the models' inability to generalise across words, treating words as discrete variables without acknowledgement of the meaning often shared between words. A solution to this problem came by thinking of words as distributed representations rather than discrete variables.

Distributional semantic models (DSM) express words as vectors or coordinates in high dimensional semantic space. Words that are close together in this space tend to have similar meaning. These models stem from the distributional hypothesis, which states that the meaning of a word can be inferred by the context in which it occurs (Joos, 1950; Harris, 1954; Firth, 1957). There are several popular methods for estimating word vectors. Global factorisation methods such as Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) or Hyperspace Analogue to Language (HAL) (Lund and Burgess, 1996) utilize the co-occurrence statistics of large text corpora. In these so-called count-based approaches, a matrix of co-occurrence counts is obtained by scanning a large text corpus and counting the times a word occurs in a document (word-document LSA) or co-occurs with other words (word-word, HAL). Dimensionality reduction is then typically applied to these co-occurrence matrices to produce vectors for each word with 100-1000 dimensions. Word to word similarity is then measured by the distance or cosine angle between their vectors. In this framework, words like *cake* and *pie* will have similar vectors because they both tend to co-occur with words like *dessert*, *sweet* and *eat*; and will be dissimilar to words like *hammer* or *alligator* which tend to occur in different contexts. The ability of these models to capture semantic associations between words was clearly demonstrated in tasks relating to text analysis and human judgement (Lund and Burgess, 1996; Foltz et al., 1998).

With increasing popularity in the use of deep learning methods to model language, count-based techniques were eventually outperformed by distributed representations of words learned by neural networks (Baroni et al., 2014), known as local context window methods. These

representations were learned not by counting words but rather predicting them. Specifically, word vectors are learned as part of a neural network, trained to predict the next word in sequence of words. Interestingly, prediction-based representations originally came about not as a means unto themselves but rather as the by-product of another goal – improving a language model performance beyond what had been previously achieved using n-grams (Bengio et al., 2003). Despite word vectors being used in series of downstream tasks (Collobert and Weston, 2008; Collobert et al., 2011) their full utility was not realised until some time after. In 2013, Mikolov and colleagues were able to efficiently estimate vector representations of words using a more simple, single-layer architecture than a full neural network structure (Mikolov et al., 2013b). Skip-gram and continuous-bag-of-words (CBOW) models were trained on a text corpus to either predict the context of a word (in the case of the skip-gram) or the a word itself, given its surrounding context (in the case of CBOW). These models outperformed vector representations from other neural networks in a series of language tasks (Mikolov et al., 2013b, 2013a). Another interesting feature of these vectors is that simple arithmetic operations can be performed on them in a way analogous to human reasoning, for example $\text{Vector}[\text{King}] - \text{Vector}[\text{Man}] + \text{Vector}[\text{Woman}] \sim \text{Vector}[\text{Queen}]$ or $\text{Vector}[\text{Germany}] + \text{Vector}[\text{Capital}] \sim \text{Vector}[\text{Berlin}]$ (Figure 2.9).

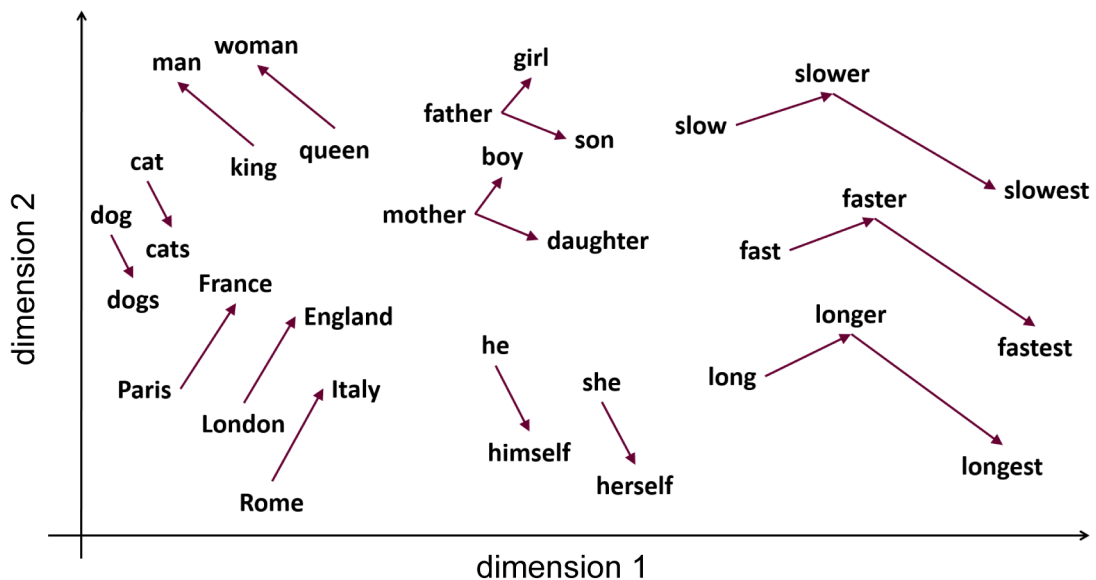


Figure 2.9 Distributional Embedding Space. Word embeddings, estimated using word2vec, projected into 2D space with t-Distributed stochastic embedding. The figure illustrates ability of word embeddings to organize concepts and implicitly learn the relationships between them.

Both global factorisation methods (HAL and LSA) and local context window methods (word2vec) suffer from different drawbacks. For global factorisation methods, high frequency

words like *the*, *with*, or *and* disproportionately influence similarity measures between word vectors despite conveying relatively little semantic information. On the other hand, local context window methods underutilize a training corpus because they train on only local windows and fail to consider the global co-occurrence statistics. Pennington and colleagues proposed a new method for word vector estimation that makes efficient use of corpus statistics. *GloVe* (**G**lobal **V**ectors) is a weighted least squares model that trains on global word-word co-occurrence counts (Pennington et al., 2014). The model benefits from count data and captures the meaningful linear substructures from prediction-based vectors like word2vec. GloVe was found to outperform all other distributional models when tested in word analogy tasks.

2.3.2 Computational Neurolinguistics

The past decade or so has seen a rise in researchers adopting methods from NLP for use in the cognitive neurosciences. This could be due in part to the increased focus in this topic area and the improved accuracy of language models; but it could also be due to the availability of such models. For example, many word embeddings like word2vec, GloVe and fastText (Bojanowski et al., 2017) can now be freely downloaded. There have been several well-regarded successes in applying these techniques to investigate language processing in the brain which has led to the emergence of a new area of focus. Some labs and scientists have begun using the term computational neurolinguistics to formally describe this research area. As depicted in figure 2.10, it sits at the intersection of several research domains.

Word embeddings have been used with fMRI data to investigate semantic processing in the brain. Early work constructed models based on distributed embeddings to predict fMRI responses to isolated words, differing from previous work which had contrasted patterns of fMRI activity associated with groups of words relating to specific semantic categories (Mitchell et al., 2008). This work added support for the part-grounding of concrete words in sensory-motor features (Binder and Desai, 2011; Meteyard et al., 2012). More recent work has extended this framework, using word embeddings to predict the activation of fMRI response related to the processing of words in sentences (Anderson, Binder, et al., 2017; Anderson, Lalor, et al., 2019) and full narratives (Wehbe, Murphy, et al., 2014; Huth, et al., 2016). These studies support a widely distributed network of brain regions involved in semantic processing. Word embeddings have also been used with fMRI to support semantic pattern completion (Kivisaari et al., 2019); the hierarchical processing of language (de Heer et al., 2017); the neural representation of context (Jain and Huth, 2018) and the full meaning of sentences (Schrimpf et al., 2019).

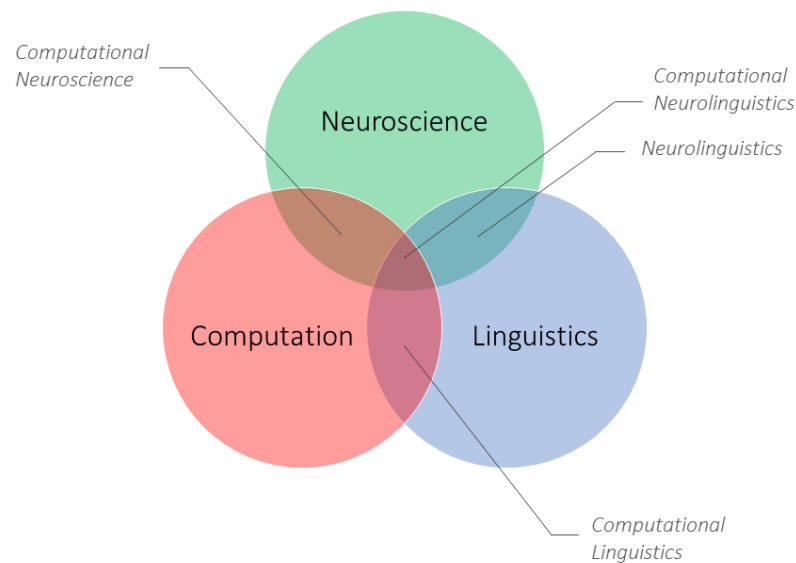


Figure 2.10 Computational Neurolinguistics. Scientists and researchers have begun adopting the term computational neurolinguistics to denote an area of research that sits at the intersection of neuroscience, linguistics and computation.

N-gram derived surprisal measures have been used to model how the brain processes language. Smith and Levy established a quantitative relationship between word expectation and reading times (Smith and Levy, 2013), building on prior work of surprisal theory (Hale, 2001; Levy, 2008). More recently, surprisal measures have been related to neuroimaging data. Frank and colleagues used the measure to predict the amplitude of N400 responses to read words (Frank et al., 2015). Surprisal values were also applied to fMRI data collected from individuals as they listened to spoken narratives. Words with higher surprisal values resulted in increased activation in areas of the temporal lobe (Willems et al., 2016). Despite a handful of studies, there is an overall lack of research investigating how these measures relate to natural speech processing in the brain, particularly using electrophysiology.

Chapter 3. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of continuous narrative speech

3.1 Introduction

People routinely hear and understand speech at rates of 120–200 words per minute (Lieberman et al., 1967; Crystal and House, 1990). Thus, speech comprehension must involve rapid, online neural mechanisms that process words’ meanings in an approximately time-locked fashion. However, electrophysiological evidence for such time-locked processing has been lacking for continuous speech. Although valuable insights into semantic processing have been provided by the N400 component of the event-related potential (Kutas and Hillyard, 1980b; Hagoort and Brown, 2000a; Kutas and Federmeier, 2011), this literature has been dominated by paradigms using incongruous words within specially constructed sentences, with less emphasis on natural, narrative speech comprehension. As discussed in section 2.2, cortical activity encodes the low-level acoustic dynamics of continuous speech (Ahissar et al., 2001; Luo and Poeppel, 2007; Lalor and Foxe, 2010). Furthermore, much psycholinguistic work (Marslen-Wilson, 1973; Simpson, 1991; Tanenhaus et al., 1995) and modelling (Levy, 2008; Pynte et al., 2009; Smith and Levy, 2013; Ettinger et al., 2016) has demonstrated how context impacts on word processing. This chapter introduces a new approach for deriving an electrophysiological correlate of natural speech comprehension.

As reviewed in section 2.3, previous work has used contextualized word embeddings to investigate how semantic knowledge is represented in the brain (Mitchell et al., 2008; Huth, et al., 2016; Anderson, Binder, et al., 2017). Inspired by this work and the brain’s sensitivity to incongruous new words (as seen in the N400; Kutas and Hillyard, 1980; Hagoort and Brown, 2000; Kutas and Federmeier, 2011), we wished to parameterize speech such that individual words

were quantified by their semantic relationship with previous context. Specifically, we measured how “semantically dissimilar” each word in our stimulus was compared to its immediately preceding context. This idea of semantic distance has previously been used in studies of reading-time effects (Pynte et al., 2009), reading comprehension (Frank and Willems, 2017), and brain imaging of speech processing (Frank and Willems, 2017). Our specific approach was based on the well-known word2vec model (Mikolov et al., 2013b), whereby each word in a speech stimulus is converted to a high-dimensional vector. Semantic dissimilarity was calculated for each content word by comparing its vector with the average of the vectors corresponding to all the preceding words in that sentence. Using system identification, semantic dissimilarity was regressed to recorded EEG data from subjects as they listened to narrative speech. This approach was then tested on a range of experimental datasets in which subjects’ speech understanding was manipulated. We investigated the impact of context on the semantic processing of speech under these different experimental conditions. Finally, we compared our measure to the N400 component, derived from the classic ERP paradigm. This work has been published as “Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech”, *Current Biology*, 28, 803–809, March 2018.

3.2 Methods

This study includes data from 5 different experiments, 4 of which had been collected and analysed for different purpose prior to this study. Data for the natural speech experiment and time-reverse speech experiment were collected for (Di Liberto et al., 2015), although data from additional subjects ($n = 9$) were collected for this study. Data for the multisensory speech-in-noise experiment were collected for (Crosse, Di Liberto and Lalor, 2016). Data for the cocktail party attention experiment were collected for (Power et al., 2012) and were analysed further in (O’Sullivan et al., 2015).

3.2.1 Subjects

Subjects in all 5 experiments were native English speakers and reported normal hearing, normal or corrected-to-normal vision, and no history of neurological disease. 19 subjects (13 male) aged between 19 and 38 years participated in the natural speech experiment. A subset of 9 of these subjects (aged 19-32 years, 7 male) participated in an additional N400 experiment. 10 subjects (7 male) aged between 21 and 32 years participated in the time-reversed speech experiment. 5 of these subjects also participated in the natural speech experiment. 21 subjects (15 male) aged between 21 and 35 years participated in the multisensory speech in noise experiment. 34 subjects (28 male) aged between 23 and 38 years participated in the cocktail party attention experiment, but data from one subject was not included in the analysis as the recordings from their mastoid

electrodes were of poor quality. All procedures were undertaken in accordance with the Declaration of Helsinki and were approved by the Ethics Committees of the School of Psychology at Trinity College Dublin, and the Health Sciences Faculty at Trinity College Dublin.

3.2.2 Stimuli and Experimental Procedure

For the natural speech experiment, subjects undertook 20 trials, each of the same length (just under 180 seconds), where they were presented with a professional audio-book version of a popular mid-20th century American work of fiction (*The Old Man and the Sea*, Hemingway, 1952). The trials preserved the storyline, with neither repetitions nor discontinuities. The average speech rate was 210 words/min. The mean length of each content word was 340ms with SD of 127ms. The time-reversed speech experiment involved the presentation of the same trials in the same order, but with each of the 20 speech segments played in reverse. Auditory stimuli for both experiments were presented diotically at a sampling rate of 44.1 kHz using HD650 headphones (Sennheiser) and Presentation software (Neurobehavioural Systems). Testing was performed in a dark, sound-attenuated room, and subjects were instructed to maintain visual fixation on a crosshair centred on the screen for the duration of each trial, and to minimize eye blinking and all other motor activities.

For the multisensory speech in noise experiment, the stimuli were drawn from a set of videos that consisted of a male speaking American English in a conversational-like manner. Fifteen 60-s videos were rendered into 1280 x 720-pixel movies at 30 frames/s and exported in audio-only (A), visual-only (V), and audio-visual (AV) format in VideoPad Video Editor (NCH Software). The soundtracks were mixed with spectrally-matched stationary noise to ensure consistent masking across stimuli (Ding and Simon, 2013; Ding et al., 2014). The noise stimuli were generated in MATLAB (The MathWorks) using a 50th-order forward linear predictive model estimated from the original speech recording. Prediction order was calculated based on the sampling rate of the soundtracks (Parsons, 1987). An SNR level of -9dB was selected based on a previous behavioural pilot study (see Crosse et al., 2016 for details). Subjects were required to detect target words, provided at the beginning of each trial, and to subjectively rate the intelligibility of the speech stimuli at the end of each 60-s trial. Intelligibility was rated as a percentage of the total words understood using a ten-point scale (0–10%, 10–20%, ..., 90–100%). The data analysed here were from the A and AV condition only. The presentation order of A, and AV repetitions was randomized across the 15 videos and across subjects. In addition, intelligibility ratings and not target detection scores were used as behavioural responses in our reanalysis. Visual stimuli were presented at a refresh rate of 60 Hz on a 19-inch CRT monitor and audio stimuli were presented diotically through Sennheiser HD650 headphones at 48 kHz.

Testing was performed in a dark, sound-attenuated room with participants seated 70 cm from the visual display.

In the cocktail party attention experiment, subjects undertook 30 trials, each of ~60 seconds in length, where they were presented with 2 classic works of fiction: one to the left ear, and the other to the right ear. Each story was read by a different male speaker. Subjects were divided into 2 groups of 17 and 16 with each group instructed to attend to the story in either the left or right ear throughout the entire 30 trials. After each trial, subjects were required to answer multiple-choice questions on both stories. Each question had 4 possible answers. A between-subjects design was used so that each subject would follow just one story to make the experiment as natural as possible. For both stories, trials were presented chronologically, with no repetitions or discontinuities. Stimulus amplitudes in each audio stream within each trial were normalized to have the same root mean squared (RMS) intensity. To minimize the possibility of the unattended stream capturing the subjects' attention during silent periods in the attended stream, silent gaps exceeding 0.5 s were truncated to 0.5 s in duration. Stimuli were presented using Sennheiser HD650 headphones and Presentation software from Neurobehavioral Systems (<http://www.neurobs.com>). Subjects were instructed to maintain visual fixation on a crosshair centred on the screen for the duration of each trial, and to minimize eye blinking and all other motor activities. At the end of each trial, subjects were required to answer between four and six multiple choice questions on both stories (i.e. on the attended story and the unattended story). The questions had four possible answers.

In the N400 experiment, 300 sentences were selected from a set of 498 sentences available online whose completions had been normed using a sample of 400 participants (Block and Baldwin, 2010). Half of the sentences were unmodified and ended with a word that was contextually congruent (high cloze probability). For the remaining half, final words were permuted across sentences so that each sentence ended with an incongruent (low cloze probability) word. To reduce the amount of testing time needed, sentences were rephrased to include 5 or 6 words. Keywords relating to sentence meaning and overall sentence coherence were preserved. Both high-cloze and low-cloze sentences had an even amount of 5-word and 6-word sentences (75 each for both cloze conditions). Subjects read the 300 sentences presented in random order. Each sentence was presented word-by-word on a screen with a fixed word duration of 650ms. Subjects were instructed to judge whether the final word was congruent or incongruent with the sentence. Testing was performed in a dark, sound-attenuated room with participants seated 70 cm from the visual display. Subjects who performed the N400 task also performed the natural speech experiment. 4 of the 9 subjects performed the N400 task before the natural speech experiment and the remaining 5 subjects performed it after.

3.2.3 Data Acquisition and Preprocessing

For all experiments, 128-channel EEG data were acquired at a rate of 512 Hz using an ActiveTwo system (BioSemi). Triggers indicating the start of each trial were sent by the stimulus presentation computer and included in the EEG recordings to ensure synchronization. Offline, the data were band-pass filtered between 1 and 8 Hz using a Chebyshev Type II filter (order 54, cutoff 0.5 Hz for high pass filtering and 8.5 Hz for low pass filtering). Passband attenuation was set to 1 dB, and stopband attenuation was set to 60 dB (high pass) and 80 dB (low pass). To identify channels with excessive noise, the standard deviation (SD) of the time series of each channel was compared with that of the surrounding channels. For each trial, a channel was identified as noisy if its SD was 2.5 times the mean SD of all other channels or less than the mean SD of all other channels divided by 2.5. Channels contaminated by noise were recalculated by spline interpolating the surrounding clean channels in EEGLAB (Delorme and Makeig, 2004). Finally, the data was then re-referenced to the average of the mastoid channels.

For the N400 experiment, data was additionally epoched into high cloze and low cloze final words, beginning 100ms before word onset and 800ms after word onset. Epochs were baseline corrected and were rejected if they exceeded 3 times the standard deviation of mean epoch response. Independent component analysis (ICA; Hyvärinen, 1999) was then performed on the epoched trials to removed eye-blinks. The eye-blink component was manually identified and removed for each subject separately. The N400 effect was then determined by subtracting the averaged event-related potential to the congruent words from the averaged event-related potentials to the incongruent words.

3.2.4 Semantic Dissimilarity Calculation

Semantic vectors for content words were derived using the state-of-the-art word2vec algorithm (Mikolov et al., 2013b). The continuous bag of words (CBOW) implementation built in (Baroni et al., 2014) was selected because this was trained on British English corpora (ukWaC, the English Wikipedia and the British National Corpus combined) which is both large and probably more reflective of the language exposure of the participants (in Dublin) than US corpora. In addition, word vectors are freely downloadable (see Baroni et al., 2014). As discussed in section 2.3.1, word2vec embodies the distributional hypothesis (words with similar meaning occur in similar contexts) using an artificial neural network approach. Practically, the approach involves sliding a fixed window of words (11 in this case, however this is a parameter set by the experimenter) over a text corpus and training a neural network to predict the word in the centre of that window. Word identity (as opposed to semantics) is uniquely encoded as a single bit set to one in a long vector of zeros (vector length is the number of words in the vocabulary). These long vectors form the basis of the input and output to the neural network. The input corresponds

to the sum of the 10-word vectors in the window, the output is the central word. Because word order is lost in this summation, the input is analogous to an unordered bag of words. The network contains an internal hidden layer of 400 dimensions. The hidden layer is fully connected to the input and output. It is in fact the weights on the connections between the input and hidden layer that are ultimately harvested to form the semantic model (the weights are a number-of-words in the vocabulary by a 400 floating point matrix) and the remainder of the network is discarded. Weights are initially set as random but are subsequently optimized to reduce error between predicted and target output. Intuitively, because words that frequently appear together in the same context window also predict similar central words, weights on these words are tuned to similar internal representations reflecting common contexts. For more details on the training procedure see (Baroni et al., 2014) and (Mikolov et al., 2013b) (note, the choice of 400 dimensions for the internal layer was arbitrary and, as described in the next paragraph, these 400-dimensional vectors are reduced to a single correlation measure).

Having obtained a vector for each word, we then quantified how semantically dissimilar each word was to the preceding words in the corresponding sentence. We did this by calculating a Pearson's correlation between the word's 400-dimensional vector and the average of the vectors corresponding to all the preceding words in that particular sentence, and subtracting this correlation from 1 (where a specific word was the first word in a sentence, we calculated the correlation between the word and the average of all word vectors in the previous sentence, before, again, subtracting that correlation from 1). It should be noted that, this kind of simple feature-wise averaging/summation of word-level semantic vectors has proven to be an effective and enduring method of modelling semantic composition in computational linguistics (e.g., Mitchell and Lapata, 2010; Kiela and Clark, 2014). It has also been proven to be a successful method for predicting fMRI activation patterns associated with sentences' meanings (Anderson, Binder, et al., 2017). However, it should also be noted that the approach is a gross oversimplification of the complexities of semantic composition in the brain (and does not take into account the effects of word order or syntax – see discussion Anderson et al., 2017). In any case, our approach produced a single semantic dissimilarity measure for each word with a value between 0 and 2. We then created a semantic dissimilarity vector at the same sampling rate as our EEG data (128 Hz) which consisted of time-aligned impulses at the onset of each word that were scaled according to the value of that word's semantic dissimilarity. The word onset times were determined by performing forced alignment of the speech files and the corresponding textual orthographical transcription using the Prosodylab-Aligner, which has been shown to produce alignments with median precision (misalignment) on the order of 10ms (Gorman et al.,

2011).

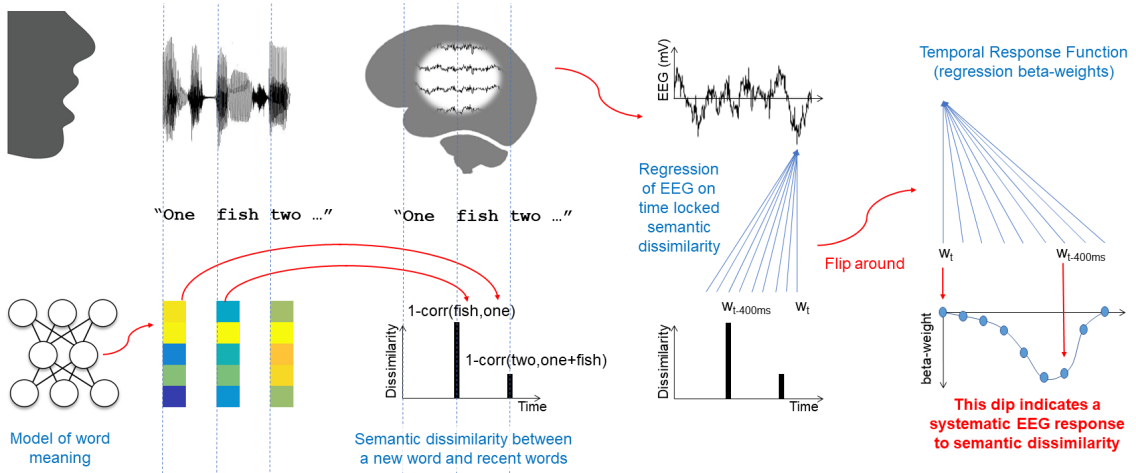


Figure 3.1 Regularized Regression Analysis for Estimating an Electrophysiological Correlate of Semantic Dissimilarity to Natural Speech. Content words from an audiobook are converted to 400-dimensional vectors using the word2vec algorithm (bottom left). The semantic similarity of each word to its preceding context is then defined by comparing (via a Pearson's correlation) its 400-dimensional vector with the average of the vectors of all the preceding words in the corresponding sentence. And the semantic dissimilarity of the word is quantified as 1 minus this correlation (bottom middle left). A vector at the same sampling rate as the recorded neural data is then created that consists of time-aligned impulses at the onset of each word that are scaled according to the value of that word's semantic dissimilarity. The ongoing EEG data are then regressed against this vector to obtain a so-called temporal response function (TRF; right) that describes via weights how fluctuations in semantic dissimilarity across words impact upon the EEG at various time lags

3.2.5 TRF Estimation and Evaluation

A system identification technique was used to compute a channel-specific mapping between the semantic dissimilarity vector and the recorded EEG data, commonly referred to as a temporal response function (TRF; Section 2.2.3). A TRF can be interpreted as a filter that describes the brain's linear transformation of a stimulus feature, $S(t)$, to the continuous neural response, $R(t)$, over a series of specified time lags, in this case 0 to 700ms

$$r(t) = TRF * s(t) \quad (3.1)$$

where '*' represents the convolution operator. Specifically, estimation of the TRF weights was performed using regularized linear regression, wherein a regularization (ridge) parameter, λ , was tuned to control overfitting. In previous work, we have attempted to cast our TRF functions with mV as their unit of measure. However, this relies on a decision to normalize the input stimulus values between some limits and, as such, has been somewhat arbitrary. In the present work, and in line with previous work from other groups, the EEG data on each channel was normalized to zero mean and unit SD (z units) prior to estimating the TRF, meaning that the TRFs are ultimately presented in arbitrary units. The colours in the TRF topographic plots can be interpreted as

follows: red at a particular latency indicates that, at that post-stimulus lag, the EEG voltage is driven in a positive direction by a unit change in semantic dissimilarity; blue means the EEG voltage at that post-stimulus lag is driven negative by a similar change. Thus, given the same normalization strategy for the various speech stimuli used in this study, the TRF responses can be compared in terms of their amplitudes, despite their description in terms of arbitrary units.

In a leave-one-out cross-validation procedure, trained TRFs are used to predict a left-out trial of EEG responses based on the corresponding stimulus. This predicted response is then compared with the recorded neural data to assess EEG prediction accuracy using Pearson's correlation. A range of TRFs were constructed using different λ values between 0.1 and 1000. The λ value corresponding to the TRF that produced the highest EEG prediction accuracy, averaged across trials and channels, was selected as the regularization parameter for all trials per subject.

3.2.6 Statistical Analysis

TRF waveforms were tested as being significantly less than zero using a running one-tailed t-test across subjects. The resulting p-values were corrected using the False Discovery Rate (FDR) method (Benjamini and Hochberg, 1995). Statistical differences between TRF waveforms for forward versus time-reversed speech, audio-only versus audio-visual conditions and attended versus unattended were tested using one-tailed t-tests. The resulting p-values were FDR corrected. To evaluate the improvement in intelligibility from audio to audio-visual conditions via self-report and word detection tasks, we used Wilcoxon signed-rank tests. To quantify effect size (d') in comparing mean TRF responses for audio versus audio-visual speech ($n = 21$) and attended versus unattended speech ($n = 33$) we used Cohen's d . The same measure of effect size was used to compare EEG prediction accuracies. The difference in magnitude of the negative component between the attended and unattended TRFs was supported by statistical testing across subjects using a paired t-test. Correlations between behavioural measures and TRF amplitude and latency were quantified using Pearson's correlation. Correlations between TRF amplitude and N400 amplitude were quantified using Pearson's correlation. Differences in N400 and TRF peak latency distributions were tested using a Wilcoxon signed-rank test.

3.3 Results

3.3.1 A neural correlate of semantic dissimilarity in natural speech

A TRF averaged over 19 subjects who each listened to ~60 minutes of an audiobook is shown in Fig. 3.2A-B. A prominent negativity is apparent over midline parietal scalp at time-lags between 200 and 500ms (Fig. 3.2A). Over this time range, this negativity was significantly less than zero

across subjects at several parietal scalp electrode sites (Fig. 3.2B; running one-tailed t -test, $p < 0.05$, FDR-corrected). To confirm that this negativity was related to the semantic content of the speech and not just the stimulus acoustics, we repeated the analysis using another dataset from 10 subjects who listened to the same audiobook, but in a time-reversed fashion. This produced TRF responses that showed no evidence of the prominent, late negativity (Fig. 3.2A, B; running one-tailed t -test, $p > 0.05$, FDR-corrected). The presence and absence of this negativity for forward and time-reversed speech respectively was evident for several individual subjects who undertook both experiments (Fig. 3.2C). These results demonstrate that electrophysiological responses to natural speech, in the form of a late, parietal negativity, reflect the semantic dissimilarity of individual words to their preceding context in a relatively tightly time-locked fashion.

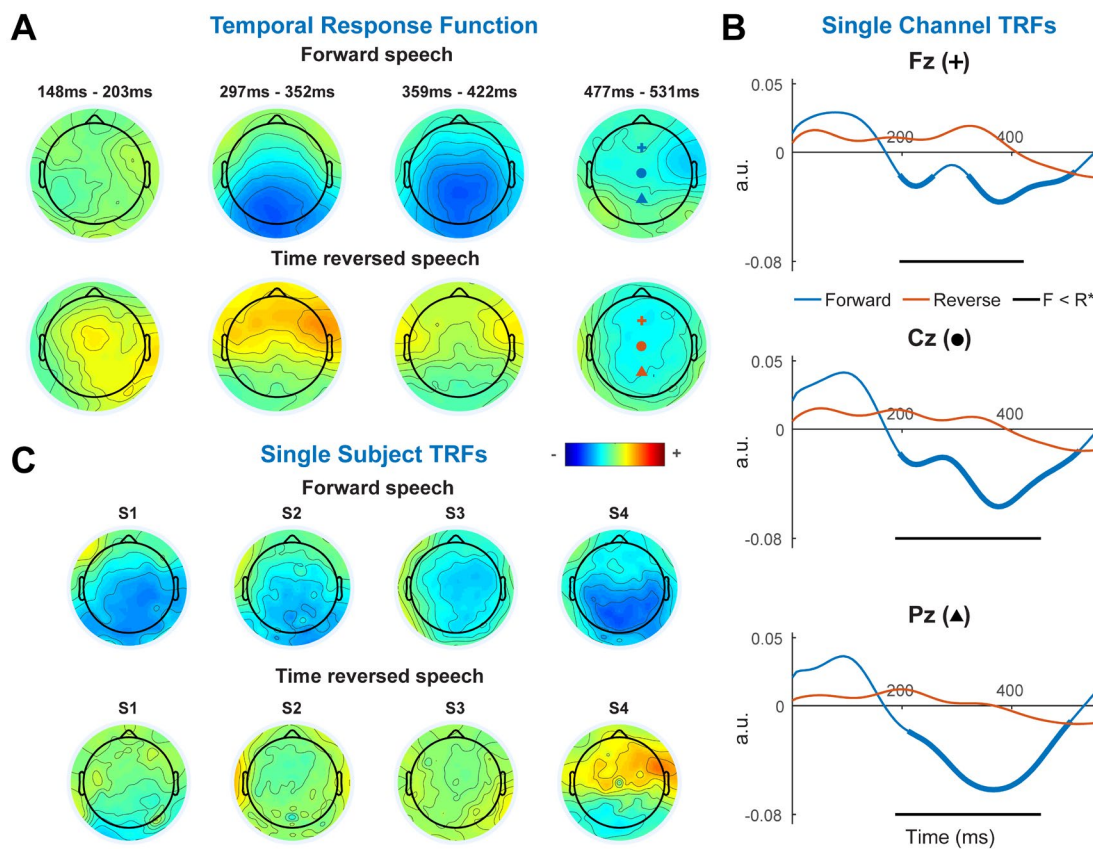


Figure 3.2 Temporal response functions for natural and time-reversed speech. (A) Topographic maps of the semantic dissimilarity TRF averaged over all trials and all subjects for natural, forward speech (top) display a marked centro-parietal negativity between ~200 and 500ms. There is no evidence of a similar negativity in the average TRF for time-reversed speech (bottom). (B) Grand average TRF waveforms at selected individual channels show the time course of the negativity related to semantic dissimilarity. Thick lines indicate a response that is statistically less than zero across subjects ($p < 0.05$, t -test, FDR-corrected). Black lines below the waveforms indicate that the TRFs for forward speech are statistically more negative than those for time-reversed speech across subjects ($p < 0.05$, t -test, FDR-corrected). (C) Topographic maps of TRFs averaged over the interval 200–500ms for selected subjects who took part in both the forward and time-reversed speech experiments. For all four subjects, a negativity is apparent for forward speech (albeit with slightly different distributions for each subject) that is absent for time-reversed speech.

The TRF analysis is based on a linear regression between the stimulus feature (i.e., semantic dissimilarity) and the EEG response. We wished to determine how reasonable this linear assumption was or whether, in fact, our TRF negativity was being driven primarily by occasional “incongruent” words with large semantic dissimilarity values. To do this, we divided the semantic dissimilarity impulses into four quartiles based on their magnitude. We then created four separate regressors corresponding to each of these subsets of impulses, where the impulses within each subset were rescaled to have unit height. The amplitude of the late negativity in the TRFs for these four quartiles showed a monotonic increase with measure of semantic dissimilarity (Figure 3.3). A one-way ANOVA revealed a significant effect of dissimilarity bin on TRF amplitude; $F(3,75) = 5.59$, $p = 0.0017$.

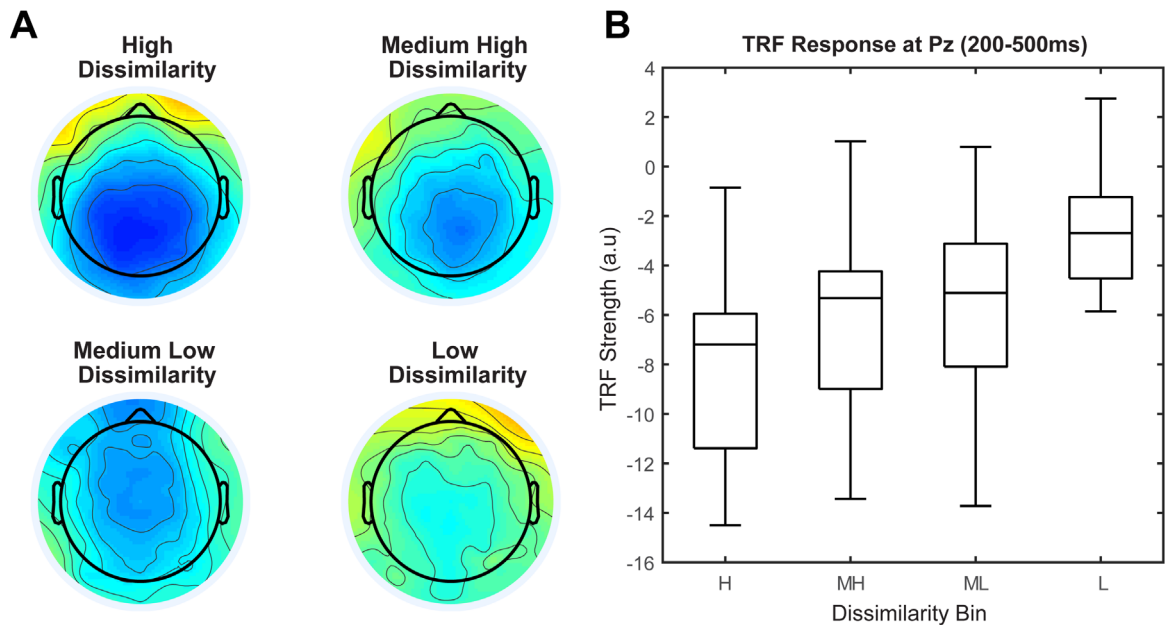


Figure 3.3 The late negativity in the TRF is linearly related to semantic dissimilarity. (A) Regressing the EEG data against each of these regressors produced a TRF featuring a negativity over centro-parietal scalp in the interval of 200-500ms. (B) The amplitude of this negativity across that time interval showed a monotonic increase with the measure of semantic dissimilarity

3.3.2 Neural signatures of semantic dissimilarity depend on intelligibility

The experiments above involved either completely intelligible or completely unintelligible stimuli. To assess how sensitive our semantic dissimilarity TRF might be to gradations of intelligibility we reanalysed data from a multisensory speech-in-noise speech in noise experiment (Crosse, Di Liberto and Lalor, 2016). Specifically, we analysed EEG data from 21 subjects who listened to two repetitions of each of fifteen 60-s segments of continuous audio speech which

were always mixed with spectrally-matched stationary noise at a signal-to-noise ratio of -9dB . Based on the well-known fact that visual speech enhances the intelligibility of speech-in-noise (Sumbly and Pollack, 1954), we manipulated intelligibility by allowing subjects to watch a video of the speaker for one of these repetitions. And while the audio-only speech was not completely unintelligible, the presence of the video led to a large, significant improvement in intelligibility as measured by both self-report ($p = 4 \times 10^{-5}$, Wilcoxon signed-rank test) and a word detection task ($p = 7.9 \times 10^{-5}$, Wilcoxon signed-rank test). This behavioural effect was mirrored by a significant difference in the semantic dissimilarity TRFs between audio-only and audio-visual conditions (Fig. 3.4A, B). This difference was most pronounced at time-lags between 380 and 600ms where it showed an effect size of $d' = 0.55$ (Cohen's d). Notably, this was substantially later than the interval for the TRF negativity during clean speech (Fig. 3.2), a fact that may have to do with the increasing difficulty of processing speech under noisy conditions (see Ding & Simon, 2013; Strauss, Kotz, & Obleser, 2013).

Semantic dissimilarity TRFs were also used to predict unseen EEG. Using cross-validation to fit and test semantic dissimilarity TRFs produced a significantly better EEG prediction for audio-visual speech than audio speech on midline parietal electrode channels (Fig. 3.4C; $p = 0.01$, Wilcoxon signed-rank test). And while the EEG predictions based on audio speech were significantly greater than zero – after all, the audio-alone speech was not completely unintelligible – the effect size of adding the visual input on these EEG predictions scores was large ($d' = 0.84$ on midline parietal electrode Pz). Overall, this demonstrates that our semantic dissimilarity TRF is sensitive to variations in the intelligibility of acoustically identical speech. Moreover, in the audio-visual speech condition, there was a significant negative correlation across subjects between the self-reported intelligibility ratings (which varied broadly) and the amplitude of the TRF negativity averaged over the interval 250–500ms (Fig. 3.4D; the more intelligible, the larger the negativity; $r = -0.5$, $p < 0.02$).

3.3.3 No evidence of contextual semantic processing for unattended speech

Over 60 years ago, it was first noted that, when attending to one of two dichotically presented speech streams, people have a very limited ability to report on the content of the speech in the unattended ear (Cherry, 1953), a phenomenon known as the cocktail party effect. Ever since then, researchers have sought to explain this phenomenon in terms of psychological models (Broadbent, 1958; Deutsch and Deutsch, 1963; Treisman, 1964) and neurophysiological data (Mesgarani and Chang, 2012; Power et al., 2012; Zion Golumbic et al., 2013). Despite these efforts, the extent to which unattended speech is semantically processed by the brain remains unclear (Lachter et al., 2004; Aydelott et al., 2015). However, given the very marked limitations in the ability of subjects to report on the content of unattended speech, we hypothesized that the

negativity in our TRF, as an index of contextual semantic processing, should be markedly reduced in unattended speech.

We derived semantic dissimilarity vectors for each of the two stories, and then regressed the EEG data against these vectors to produce two TRFs - one for the attended story and one for the unattended story. Consistent with previous studies, the behavioural effect of attention was very strong (80% correct answers for the attended story and 29% for the unattended story - chance was 25%). Mirroring this large behavioural effect, the TRF corresponding to the attended story showed a clear and prominent negativity over midline parietal scalp that was significantly larger than the corresponding TRF for the unattended speech, again at a rather long latency of 380–600ms (Figures 3.4E and 3.4F; paired t-test, $p = 9.3 \times 10^{-8}$; effect size $d' = 2.0$). Although this does not entirely rule out some level of semantic processing in unattended speech - after all, our regressors are based on one particular computational measure of linguistic processing - it does present strong evidence of a pronounced reduction in the processing of unattended words relative to their context. Once more, using cross-validation to fit and test a semantic dissimilarity TRF produced a significantly better EEG prediction for the attended speech than the unattended speech (Fig. 3.4G; $P = 9.36 \times 10^{-7}$, Wilcoxon signed-rank test). And while the EEG predictions based on unattended speech were significantly greater than zero – possibly as a result of weak correlations between the semantic dissimilarity impulses and acoustic energy changes at word onsets – the effect size of attention on these EEG prediction scores was as large as that on the TRFs themselves ($d' = 2.0$ at electrode Pz). Unlike for the audiovisual speech-in-noise experiment above, we found no relationship between the amplitude of the TRF and performance on the attended questions. This was unsurprising given that the to-be-attended speech stream was always intelligible. Instead, we found that the peak latency of the TRF negativity was significantly negatively correlated with performance on the questions across subjects (Fig. 3.4H; $r = -0.7$, $P = 1.95 \times 10^{-5}$). In other words, the earlier a subject's TRF peak, the better that subject did on the task. We interpret this as evidence that people who can successfully sustain their attention and/or suppress distracting information can more efficiently process the behaviorally relevant speech – or vice versa. This notion of more efficient semantic processing of words in their recent historical context aligns with the well-known link between working memory and cocktail party attention performance (Conway et al., 2001).

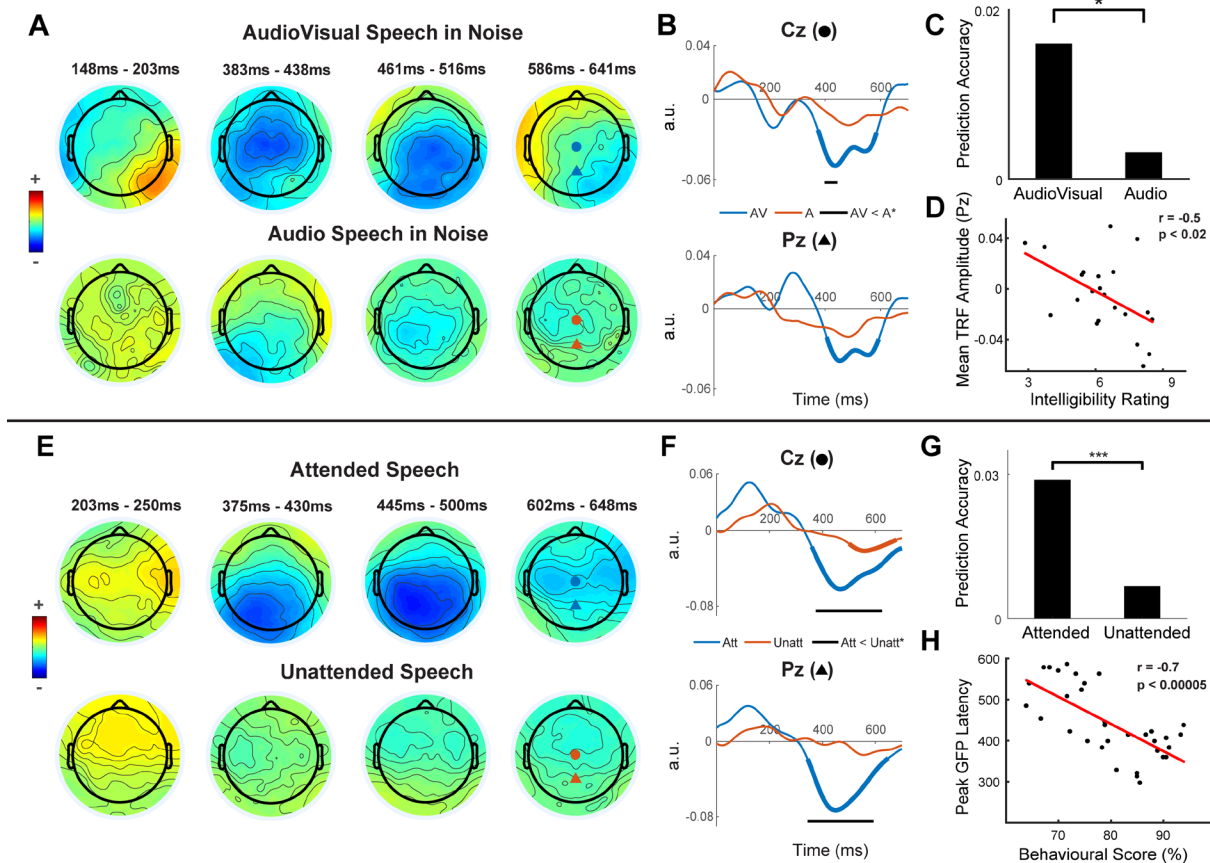


Figure 3.4 Assessing the effect of comprehension on the electrophysiological index of semantic dissimilarity. (A) Topographic maps of the semantic dissimilarity TRF averaged over all trials and all subjects for audiovisual speech in -9 dB of acoustic background noise display a centro-parietal negativity between ~ 400 and 600 ms. This negativity is significantly reduced in the average TRF for audio-only speech in the same level of background noise, which was much less intelligible. **(B)** Grand average TRF waveforms for audiovisual and audio-only speech over two selected midline electrodes. Thick lines indicate a response that is statistically less than zero across subjects ($P < 0.05$, running t -test, FDR-corrected). Black lines below the waveforms indicate that the TRFs for audiovisual speech are statistically more negative than those for audio-only speech across subjects ($P < 0.05$, running t -test, FDR-corrected). **(C)** EEG prediction accuracy for audiovisual speech was significantly greater than that for audio-only speech ($p < 0.01$, t -test). **(D)** Across subjects, the amplitude of the semantic dissimilarity TRF over midline parietal scalp was significantly correlated with self-reported intelligibility rating of audio-visual speech ($p < 0.02$, Pearson's correlation). **(E)** Topographic maps of the semantic dissimilarity TRF averaged over all trials and all subjects for attended speech in a dichotic cocktail party paradigm display a centro-parietal negativity between ~ 300 and 600 ms. This negativity is not apparent in the average TRF for unattended speech. **(F)** Grand average TRF waveforms for attended and unattended speech over two selected midline electrodes. Thick lines indicate a response that is statistically less than zero across subjects ($P < 0.05$, running t -test, FDR-corrected). Black lines below the waveforms indicate that the TRFs for attended speech are statistically more negative than those for unattended speech across subjects ($p < 0.05$, running t -test, FDR corrected). **(G)** EEG prediction accuracy for attended speech was significantly greater than that for unattended speech ($p < 1 \times 10^{-6}$, t -test). **(H)** Across subjects, the latency of the peak in the global field power (GFP) of the semantic dissimilarity TRF was significantly negatively correlated with the number of questions answered correctly on the attended speech ($P < 5 \times 10^{-5}$, Pearson's correlation).

3.3.4 On the relationship of the semantic dissimilarity TRF negativity to the N400 component

The dominant feature of our semantic dissimilarity TRF is a negativity over centro-parietal scalp, a feature shared by the EEG measure that has most strongly been linked with semantic processing of words in context - the N400 component. Although this correspondence is not terribly surprising given that the derivation of our semantic dissimilarity measure was motivated by the fact that the N400 is elicited by words that are semantically incongruent with their context, we have refrained from referring to our response as being the classic N400. The reason for this is that the specific assumptions about semantic dissimilarity that underlie our TRF analysis are not precisely the same as the differences in predictability (cloze procedure, Taylor, 1953) that drive the development of most N400 stimuli (Kutas and Federmeier, 2011). As such, although it is possible that our approach is simply another way to derive the classic N400, it is also possible that the two measures may reflect at least partially dissociable processes. As a first attempt to examine the relationship between the two responses more directly, we compared ERP responses from the N400 experiment with semantic dissimilarity TRF for the 9 subjects who participated in both experiments. The two responses displayed somewhat similar time courses over midline parietal scalp (Figure 3.5A; although the TRF peaked significantly earlier than the N400; $p = 0.012$; Fig. 3.5C), as well as similar topographical distributions at 375 - 425ms (Figure 3.5B). The peak amplitude of the N400 component was also significantly correlated with the peak amplitude of the semantic dissimilarity TRF across the nine subjects ($r = 0.751$, $p = 0.02$; Figure 3.4C).

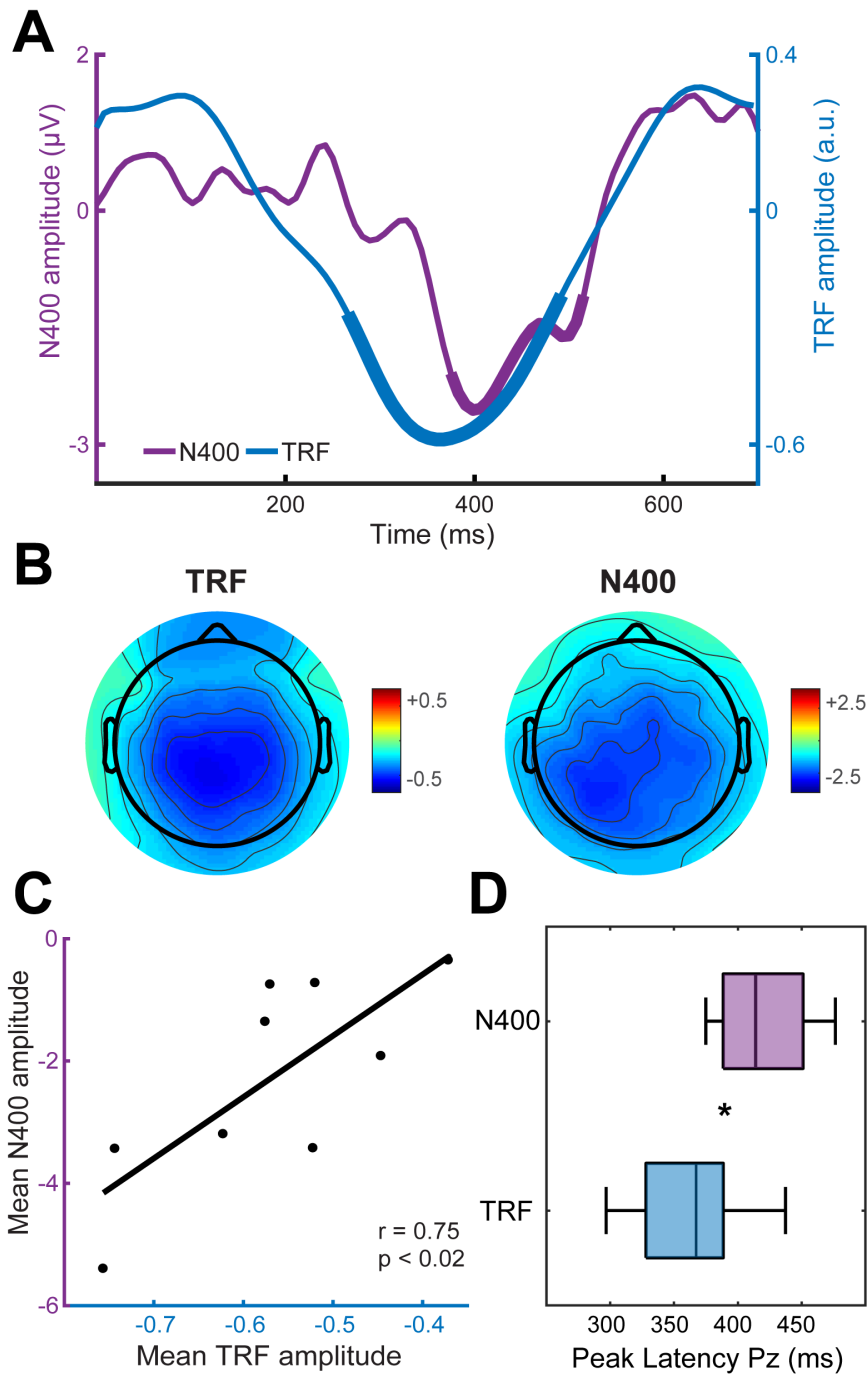


Figure 3.5 Comparison of semantic dissimilarity TRF and the classic N400 event-related potential component. (A) Grand average waveforms from a midline-parietal scalp electrode for the classic N400 component and the semantic dissimilarity TRF. Thick lines indicate a response that is statistically less than zero across subjects ($P < 0.05$, t-test). (B) Topographic maps of both the N400 and the semantic dissimilarity TRF over the interval 375–425ms. (C) The amplitude of the N400 component in the interval 390–450ms was correlated with the amplitude of the semantic dissimilarity TRF in the interval 330–390ms across the 9 subjects ($P < 0.02$, Pearson’s correlation). Importantly, the intervals for the two components were chosen based on the distribution of the peak latency for each response type. Specifically, these intervals represented the 25th to 75th percentiles of those distributions. (D) The latency of the peak of the negativity in the semantic dissimilarity TRF is significantly earlier than the peak of the N400 across subjects (Wilcoxon signed-rank test, $P = 0.0117$).

3.4 Discussion

In this chapter we have shown that, when listening to natural speech, the ongoing dynamics of cortical activity reflect the semantic processing of words in context in a rapid, time-locked fashion. And we have shown that indices of this processing are robustly affected by whether subjects understand the speech they hear and whether they are paying attention to that speech. This approach adds an extra dimension to research on the neural tracking of natural speech dynamics by directly linking a new component of that tracking to the contextual semantic processing of speech. Further work will be necessary to characterize this online semantic processing more fully. This will include investigating whether other types of language knowledge contribute to our measures (Frank and Willems, 2017), assessing whether unattended speech is processed at a lexical-semantic level that depends less upon context than our dissimilarity measure, and modelling semantic representation using alternative neurobiologically motivated approaches (Anderson, Binder, et al., 2017; Anderson, Kiela, et al., 2017). By incorporating other computational models into the framework we have outlined, we would expect that EEG, electrocorticography (ECoG), and magnetoencephalography (MEG) could be very useful in answering these questions.

Gaining deeper insights on these issues will also be helped by optimizing our analysis. For example, it is likely that generating regressors by aligning impulses to word onsets, as we have done, is suboptimal. This is because the time taken to semantically process different words necessarily depends on the words themselves, their context, and listening conditions. That said, the assumption underlying our analysis is that for the successful comprehension of natural speech to be possible at all, meaning must typically be extracted in a relatively online manner - otherwise, we would constantly lose track of what we are hearing. And this assumption appears to be borne out by existence of the robust TRFs we observe. We hope that the framework we have introduced here will allow researchers to estimate more optimal locations for the regressor events and/or regressor events other than impulses. For example, it may be possible to use a reverse approach to discover such optimal regressors (Huth, Lee, et al., 2016).

This work fits in with a line of recent studies employing the TRF which have sought to characterise speech beyond the fluctuations of its acoustic signal, captured by the speech envelope. Motivated by the notion that language processing is hierarchically organised in the brain (section 2.1.1), speech has been represented in terms of its spectrotemporal properties (Ding and Simon, 2012b; Daube et al., 2019); its phonemes and phonetic features (Di Liberto et al., 2015); the transitional probability between phonemes (Brodbeck, Hong, et al., 2018; Di Liberto et al., 2019; Donhauser and Baillet, 2019) and its lexical properties (Brodbeck, Presacco and Simon, 2018). Although, the speech envelope is an important cue for intelligibility, these precise

speech feature estimates exemplify a more optimal approach to address questions about processing at specific linguistic levels. Here, our measure of semantic dissimilarity isolates the semantic processing of words in context and, as a result, we can derive strong behavioural correlates of intelligibility and comprehension (figure 3.4D & 3.4G).

One of the most striking aspects of the semantic dissimilarity TRF is that its morphology, topographical distribution, and peak latency is highly similar to that of the N400 component of the event related potential (Kutas and Hillyard, 1980b). The functional basis of the N400 has been long debated, but one prominent theory that has gained considerable popularity is the idea that the component reflects predictive processing of language (DeLong et al., 2005; Kutas et al., 2011). An influential study by Federmeier and colleagues gave evidence in favour of this, showing how the semantic features of predicted words are preactivated by their context (Federmeier and Kutas, 1999). This has been used to support the notion that brain uses all types of available information from different linguistic levels to predict upcoming words (Kuperberg and Jaeger, 2016). It is plausible that our model of semantic dissimilarity captures this process of preactivation at the level of semantic features. Future work will seek to delineate this measure from lower, more surface-form levels of prediction. It will also be important to more fully examine the relationship between our TRF and the N400. It is entirely possible that they are effectively functionally equivalent. That said, although the amplitude correlations between the two measures were significant, they were decidedly imperfect, and their peak latencies also differed significantly. These differences may be due to the simple fact that the TRF was derived using audio speech, whereas the N400 was elicited using reading. But it also remains possible that the differing assumptions used in producing each measure mean they are reflecting different sub-processes involved in semantic understanding and predictive processing. Future work is needed to answer this question. This includes efforts to determine what sub-processes the TRF negativity represents, similar to efforts involving the N400 (Lau et al., 2008, 2013). Either way, it remains the case that our study has introduced a novel framework for extracting EEG responses that reflect the semantic processing of natural, running speech.

The effects of attention and intelligibility on our electrophysiological measures of semantic similarity closely correlated with the behavioural phenomena in these experiments. And, importantly, these EEG measures were quite robust at the level of individual subjects (Figure 3.2C). This suggests that these indices could be useful in a broad range of basic, applied and clinical research areas. This could include basic research on infant language development, language learning, and the effects of cocktail party attention on speech processing. It could be helpful in arbitrating between computational models of natural speech processing. It could be of benefit to researchers studying language impairments in different clinical cohorts, in clinical

testing of disorders of consciousness, and, given that it relies on relating the processing of words to their recent context, as an assay in people at risk for cognitive and memory decline.

Chapter 4. Neural Markers of Comprehension for Different Levels of Speech Scrambling

4.1 Introduction

In the previous chapter we derived a neural correlate of continuous speech processing that robustly reflected a listener's speech understanding. We investigated the sensitivity of this measure by comparing forward speech with time-reversed speech; by investigating speech processing under different levels of noise; and with tasks where speech could be attended or unattended. We have proposed that this measure reflects the semantic processing of words in context. However, as discussed in chapter 2, the human brain processes speech at multiple linguistic levels, including the analysis of words' phonological, lexical, and syntactic properties (Davis and Johnsrude, 2003; Hickok and Poeppel, 2007; Price, 2010; Poeppel et al., 2012; de Heer et al., 2017) and it's possible that our TRF measure could reflect processing at some level lower than semantics, where words are recognised but their true underlying meaning isn't accessed from or integrated with previous context. The morphology and topographical distribution of the TRF weights shared common characteristics with the N400, a component strongly linked with the processing of word meaning. However, the N400 component is not only sensitive to semantic properties of words in context but can also show sensitivity to lower-level lexical properties like word frequency or neighbourhood density (Kutas, 1993). Therefore, we wished to further examine the measure's sensitivity to the processing of clear speech where all words were lexically identifiable, but the overall meaning of the speech was not understood. In this chapter, our main research question was whether our TRF measure was specifically sensitive to the processing of words following a coherent, predictive context, thus reflecting semantic levels of processing, or whether it could be elicited by the same speech where word order had been manipulated.

To manipulate the coherence of our stimuli while keeping lexical properties consistent, we systematically randomised the order of words for a piece of narrative text to create 5 different levels of randomisation, or scrambling. This text was then converted to a speech signal which we presented to subjects. Similar forms of linguistic degradation have been used in previous studies to investigate syntactic and combinatorial semantic processing (Humphries et al., 2006; Mollica et al., 2018). Inspired by previous studies, where local temporal information in the acoustic (Saberri and Perrott, 1999; Kiss et al., 2008) or linguistic (Lerner et al., 2011) speech stream is manipulated at gradually increasing temporal windows, we varied the size of a temporal window in which words could be scrambled to gradually increase the comprehension difficulty of the speech signal. This created graded levels of speech understanding with which we could test the sensitivity of our newly derived indices of semantic processing.

We additionally wished to test the impact of speech comprehension on neural responses relating to the speech envelope. The amplitude envelope of the speech waveform has been highly emphasised in the literature (Luo and Poeppel, 2007) and is an important cue for speech perception (Shannon, 1951; Drullman et al., 1994b, 1994a). This has led researchers to use neural indices of envelope tracking as dependent measures of speech comprehension (Octave and Reichenbach, 2019; Verschueren et al., 2019). However, reliable envelope tracking has also been shown for an array of signals that do not allow comprehension (Lalor et al., 2009; Howard and Poeppel, 2010; Peña and Melloni, 2012; Doelling and Poeppel, 2015). Fewer studies have compared envelope tracking of speech that is entirely recognisable to the listener but varies in its degree of comprehensibility. Therefore, we wished to examine the suitability of the envelope tracking measure as an index of the higher-level linguistic processing of the speech signal.

4.2 Methods

4.2.1 Subjects

15 subjects (7 female) aged between 19 and 29 participated in the study. All participants were native English speakers, had self-reported normal hearing, were free of neurological diseases, and provided written informed consent. All procedures were undertaken in accordance with the Declaration of Helsinki and were approved by the Ethics Committees of the School of Psychology at Trinity College Dublin, and the Health Sciences Faculty at Trinity College Dublin.

4.2.2 Stimuli and Experimental Procedure

Stimuli were acquired from a children's novel (White, 1951). Text from chapters 3-10 of the novel were split into 60 segments corresponding to the 60 trials in the experiment. Each segment then underwent a scrambling procedure to generate 5 different versions. This was done by

grouping consecutive words into windows of length w and randomly permuting the words within each window. The window lengths were selected as $w = 1$ (unscrambled), 2, 4, 7, 11 so that the interval between window lengths increased by 1 with each w (i.e. $2 - 1 = 1$, $4 - 2 = 2$, $7 - 4 = 3$ etc.). Each segment could be heard as one of the five scrambled versions. Audio versions of each text segment (including the unscrambled text segments) were then generated using Google's Text-to-Speech. This software is powered by WaveNet (Oord et al., 2016), a generative model for raw audio based on a deep neural network. It generates realistic human-like sounding voices and has been rated significantly more natural sounding than the best parametric and concatenative systems for English. Each speech segment, or trial, lasted ~60 seconds.

Two sets of questionnaires were generated for each trial. The first was a comprehension questionnaire that consisted of 6 questions with 2 possible, multiple choice answers, pertaining to the content of the trial. The second was a lexical identification questionnaire. 6 words were presented with yes/no choices as to whether the word appeared in the trial. 2-4 of these words were nouns/verbs which had appeared trial. The remainder were nouns/verbs which were randomly selected from the rest of the text and did not appear in the presented trial. No subject reported listening to or reading the novel within at least 3 years of participating in the study.

For the EEG experiment, trials were presented chronological to the story in 12 blocks of 5 trials. Each block contained the 5 different scrambling conditions. Condition order was randomised for each subject and for each block, so that, for example, subject 1 might hear trial 1 with a scrambling window 11 whereas subject 2 might hear trial 1 with a scrambling window of 4. After each trial subjects were presented with a comprehension and lexical identification questionnaire. Subjects were encouraged to take breaks when needed between trials. Stimuli were presented diotically at a sampling rate of 44.1 kHz using HD650 headphones (Sennheiser) and Presentation software (Neurobehavioural Systems). Testing was performed in a dark, sound-attenuated room, and subjects were instructed to maintain visual fixation on a crosshair centred on the screen for the duration of each trial, and to minimize eye blinking and all other motor activities.

4.2.3 Data Acquisition and Preprocessing

128-channel EEG data were acquired at a rate of 512 Hz using an ActiveTwo system (BioSemi). Offline, the data were downsampled to 128Hz and bandpass filtered between 0.5 and 8Hz using a zero-phase shift Butterworth 4th order filter. To identify channels with excessive noise, the standard deviation of the time series of each channel was compared with that of the surrounding channels. For each trial, a channel was identified as noisy if its standard deviation was more than 2.5 times the mean standard deviation of all other channels or less than the mean standard deviation of all other channels divided by 2.5. Channels contaminated by noise were recalculated by spline interpolating the surrounding clean channels. Independent component analysis (ICA)

(Hyvärinen, 1999) was then performed on the EEG data in order to remove eye blinks. EEG data was transformed into component space and components relating to eye-blinks were identified based on their topographical distribution and component time-series. These components were removed, and the data were then transformed back to EEG space. Data were then referenced to the average of the 2 mastoid channels and were normalized to zero mean and unit SD (z units).

4.2.4 Stimulus Characterisation

We wished to test the effect of scrambling condition on the neural encoding of low-level and high-level properties of the speech signal. For the low-level representation we chose the speech envelope. This is one of the most commonly utilized speech representations with respect to the TRF (Crosse, Di Liberto, Bednar, et al., 2016). It is calculated as the absolute value of the Hilbert transform of the broadband (80 to 3000Hz) speech signal.

We chose semantic dissimilarity as the high-level representation of speech. This was calculated in a similar way to the previous chapter with some slight modifications. First, we used a different word embedding model (GloVe; Pennington et al., 2014) to represent words in high dimensional space. It has been shown that GloVe outperforms word2vec on similarity and named entity recognition benchmark tasks in computational linguistics (Pennington et al., 2014). With respect to the neural encoding of speech, we compared semantic dissimilarity TRFs based on both embedding models and found GloVe also outperforms word2vec in capturing the semantic processing of words in context (Appendix I). Second, semantic dissimilarity had been estimated from the correlation of a word's vector and the average of all preceding word vectors in the same sentence. However, given that, for scrambled speech, sentence boundaries were randomised, we chose to instead estimate semantic dissimilarity by comparing a word vector with the averaged vector of its 10 preceding words. Finally, we included an additional feature of word onset as input to the TRF. This is an impulse vector, with impulses at the beginning of each content word, whose height is a constant value of the average of the semantic dissimilarity values in the same trial. The purpose of this feature was to soak up any variance in the EEG response that related the auditory processing of the word onset.

Speech waveforms for trial and each scrambling condition (60 trials x 5 conditions) were generated first (using WaveNet) before the speech envelope or semantic dissimilarity features were estimated.

4.2.5 TRF Estimation and Evaluation

Channel by channel TRFs were estimated for each subject in the same way as the previous chapter using ridge regression. TRFs were estimated separately for each scrambling condition by grouping trials from each condition (12 trials per condition) together. Separate envelope TRFs

and semantic dissimilarity TRFs were estimated using the envelope and semantic dissimilarity speech representations, respectively. For the semantic dissimilarity TRF, a range of time-lags from 0-700ms was selected. For the envelope TRF, the time-lag range was selected between -100 and 400ms.

TRFs were evaluated using a leave-one-out, cross validation analysis. For each condition and speech representation, TRFs trained on 11 out of 12 trials were used to predict the EEG data of the held-out trial. This was compared with the actual recorded EEG data to obtain prediction accuracy, measured as the Pearson's correlation between the predicted EEG and the recorded EEG. For the semantic dissimilarity TRF some additional steps were implemented in order to directly test how well semantic dissimilarity captured the semantic processing of words relative to their context. First, TRFs were trained and tested on a later window of time-lags, from 200-600ms, rather than the full 0-700ms time-lag window. This was done ensure that EEG prediction accuracy reflected only neural activity relating to higher level linguistic processing. Although the plotted TRFs (figure 4.2) show the entire time-lag window (0-700), prediction accuracies (figure 4.3) were obtained from this narrower window. Second, we wished to establish a more stringent baseline for which EEG prediction accuracies from the semantic dissimilarity TRF could be tested against. For each stimulus trial and scrambling condition, 5 null semantic dissimilarity representations were generated. This was done by randomly permuting the heights of the impulses in the semantic dissimilarity vectors. In the testing phase of the cross-validation procedure, trained TRFs were used to predict EEG based on the 5 null representations for every recorded trial. The predicted EEG of these 5 null representations was then compared the actual EEG with Pearson's correlation. The EEG prediction accuracies of these null representations formed a baseline against which the accuracy of the true representation was compared. Specifically, true EEG prediction accuracy was then calculated by subtracting the mean prediction accuracies from the 5 null representations from the prediction accuracy obtained for the true semantic dissimilarity representation. EEG prediction accuracy for semantic dissimilarity, therefore, refers to the prediction accuracy *difference* between true speech feature and the average of the 5 null speech feature representations.

4.2.6 Statistical Analysis.

For the comprehension and lexical identification questionnaires, chance level was estimated using the binomial inverse cumulative distribution function in MATLAB (Combrisson and Jerbi, 2015), which determines chance level based on the number of trials ($n = 60$) and significance threshold ($\alpha = 0.05$). Comprehension and lexical identification scores and EEG prediction accuracies were compared using one-way ANOVAs. TRF waveforms were tested as being significantly less than zero using a running t-test across subjects. The resulting p values were FDR corrected (Benjamini

and Hochberg, 1995). Average TRF waveforms differences were measured using a running ANOVA, where TRF weights at each time lag were compared across scrambling conditions. This produced F-values as a function of time-lag. P values corresponding to the running F-values were FDR corrected.

4.3 Results

4.3.1 Scrambling window impacts comprehension despite stability of attention

Subjects were required to answer comprehension and lexical identification questionnaires after each experimental trial. The purpose of these questionnaires was to test subject's understanding of the speech segment and attention levels during each trial. Subjects were given multiple choice questions with 2 possible answers to each question. The 95% confidence interval exceeding chance was therefore established as 59.7% using the inverse of the binomial cumulative distribution function. Figure 4.1 shows the comprehension and lexical identification questionnaire scores for each condition. Comprehension accuracy was significantly greater than chance for scrambling windows of 1 (unscrambled; $t_{14} = 13.6$, $p = 8.94 \times 10^{-9}$) and 2 ($t_{14} = 4.2$, $p = 2.3 \times 10^{-3}$), but was not significantly greater than chance for windows of 4 ($t_{14} = 0.78$, $p = 0.54$), 7 ($t_{14} = -0.3$, $p = 0.77$) or 11 ($t_{14} = -0.86$, $p = 0.55$; one-sample t-test, corrected for multiple comparisons). Lexical identification accuracy was significantly greater than chance for all windows (1: $t_{14} = 11.6$, $p = 7.4 \times 10^{-8}$, 2: $t_{14} = 9.1$, $p = 6.6 \times 10^{-7}$, 4: $t_{14} = 7.4$, $p = 3.2 \times 10^{-6}$, 7: $t_{14} = 7.6$, $p = 3.2 \times 10^{-6}$, 11: $t_{14} = 8.9$, $p = 6.6 \times 10^{-7}$; one-sample t-test, corrected for multiple comparisons). These behavioural results suggest that, although subjects failed to understand at higher scrambling windows, they remained engaged with the stimulus and were able to successfully identify words that appeared in the trial.

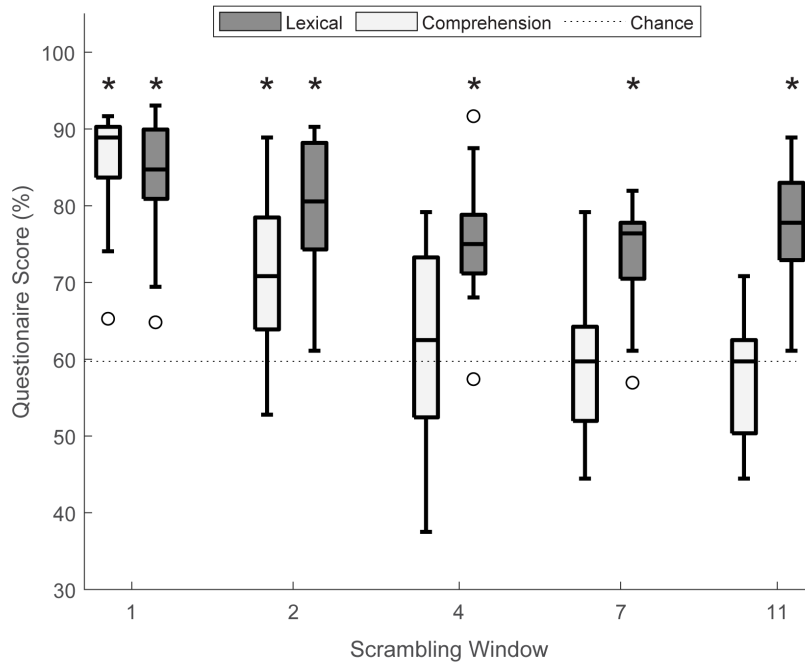


Figure 4.1 Behavioural effects for each scrambling condition. Comprehension (light) and lexical identification (dark) questionnaire scores for each scrambling window. The 95% confidence interval exceeding chance level accuracy is indicated by the dashed horizontal line. Stars indicate the scrambling condition where average subject score was significantly above chance. Lexical identification scores were significantly greater than chance across subjects for all scrambling windows. Conversely only scrambling windows of 1 and 2 had comprehension scores that were significantly above chance across subjects.

Separate one-way ANOVAs were carried out on the questionnaire scores as the questions probed different knowledge about a trial. There was a main effect of scrambling condition on comprehension question score; $F(4, 70) = 20.79$, $p = 2.5 \times 10^{-11}$, one-way ANOVA. Pairwise comparisons using Tukey's Honestly Significant Difference (HSD) procedure indicated that scores for a scrambling window of 1 were significantly higher compared to all other scrambling windows ($p < 0.05$). Scores for a scrambling window of 2 were significantly higher for windows of 7 and 11 ($p < 0.05$) but not for 4 ($p = 0.1$). Comprehension scores for all other windows were not significantly different from each other ($p > 0.05$). There was an additional effect of scrambling condition on lexical identification questionnaire score; $F(4, 70) = 3.6$, $p = 0.01$, one-way ANOVA. Pairwise comparisons revealed significant differences between a scrambling window of 1 and scrambling windows of 4 and 7 ($p < 0.05$). There were no other significant differences between scrambling conditions ($p > 0.05$).

4.3.2 Semantic TRF weights pattern with comprehension

We next tested whether the semantic dissimilarity TRF was sensitive to a subject's understanding of the speech segment they were hearing or whether the measure reflected the more lower level lexical processing of individual words without a coherent context. We first examined the weights

of the TRF over time-lags, averaged over parietal channels. Figure 4.2A-B show the time course of TRF weights. TRFs relating to conditions with less scrambling ($w = 1$ and 2) showed a morphology strongly consistent with TRF responses from the previous chapter. Weights for these scrambling windows were significantly greater than zero across subjects (Fig. 4.2B horizontal line, running one-sample t-test, FDR corrected). A running one-way ANOVA was conducted on each time time-lag, comparing TRF weights across scrambling conditions. Plotted below figure 4.2A is the time series of F-values corresponding each comparison. TRF weights were significantly different across scrambling conditions between 400 and 500ms ($p < 0.01$). However, these F-values did not survive significance testing after correcting for multiple comparisons ($p < 0.1$). The topographical distribution of the TRF weights averaged over later time-lags (300-600ms; Figure 4.2C) further reveal responses that shared similar characteristics of the N400 component. Crucially, this pattern of responses was weaker or absent for TRFs derived from conditions with higher word scrambling. This indicates that semantic dissimilarity TRFs do not simply show responses to any word but rely on the subject's ability to relate a word to its previous context.

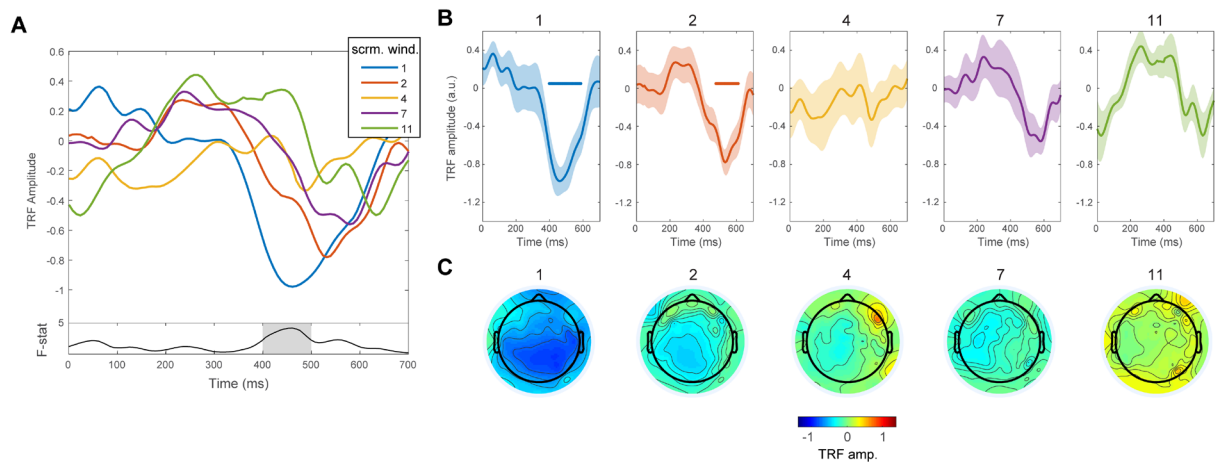


Figure 4.2 Semantic dissimilarity TRFs for speech at each scrambling window (A) TRF weights averaged over parietal electrodes and across subjects for each scrambling condition. Plotted below is the F-values of a running one-way ANOVA, comparing weights across conditions at each individual time-lag. Significant differences across scrambling conditions were observed over time-lags between 400 and 500ms ($p < 0.01$; not corrected for multiple comparisons) (B) The same TRF weights plotted in A in separate plotting windows. Horizontal lines in each window indicate the time points where TRF weights were significantly different from zero (FDR-corrected). (C) Topographical plots show weight values averaged over later time-lags (300-600ms) and across subjects.

Next, we tested how well trained TRFs for each scrambling condition could predict unseen neural responses above a permuted null baseline. Figure 4.3 shows EEG prediction accuracies over parietal electrodes for each scrambling window. Below the box plots are topographical plots showing prediction accuracy, averaged across subjects, for each individual channel. The

unscrambled speech condition was the only one in which trained TRFs could accurately predict EEG above the null baseline ($t_{14} = 3.32$, $p = 0.025$; one-sample t-test, corrected for multiple comparisons). Although the TRF weights for a scrambling window of 2 were significantly greater than zero over later time-lags (Figure 4.2B), they were unable to significantly predict EEG above the null baseline ($t_{14} = 0.67$, $p = 0.71$; one-sample t-test, corrected for multiple comparisons). EEG prediction accuracies for the other scrambling conditions were also non-significant (4: $t_{14} = 0.43$, $p = 0.71$; 7: $t_{14} = 1.7$, $p = 0.26$; 11: $t_{14} = -0.37$, $p = 0.71$; one-sample t-test, corrected for multiple comparisons). A one-way ANOVA was used to test the difference in prediction accuracy between scrambling conditions where a significant effect was found; $F(4, 70) = 3.83$, $p = 0.007$. Pairwise comparisons using Tukey's HSD procedure indicated that prediction accuracy for a scrambling window of 1 was significantly higher compared to a scrambling window of 2, 4 and 11 ($p < 0.05$) but not 7 ($p = 0.17$). There were no other significant differences between scrambling conditions ($p > 0.05$).

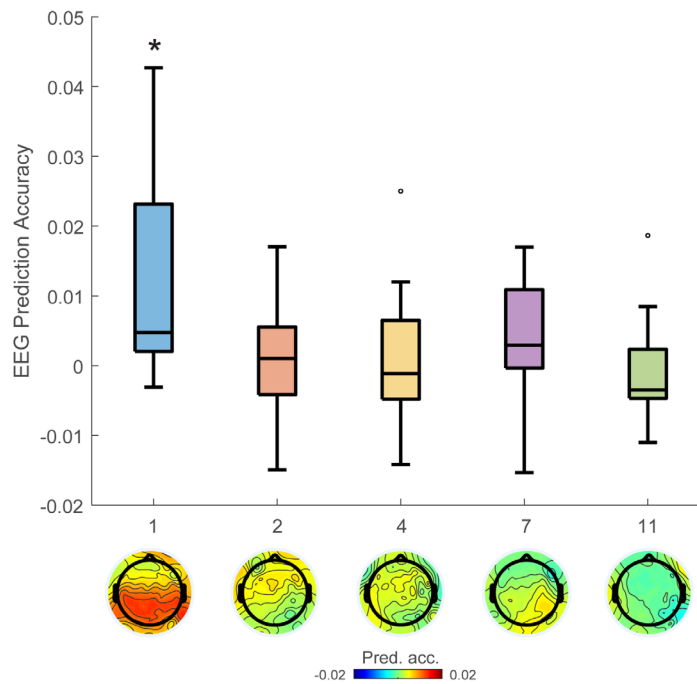


Figure 4.3 EEG prediction accuracy for semantic dissimilarity. Prediction accuracy was estimated as a trained TRFs ability to predict EEG from a true semantic dissimilarity representation over its ability to predict EEG for null semantic dissimilarity representation. EEG prediction accuracies averaged over parietal electrodes are given for each scrambling window (1, 2, 4, 7, 11). Asterix indicates prediction accuracies that were significantly greater than zero. Below are topographical plots of prediction accuracies at each channel averaged across subjects for each scrambling window.

4.3.3 Envelope tracking is unaffected by speech understanding.

We tested whether the envelope-based TRF would show a similar sensitivity to scrambling condition and, hence, also pattern with speech understanding. Figure 4.4A shows the TRF weights

averaged over temporal electrodes and across subjects for each scrambling window. The weights show a morphology characteristic of speech envelope based TRFs from previous studies (Lalor and Foxe, 2010; Di Liberto et al., 2015). Unlike the semantic dissimilarity TRF weights, the morphology of TRF weights for all scrambling conditions shared common characteristics. As before, a running one-way ANOVA was conducted on each time time-lag separately, comparing envelope based TRF weights across scrambling conditions. Plotted below figure 4.4B is the time series of F-values corresponding to each comparison. TRF weights were significantly different across scrambling conditions in a short window between 200 and 220ms ($p < 0.05$). These F-values did not survive significance testing after correcting for multiple comparisons ($p > 0.05$).

We tested each trained TRF's ability to predict unseen EEG. Figure 4.4B shows prediction accuracy, averaged across temporal electrodes. Below the boxplots shows the topographical distribution of prediction accuracies, average across subjects. TRFs from all 5 scrambling conditions were able to significantly predict EEG (1: $t_{14} = 8.9$, $p = 8.2 \times 10^{-7}$, 2: $t_{14} = 8.7$, $p = 8.2 \times 10^{-7}$, 4: $t_{14} = 6.2$, $p = 2.2 \times 10^{-5}$, 7: $t_{14} = 10.9$, $p = 1.6 \times 10^{-7}$, 11: $t_{14} = 8.4$, $p = 9.9 \times 10^{-7}$; one-sample t-test, corrected for multiple comparisons). Differences in prediction accuracy between scrambling conditions were tested using a one-way ANOVA. No significant differences were found for mean prediction accuracy across scrambling conditions; $F(4, 70) = 0.8$, $p = 0.53$. This indicates that, although subjects could not grasp the meaning of the speech they were hearing, their brains still reliably tracked the acoustic properties of the speaker.

EEG prediction accuracies of envelope based TRF were estimated based on a broad window of time-lags (-100 to 400ms). However, we observed small differences between scrambling conditions at peaks in the TRF for later time-lags (200-220ms). Therefore, it is possible that this window could represent some locus of comprehension and, thus, training and testing TRFs in this window might reveal differences in EEG prediction accuracies relating to speech understanding. We trained envelope TRF using a time-lag window centred around the P2 peak (175 to 225ms) and found that this was not the case. Prediction accuracies for all scrambling conditions were reduced but significantly greater than zero ($p < 0.05$; corrected for multiple comparisons). There was no significant effect of scrambling condition on EEG prediction accuracy; $F(4,70) = 1.7$, $p = 0.16$, one-way ANOVA.

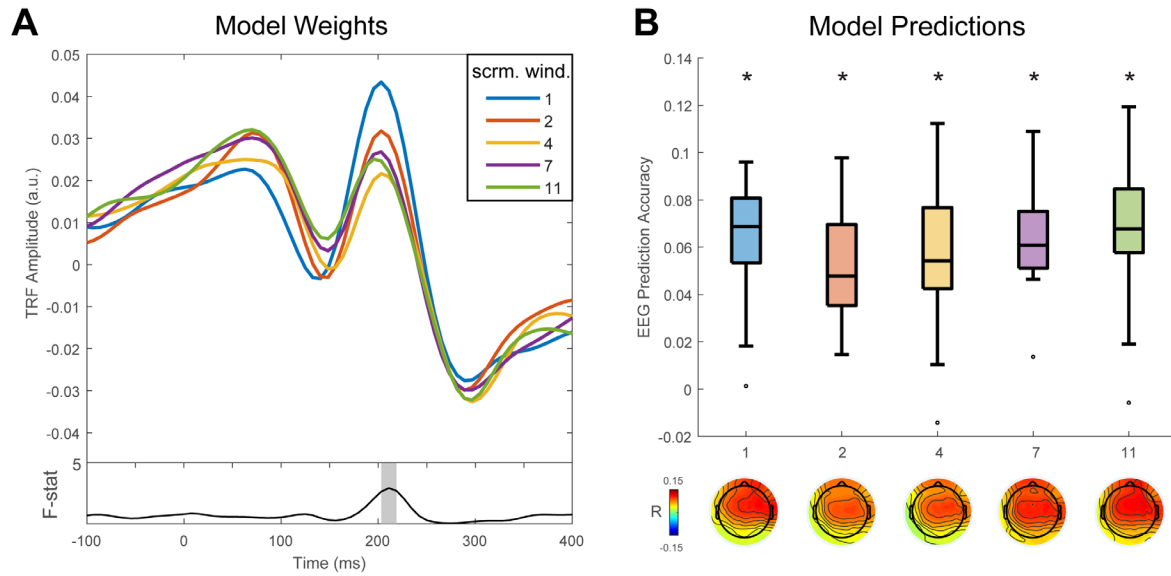


Figure 4.4 Envelope TRFs for speech at each scrambling window. (A) TRF weights averaged over temporal electrodes and across subjects for each scrambling condition. Plotted below is the F-values of a running one-way ANOVA, comparing weights across conditions at each individual time-lag. (B) EEG prediction accuracies for TRF trained on the envelope. Asterix indicates prediction accuracies that were significantly greater than zero. Below are topographical plots of prediction accuracies at each channel averaged across subjects for each scrambling window.

4.3.4 Semantic Dissimilarity TRF predicts comprehension scores

Finally, we wished to test whether the TRF measures could reliably predict an individual's comprehension score for each scrambling condition. As seen above, the averaged weights and prediction accuracies for the semantic dissimilarity TRF indeed reveal responses that patterned with scrambling condition. However, we wished to examine where comprehension scores at the level of individual subjects and individual trials could be predicted using our neural measures. We therefore generated linear mixed effect (LME) models, which took prediction accuracies from the semantic dissimilarity based TRFs and envelope based TRFs as independent variables and comprehension scores as dependent variables, with random effects of individual subjects. One of the LME models predicted comprehension scores averaged across trials for each scrambling condition and the other predicted comprehension scores at the level of individual trials. Predictor variables were normalized before being inputted to the model. Table 4.1 shows the predictor variable coefficients for the first model. There was a significant main effect of average semantic dissimilarity prediction accuracy on comprehension score ($t = 2.93$, $p = 0.005$) and no main effect of average envelope prediction accuracy ($t = -0.76$, $p = 0.44$). Furthermore, we tested the interaction between predictor variables which had no effect of comprehension score ($t = -0.39$, $p = 0.7$)

	Coefficient	S.E.	t statistic	p
Sem. Dissim.	4.62	1.59	2.93	0.005
Env.	- 2.1	2.75	- 0.76	0.44
Sem. Dissim. * Env.	- 0.56	1.45	- 0.39	0.7

Table 4. 1 Subjects LME Model. Model weights for the linear mixed effects model predicting comprehension scores at the level of subjects

Next, we looked at the TRF output measures of individual trials and their ability to predict comprehension success. Table 4.2 predictor variable coefficients. A significant main effect of semantic dissimilarity prediction accuracy was found at the level of individual trials ($t = 2.08$, $p = 0.04$). These results indicate that an individual's performance on the comprehension task overall and at the level of individual trials could be predicted from their neural responses to speech.

	Coefficient	S.E.	t statistic	p
Sem. Dissim.	0.098	0.047	2.08	0.04
Env.	0.063	0.052	1.22	0.22
Sem. Dissim. * Env.	- 0.013	0.044	-0.27	0.77

Table 4. 2 Trial LME model. Model weights for the linear mixed effects model predicting comprehension scores at the level of individual trials.

4.4 Discussion

In this chapter we further examined whether TRFs derived from semantic dissimilarity related specifically to the semantic processing of words following coherent, supportive context. For the unscrambled speech condition, we found reliable responses that were similar to responses derived in the experiments in the previous chapter. Like the previous chapter, components of the TRF to unscrambled speech shared characteristics with the N400 ERP component. These neural indices of speech processing became weaker as soon as any scrambling of the speech signal was introduced. This pattern of neural responses was largely consistent with speech comprehension accuracy where, at higher instances of scrambling, subjects failed to comprehend the overall narrative they were listening to. Lexical identification questionnaire scores were similar across scrambling conditions, allowing us to confidently rule out the possibility of large changes in stimulus engagement across scrambling conditions. Subjects were able to accurately identify words that appeared in a presented trial, indicating that they could recognise the words despite a drop in their overall understanding.

Our findings give further credence to the idea that the semantic dissimilarity derived TRF reflects the semantic processing of words relative to their preceding context. When listening to speech, we incrementally build an internal representation of the current event, discourse, or context (Marslen-Wilson, 1973, 1975; Eberhard et al., 1995; Tanenhaus et al., 1995) that will influence how subsequent information is processed (Oleser, 2014). Such is the basis for language understanding. Our behavioural results outline the importance of a coherent context for the comprehension of words and full speech excerpts. Furthermore, with our methodological framework presented in chapters 3 and 4, we have derived neural correlates that are sensitive to contextual effects on language understanding.

Unlike the semantic dissimilarity TRF, TRFs based on the speech envelope were unable to distinguish between scrambling conditions despite the behavioural difference in speech comprehension across conditions. Previous work has indicated that neural tracking of the speech is foundational for speech comprehension (Pelle and Davis, 2012). However, our findings further highlight the fact that there is not a one-to-one mapping between the two measures (Brodbeck and Simon, 2020). Our results suggest that there is a limit to the stages of higher-level linguistic processing that envelope tracking measure can successfully index. Although they likely reflect the recognition and parsing of individual words in the continuous speech stream (Ding et al., 2015), envelope tracking measures provide little evidence for processes like accessing word meaning or relating it with its previous context, that reflect true comprehension. In contrast, TRFs based on semantic dissimilarity show a closer correspondence with behavioural measures of comprehension. This is unsurprising, given that the measure reflects the relationship between a word's semantic features and the features activated by its previous context.

Previous work examining the relationship between speech envelope tracking and comprehension have found stronger tracking when speech is comprehended (Ahissar et al., 2001; Pelle et al., 2013; Ding et al., 2014; Vanthornhout et al., 2018). However, these studies typically manipulate a listener's understanding by changing the acoustics of the signal. Applying noise vocoding or lower SNR to a signal affects a listener's ability to understand speech but also affects their ability to recognise words (Shannon et al., 1995). Therefore, it has been challenging to dissociate processes of recognition and comprehension using this type of speech manipulation. Of course, neural indices of envelope tracking have provided great insight into how speech is processed from early acoustic stages (Lalor et al., 2009; Ding and Simon, 2013; Ding et al., 2014; Drennan and Lalor, 2019) to stages of lexical identification (Pelle et al., 2013; Ding et al., 2015). However, our results cast doubt on efficacy of such measures as indices for higher-level processes such as comprehension. Researchers should take care when choosing this measure if their main goal is to investigate true speech understanding.

One additional observation for the envelope TRF was the significant (but uncorrected) difference between weights at a time lag of ~200ms. This finding corresponds with results from Power and colleagues who estimated envelope TRF weights for individual's participating in a dichotic listening paradigm (Power et al., 2012). They found a late locus of selective attention at a time-lag of 200ms, where TRF weights for the attended speaker's envelope were more positive than the unattended speaker's envelope. The difference between weights for attended and unattended speakers in this study seems more pronounced than the difference between weights for our scrambling conditions. Furthermore, later studies using the same dataset could effectively distinguish between attended and unattended speakers using the TRF approach (O'Sullivan et al., 2015), whereas our envelope based TRF could not distinguish between scrambling conditions. Nevertheless, it's possible that the differences between the envelope TRF weights across scrambling conditions in this window could reflect a filtering process located at the level of early semantic analysis (Treisman et al., 1964) that is better facilitated by higher contextual support. Future work should more directly investigate the relationship between output TRF measures and the speech representations they are based on.

Chapter 5. Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy for younger and older adults

5.1 Introduction

Healthy ageing is accompanied by a myriad of sensory and cognitive changes. This includes a decline in working memory (Salthouse, 1994) and episodic memory (Burke and Mackay, 1997) as well as hearing loss (Humes et al., 2012) and a slowing in processing across cognitive domains (Salthouse, 1996). Given that spoken language comprehension is a multifaceted cognitive skill involving all these processes, it is remarkable that it remains relatively stable across a healthy adult's lifespan. An interesting question, therefore, is whether the neural systems supporting successful language comprehension undergo a strategic shift with age to maintain preservation in the face of decline (Cabeza et al., 2002; Wingfield and Grossman, 2006; Peelle, 2019), resulting in measurable differences between younger and older adults engaged in comprehension tasks. Furthermore, a related question is whether such differences play into the reported extra difficulties that older adults experience in trying to follow everyday conversational speech, especially in challenging listening environments (Plomp and Mimpen, 1979; Peelle et al., 2010; Presacco et al., 2016).

Electrophysiology studies have indicated that age-related differences exist in the neural signatures relating higher level linguistic processing. This has been shown consistently in studies examining the N400 component (Kutas and Iragui, 1998; Federmeier et al., 2002, 2003). As discussed in section 2.2.2, one potential unifying theory is that this component reflects the stimulus induced change in a multimodal neural network, wherein an implicit and probabilistic representation of

sentence meaning is held (Kutas and Federmeier, 2011; Rabovsky et al., 2018). Importantly, the state of this internal network can be shaped by predictions, such that information can be partially or fully activated before the arrival of bottom-up input (Altmann, 1999). This idea relies on the suggestion that listeners process speech predictively. In particular, it has been suggested that listeners use context to predictively pre-activate information at multiple representational levels during language comprehension (Kuperberg and Jaeger, 2016). At a *lexical* surface level, this could include the activation of representations of word identity (Van Petten et al., 1999b; DeLong et al., 2005), whereas a higher *semantic* level relates to the activation of an upcoming word's semantic features (Federmeier and Kutas, 1999).

Previous cognitive electrophysiology studies have consistently shown age-related differences in the amplitude and latency of the N400 component, indicating that changes do indeed occur in how older adults use context to facilitate the processing of words (Kutas and Iragui, 1998; Federmeier et al., 2003, 2010; Federmeier and Kutas, 2005; DeLong et al., 2012; Wlotko et al., 2012; Payne and Federmeier, 2017). An important study by Federmeier and colleagues found that, for younger adults, N400 responses showed a graded facilitation for words depending on their semantic relationship with some predicted target. For example the sentence “I take my coffee with cream and ...”, showed the largest N400 reduction for expected completions (sugar), but also showed reductions for unexpected but semantically related words (salt) compared to unexpected and unrelated words (socks) (Federmeier and Kutas, 1999; Federmeier et al., 2002). This was taken as strong evidence that the N400 component reflects, in part, the predictive preactivation of semantic features for upcoming words ahead of their bottom-up processing (Federmeier, 2007). Crucially, older adults, as a group, showed weaker facilitation effects for unexpected but semantically related target words, although a cohort of older adults with higher verbal fluency scores showed more younger-like responses (Federmeier et al., 2002).

Several related N400 studies have similarly concluded that prediction plays an overall reduced role in language comprehension for older adults (Wlotko et al., 2010). This strikes somewhat of a discord with findings from cognitive audiology (Pichora-Fuller et al., 1995), behaviour (Burke and Peters, 1986; Stine-Morrow et al., 1999, 2008) and eye-tracking (Kliegl et al., 2004; Rayner et al., 2006; DeDe, 2014) literatures that have, in many cases, shown the preservation of prediction mechanisms or a higher reliance on context in the brains of older adults (recently reviewed by Payne and Silcox, 2019). However, as mentioned above, it is possible that the N400 reflects contributions from predictive processing at multiple linguistic levels. Therefore, the notion that the preactivation of information at distinct levels is differentially affected by ageing has the potential to reconcile these apparently contrasting literatures.

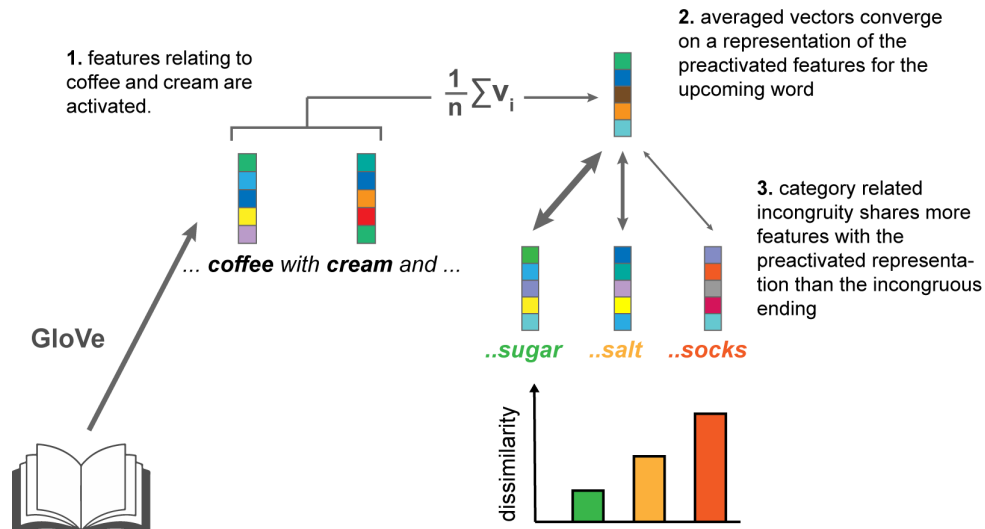
To explore this possibility, we leveraged our developed framework in Chapters 3 and 4 to isolate neural correlates of prediction from these different levels in younger and older adults as they listened to natural, continuous speech. To model predictive processing at the lexical level, we estimated 5-gram surprisal: an information theoretic measure of the inverse of the probability of encountering a word, given the ordered sequence of the 4 preceding words (Buck et al., 2014). The surprisal estimate itself captures nothing about what individual words mean, in the sense that it supplies no measure of whether *cats* and *tigers* are categorically similar, or whether either are domestic. However, word symbol sequences do in part reflect the structure of meaningful events in the real world (*cats* chase *mice*). They also reflect grammatical constraints on permissible symbol sequences (“*The jumped the a cat*” is nonsensical). Lexical surprisal is highly correlated with cloze probability and can predict word reading times (Smith and Levy, 2011, 2013).

In contrast, the semantic dissimilarity measure captures differences in the semantic categories that words belong to, and in the contexts that words appear in (*cats* and *tigers* are both felines, but *tigers* rarely occur in domestic contexts). Distributional modelling approaches like this support the construction of conceptual knowledge hierarchies, e.g., a dragonfly is an insect is an animal (Fu et al., 2014), and would be expected to capture similarities between words belonging to similar categories, such as sugar and salt, and their difference to, say, socks. Thus, semantic dissimilarity would predict a greater N400 for “I take my coffee with cream and *socks*”, than for “I take my coffee with cream and *salt*” (Ettinger et al., 2016). In contrast, a 5-gram surprisal model would likely regard *salt* and *socks* as being equally unexpected at the lexical level because their occurrences are both, presumably, non-existent in a text corpus. So, using this model, one might not expect to see much difference in the N400 for *salt* vs *socks*. (Figure 4.1).

Given the differences in how these language models operate, we hypothesized that they could dissociate predictive processes at lexical and semantic representational levels in terms of how they contribute to the N400. Additionally, based on previous N400 literature (Federmeier et al., 2002), we hypothesized that older participants would show a specific detrimental effect in their ability to preactivate semantic features of upcoming words, and that this effect would correlate with behavioural measures of verbal fluency - for example, the number of words they could produce belonging to a specific category in fixed amount of time.

Example sentence: I take my coffee with cream and *sugar* / *salt* / *socks*

dissimilarity model - preactivation of semantic features



surprisal model - probabilistic prediction

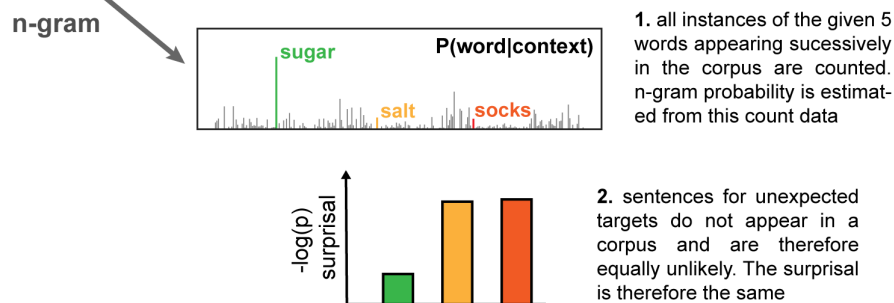


Figure 5.1 Computational models of predictive processing at lexical and semantic level. To illustrate the idea of prediction operating across multiple representational levels, consider the sentence I take my coffee with cream and... which ends with either an expected completion (*sugar*), an unexpected but semantically related completion (*salt*) or an unexpected and semantically unrelated completion (*socks*). At a lexical level, salt is unexpected because it is extremely rare that this sequence of words is heard or read. Processing of this word is therefore assumed to be no different from the processing of other unexpected words (i.e. *socks*). Conversely, at a higher semantic level, salt is relatively more likely, because sugar and salt share common features, both being powders and condiments; edible; white etc. We used two models of lexical surprisal and semantic dissimilarity to disentangle the contributions of prediction at lexical and semantic levels, respectively. Top: For the semantic dissimilarity model, vector representations of previous words in the sentence are averaged to form an estimation of the event context. The latent semantic features of the averaged vector converge on a representation similar to the predicted target “sugar” which, consequently, is more similar to words from the same category (e.g. “salt”) than different categories (e.g. “socks”). Bottom: Conversely, the lexical surprisal model does not distinguish between unexpected words based on their semantic category as it only reflects the probability of encountering either sequence of words in the training corpus, which is either rare or non-existent.

5.2 Methods

5.2.1 Subjects

EEG data from 38 individuals (19 younger (6 female), age 19-38 years, $M=26.8$ years $\pm$ s.d. = 5 years; 19 older (12 female), age 55-77 years, $M = 63.9$ years $\pm$ s.d. = 6.7 years) was reanalysed in the study. Data from the younger subjects that was analysed in chapter 3 was reanalysed. However, in the current analysis, only a portion of each subjects' EEG (the first 12 minutes) was used in order to match the data that was recorded for the older participants. Data collected from the 19 older subjects had not been previously published. Both studies were undertaken in accordance with the Declaration of Helsinki and were approved by the Ethics Committee of the School of Psychology at Trinity College Dublin. Each subject provided written informed consent. Subjects reported no history of hearing impairment or neurological disorder.

5.2.2 Stimuli and Experimental Procedure

The stimuli were taken from the same audio-book as the natural speech experiment in chapter 3. The first 12 minutes of the audiobook was divided into 4 trials, each 3 minutes in duration. Trials were presented chronologically to the story with no repeated trials. All stimuli were presented monophonically at a sampling rate of 44.1 kHz using Sennheiser HD650 headphones and Presentation software from Neurobehavioural Systems. Testing was carried out in a dark, sound attenuated room and subjects were instructed to maintain visual fixation on a crosshair centred on the screen for the duration of each trial, and to minimise eye blinking and all other motor activities. Procedural instructions did not differ between younger and older subject groups.

Older participants were additionally tested with 2 verbal fluency (VF) tasks. Letter verbal fluency was measured by asking participants to name as many words beginning with the letter 'F' as they could in 60 seconds (Borkowski et al., 1967). Similarly, semantic verbal fluency was measured by asking participants to name as many animals as they could in 60 seconds (Benton and Hamsher, 1967). These measures have been shown as reliable indicators of verbal fluency in older adults (Kozora and Cullum, 1995; Bolla et al., 2010). Audio recordings of participant responses were acquired and transcribed verbatim. Average subject letter fluency was 14.63 ± 4.27 S.D. Average subject category fluency score was 20.7 ± 3.75 S.D.

Prior to the verbal fluency task, participants were screened using the Montreal Cognitive Assessment (MOCA). Mean MOCA score was 26.7 ± 1.9 S.D. Verbal fluency is usually assessed as part of MOCA using the letter fluency task. A point is added to the overall MOCA score if a subject correctly names 11 or more words beginning with 'F'. To avoid potential repetition effects, subjects performed the letter fluency task only once and both MOCA and letter fluency scores were obtained from this task.

5.2.3 Data Acquisition and Preprocessing

128-channel EEG data were acquired at a rate of 512 Hz using an ActiveTwo system (BioSemi). Offline, the data were downsampled to 128Hz and bandpass filtered between 0.5 and 8Hz using a zero-phase shift Butterworth 4th order filter. To identify channels with excessive noise, the standard deviation of the time series of each channel was compared with that of the surrounding channels. For each trial, a channel was identified as noisy if its standard deviation was more than 2.5 times the mean standard deviation of all other channels or less than the mean standard deviation of all other channels divided by 2.5. Channels contaminated by noise were recalculated by spline interpolating the surrounding clean channels. Data were then referenced to the average of the 2 mastoid channels.

Finally, we applied multiway canonical component analysis (MCCA) to denoise the data. MCCA is a technique that seeks to extract canonical components across subjects (de Cheveigné et al., 2019). Like canonical component analysis (CCA), which is applied to single subjects, it can be used to find linear components that are correlated between stimulus and response. However, rather than analysing the components directly, EEG can be denoised by projecting it to the overcomplete basis of canonical components, selecting a set of components and then projecting back to EEG space. We denoised each age-group separately, with the prior hypothesis that latency and morphology of the group responses would be different. For each group, we chose parameters of 40 principal components for the initial principal component analysis and then 110 canonical components. These chosen values were based on the parameters that were recommended for denoising speech related EEG (de Cheveigné et al., 2019); however, we tried several different parameter pairs and tested their effect on the prediction accuracy of EEG from the speech envelope. We found that the recommended parameters returned the optimal denoising for both groups as determined by prediction accuracy of EEG based on the speech envelope (Figure 5.4).

5.2.4 Speech Representations

Semantic Dissimilarity was estimated from word embeddings derived using GloVe (Pennington et al., 2014) which was trained on a large corpus (Common Crawl). Like previous chapters, the measure was represented as an impulse vector, the same length as a presented trial, with impulses at the onset of each content word. The height of each impulse is 1 minus the Pearson's correlation between that word's vector and average of all preceding word vectors in the same sentence. Semantic dissimilarity values have a mean of $0.48 \pm \text{s.d.} = 0.17$.

Lexical surprisal was calculated using a Markov model trained on the same corpus as GloVe (common crawl). These models, commonly referred to as n-grams, estimate the conditional probability of the next word in a sequence given the previous n-1 words. We applied a 5-gram model that was produced using interpolated modified Kneser-Ney smoothing (Chen and

Goodman, 1996; Buck et al., 2014). Surprisal vectors were calculated as impulses at the onset of all words whose heights were scaled according to the negative log of a word's 5-gram probability. Like the semantic dissimilarity vector, impulses for any non-content words were removed resulting in the removal of any impulses that were not common between dissimilarity and surprisal vectors. Surprisal values were normalised to match the distribution of dissimilarity values with a mean of $0.48 \pm \text{s.d.} = 0.16$. Lexical surprisal and semantic dissimilarity were weakly correlated (Pearson's $R = 0.14$, $p = 1.2 \times 10^{-5}$, $n = 916$, Fig. 5.2) indicating that they captured distinct features of the stimulus. Figure 5.2 provides example sentences from the experimental stimulus with lexical surprisal and semantic dissimilarity values of the final word.

An *onset vector* was used as a nuisance regressor to capture variance relating to any acoustic onset responses in the EEG signal. This is a vector with impulses at the onset of each content word whose heights are constant and set to the mean dissimilarity and surprisal value (0.48)

The broadband amplitude *envelope* was calculated as the absolute value of the Hilbert transform of the speech signal.

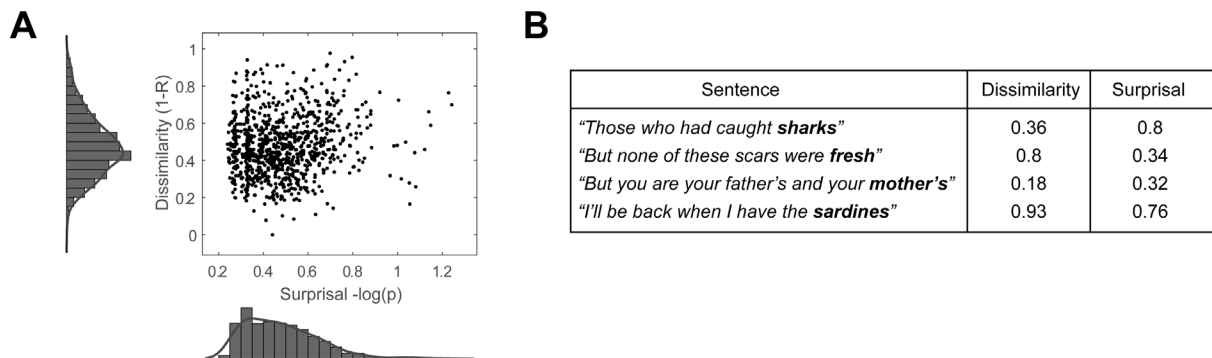


Figure 5.2 Lexical surprisal and semantic dissimilarity values of individual words. (A) Marginal and joint distributions of the normalised surprisal and dissimilarity values for individual words. These values were weakly correlated (Pearson's $R = 0.14$, $p = 1.2 \times 10^{-5}$, $n = 916$). (B) Example sentences with lexical surprisal and semantic dissimilarity values for the final word (indicated in bold) were drawn from the experimental stimulus. Four sentences with combinations of high and low surprisal and high and low dissimilarity word values were selected as examples.

5.2.5 TRF Estimation and Evaluation

Lexical surprisal, semantic dissimilarity and onset vectors were regressed simultaneously to the EEG signal of each individual subject using the same methods as previous chapters but including multiple speech multiple features (mTRF). The models were trained and tested using a 4-fold cross-validation procedure: 3 of the 4 trials were used to train the TRF which predicted the EEG of the remaining trial. As before, a range of TRFs were constructed using different λ values between 0.1 and 1000. The λ value corresponding to the TRF that produced the highest EEG

prediction accuracy, averaged across trials and channels, was selected as the regularisation parameter for all trials per subject.

To test directly how well each feature accurately captured neural activity for each subject we measured the model's ability to predict EEG based on the true feature representation above null feature representations. Specifically, the heights of the impulses for the dissimilarity and surprisal measures were randomly shuffled to produce permuted dissimilarity or permuted surprisal vectors. In the testing phase of the cross-validation procedure, a trained TRF would attempt to predict the neural response to the permuted features, while all other features remained constant. This was repeated for 5 permutations of each stimulus feature. Hence, prediction accuracy for semantic features refers to the prediction accuracy difference between true speech feature and the average of the 5 null speech feature representations.

In addition, we extracted properties of the model weights themselves. N400 peak latency was calculated automatically for each subject as the time lag with the lowest peak weight within a window of 200-600ms after a time-lag of zero. The response peak delay between groups was calculated as the difference between group averaged peak delays.

5.2.6 Statistical Analysis

For every statistical comparison, we first verified whether the distribution of the data violated normality and was outlier free. This was determined using the Anderson-Darling test for normality and 1.5 IQR criterion, respectively. We used parametric tests (t-test, paired t-test) for data which satisfied these constraints and non-parametric tests for data which violated them.

TRF waveforms were tested as being significantly less than zero using a running two-tailed t-test across subjects. The resulting p-values were corrected using the False Discovery Rate (FDR) method (Benjamini and Hochberg, 1995).

5.3 Results

5.3.1 Age-related dissociable neural correlates of lexical surprisal and semantic dissimilarity

Lexical surprisal and semantic dissimilarity were regressed simultaneously to the recorded EEG signal of each individual participant. This produced separate temporal weights for surprisal and dissimilarity. Figure 5.3A shows the surprisal TRF weights for older and younger groups at midline parietal electrodes with scalp weight topographies at selected time windows (300ms, 400ms and 500ms; window width of 50ms) plotted above and below. Both groups show a prominent negative component, characteristic of the classic N400 ERP. We found that the latency

of this component was significantly delayed by 74ms in the older group ($T = 3.5$, $p < 0.005$, 2 sample t-test, Cohen's $d = 1.13$) and observed a correlation between age and response peak latency within the older group (Pearson's $r = 0.46$, $p = 0.047$). Figure 5.3B shows the TRF weights for the semantic dissimilarity feature in older and younger subjects. Younger subjects showed comparable responses for dissimilarity and surprisal feature weights. In contrast, dissimilarity weights were significantly weaker than surprisal weights for older subjects ($p < 0.05$ running paired t-test, FDR corrected).

The performance of a model is also assessed by its ability to predict unseen neural data. Employing a cross validation procedure, we trained TRFs on a subset of each subject's EEG data. This trained model was used to predict EEG that was compared with the remaining, held out EEG data. To test the predictive strength of surprisal and dissimilarity individually, we compared prediction accuracy of the full model (including dissimilarity and surprisal) with 5 null models where either surprisal or dissimilarity values were randomly permuted. Figure 5.3C shows the prediction accuracy (r) of each feature relative to the average null model predictions over midline parietal channels for younger and older subjects. Figure 5.3D shows the topographical distribution of r values. For both groups, surprisal and dissimilarity could significantly predict EEG above this baseline (Younger subjects: $p = 0.0005$ and $p = 0.0011$ for dissimilarity and surprisal, respectively; Wilcoxon signed-rank test. Older subjects: $p = 0.022$ and $p = 0.0002$, for dissimilarity and surprisal, respectively; Wilcoxon signed-rank test). Consistent with the feature weights of the TRF, there were no significant differences between the prediction accuracies for dissimilarity and surprisal for younger subjects ($p = 0.28$, Wilcoxon signed-rank test). However, importantly, older subjects showed significantly higher prediction accuracy for surprisal compared to dissimilarity ($p = 0.0048$, Wilcoxon signed-rank test). Younger subjects also showed significantly higher prediction accuracy for dissimilarity than older subjects ($p = 0.033$, Mann-Whitney U-test), whereas no significant age-related difference was found for lexical surprisal prediction accuracies ($p = 0.91$, Mann-Whitney U-test). 2 of the 19 subjects scored below 25 (scores of 24 and 23) in a prior screening Montreal Cognitive Assessment (MOCA). Although 25 is typically set as the inclusion criteria (Poullisse et al., 2019), we chose to include these subjects in the main analysis because the MOCA is not a direct measure of language function. However, excluding these subjects did not affect the within group difference of semantic dissimilarity and lexical surprisal models in older adults ($p = 0.015$, Wilcoxon signed-rank test) or the between group difference for semantic dissimilarity ($p = 0.043$, Mann-Whitney U-test).

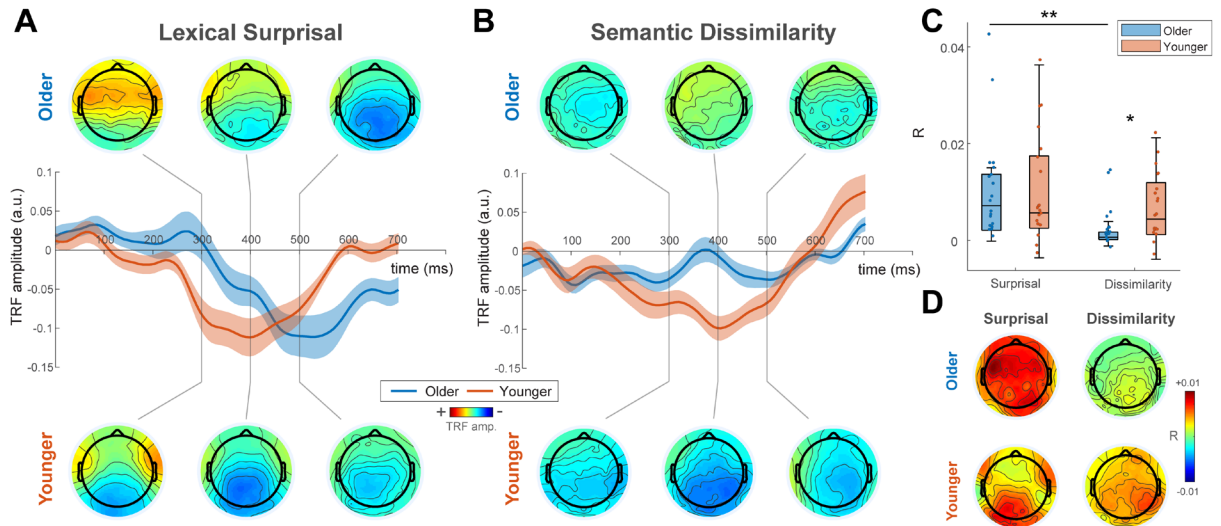


Figure 5.3 TRF weights and prediction accuracies. (A) Lexical surprisal TRF weights averaged over parietal electrodes and across older (blue) and younger (red) subjects. Shaded areas show s.e.m. across subjects. Scalp weight topographies at selected time windows (300ms, 400ms and 500ms; window width of 50ms) plotted above and below the channel plots for older and younger groups, respectively. N400 components were seen in the TRF weights at later time-lags for both groups and the peak latency of this component was significantly delayed for the older group. (B) Semantic dissimilarity TRF weights averaged over parietal electrodes and across older (blue) and younger (red) subjects. Shaded areas show s.e.m. across subjects. Scalp weight topographies at selected time windows (300ms, 400ms and 500ms; window width of 50ms) plotted above and below the channel plots for older and younger groups, respectively. In contrast to lexical surprisal, semantic dissimilarity TRF weights were significantly weaker at later time lags for the older group compared to the younger group. (C) Boxplots of EEG prediction accuracy corresponding to each linguistic feature (lexical surprisal and semantic dissimilarity) and age-group. Dots indicate individual subjects. Consistent with the feature weights of the TRF, there were no significant differences between the prediction accuracies for dissimilarity and surprisal over parietal channels for younger subjects. However, older subjects showed significantly higher prediction accuracy for surprisal compared to dissimilarity at these channels. (D) Topographical plots of prediction accuracy for both age groups and both models.

From these results it is evident that semantic dissimilarity is weaker at explaining the neural responses for older subjects compared to younger subjects. However, this difference in model performance could conceivably be due to the particular way in which we have computed semantic dissimilarity. For instance, it has been shown that older adults have reduced working memory capacity (Salthouse, 1994), and thus for older adults it may be more appropriate to compute dissimilarity using a smaller window of previous words. To safeguard against this possibility, we tested several semantic dissimilarity vectors, where dissimilarity was estimated by comparing a word with a fixed number of previous words. We used context window sizes of 3, 5, 7, 9 and 11 words. We found no differences between models with different context window sizes or the model with a sentence context window ($p = 0.61$ for the older group, $p = 0.79$, for the younger group, Kruskal-Wallis test), indicating that the difference in brain responses between younger and older participants was not the result of the selected parameters.

5.3.2 Stronger envelope tracking for older subjects despite weakened semantic responses

We investigated the low-level acoustic tracking of the speech envelope in both groups to check if our results might be explained based on differences in low-level encoding of the speech signal. Envelope tracking is sensitive to factors like attention and perceived speech intelligibility (Ding and Simon, 2012a; Power et al., 2012; O’Sullivan et al., 2015; Vanthornhout et al., 2018) and weaker tracking measures in the older group could indicate that these processes negatively impact subjects’ ability to preactivate the semantic features of upcoming words. Consistent with previous reports (Presacco et al., 2016; Brodbeck, Presacco, Anderson, et al., 2018), we found significantly *stronger* tracking of the speech envelope in older adults. Figure 5.4A shows the average weights for older and younger participants over bilateral temporal electrodes. Plotted above and below are the topographical distributions of these weights at selected time windows. For both groups, the temporal weights showed a morphology, consistent with previous studies (Di Liberto et al., 2015). However, older subjects showed modulated peak responses relative to the younger group. P1, N1 and P2 components peaks were significantly larger for older subjects ($p < 0.05$ 2 sample t-test, FDR corrected). We also found age-related latency delays of 27ms, 10ms and 17ms for P1, N1 and P2 respectively ($p < 0.001$ 2-sample t-test). Figure 5.4B shows the prediction accuracy (R) of the model for older and younger subjects at each channel (left) and for each subject averaged over 12 bilateral temporal electrodes (right). This shows an increase in prediction accuracy for older compared to younger participants ($T = 3$, $p < 0.01$ 2-sample t-test, effect size (d') = 0.98). These findings are consistent with previous studies, showing stronger tracking of the speech envelope in older adults (Presacco et al., 2016; Brodbeck, Presacco, Anderson, et al., 2018; Zan et al., 2019).

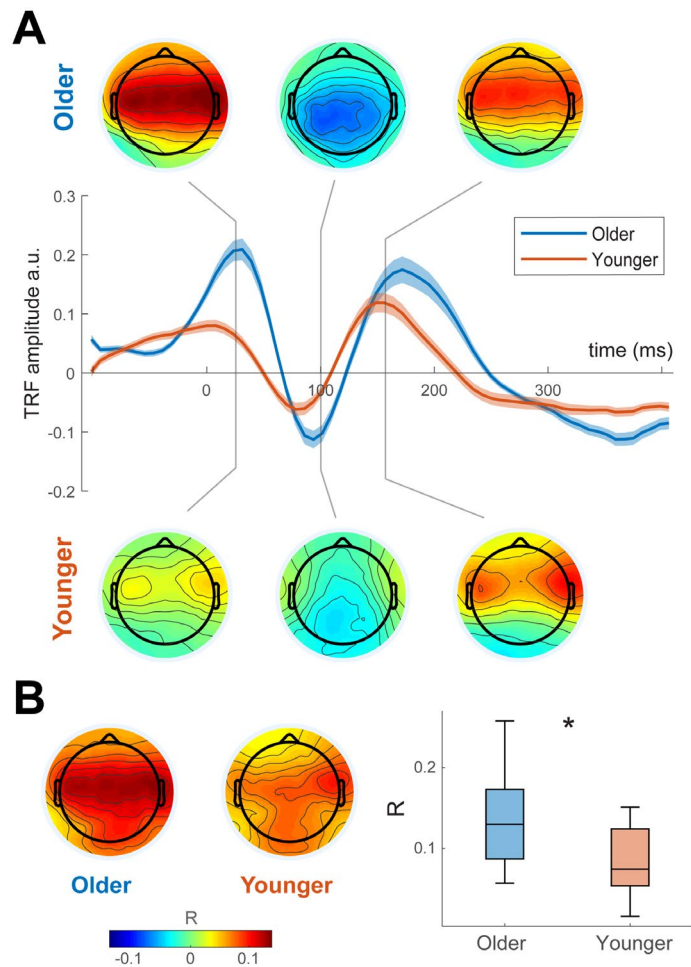


Figure 5.4 Envelope in older and younger subjects. (A) Envelope TRF weights, averaged across subjects and over temporal electrodes. Topographies at selected time windows (~30ms, ~100ms and ~150ms) shows the same morphology of responses for each group. However, the response is stronger for older subjects. (B) The trained models were used to predict unseen EEG in a cross-validation procedure. Older subjects had higher EEG prediction accuracy (R) compared to younger subjects.

5.3.3 Semantic dissimilarity TRF prediction accuracy in correlated with verbal fluency score

Previous work has indicated that older adults with higher verbal fluency scores were more likely to engage predictive processes at the level of semantics, resulting in N400 response patterns that were more similar to their younger counterparts (Federmeier et al., 2002). On this basis, we tested whether semantic dissimilarity model performance could predict verbal fluency in our older subjects. Semantic verbal fluency (VF) scores were collected from all but 2 of the older subjects. We found that, when controlling for age, model prediction accuracies were positively correlated with semantic verbal fluency scores across subjects (Pearson's $R = 0.63$, $p = 0.005$, Fig. 5.5). This reveals that semantic dissimilarity was more accurately modelled for older subjects with higher semantic fluency. The model accuracy for surprisal was not predictive of this measure (Pearson's $R = -0.04$, $p = 0.88$). In addition, model accuracies for surprisal and dissimilarity were

not significantly correlated with letter fluency scores ($R = -0.12$, $p = 0.67$; semantic dissimilarity, $R = -0.44$, $p = 0.07$; lexical surprisal). Although the sample of the correlation was relatively low ($n = 19$) and verbal fluency measures are typically more robust when more than one construct is assessed per measurement (Waters and Caplan, 2003), this significant correlation represents a positive advance in the ability to acquire cognitive estimates from an individual's neural signal under naturalistic listening conditions.

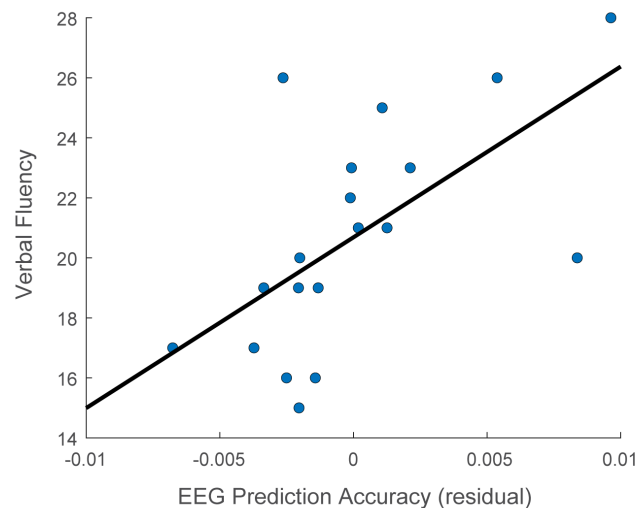


Figure 5.5 Correlation of TRF measures with verbal fluency. Within the older subjects, semantic category fluency was positively correlated with the semantic dissimilarity model performance when controlling for age ($R = 0.63$ $p = 0.005$).

5.4 Discussion

This study has revealed differences between younger and older adults' electrophysiological responses to natural, narrative speech. In both young and old, a joint model capturing lexical surprisal and semantic dissimilarity produced N400 component responses in its temporal weights. While the lexical surprisal measure was robust in older adults, its peak negativity was delayed in a way that is similar to previous reports based on the N400 (Kutas and Iragui, 1998). In contrast, the semantic dissimilarity component was much reduced in the older subjects. We interpret this as evidence for two distinct contributions to the N400 that reflect predictive processing at the level of lexical and semantic representation. Furthermore, the pattern of results suggests that while older subjects maintain a robust ability to utilize lexical predictions during language processing, their ability to do so based on semantic representations appears to be impaired. Importantly, this interpretation was supported by the fact that the performance of the semantic dissimilarity model in older adults reflected a semantic behavioural measure of their categorical verbal fluency. These results extend basic scientific understanding of neurophysiological changes

that accompany ageing and could have implications for research into naturalistic measures of brain health, as we discuss below.

The notion that our measures reflect predictive processing at different lexical-semantic levels fits with results from previous modelling studies on young adults. Surprisal and dissimilarity measures were jointly modelled on both EEG and fMRI responses in younger adults during sentence reading and narrative speech comprehension respectively (Frank and Willems, 2017). The measures both produced similar N400 responses in the sentence reading EEG data, as they have done for the younger adults in our study (Figure 5.3). However, the fMRI results provided evidence that distinct brain regions were involved in processing the two different aspects of the speech input. In particular, visual word form areas reflected surprisal (i.e., lexical processing) and areas of the semantic network (Binder et al., 2009) reflected dissimilarity (i.e., semantic processing). However, in that study, the distinct contributions of lexical and semantic processing were not dissociable in the N400 data because they were both strongly represented in younger adults. Our study goes beyond this, in 1) revealing that electrophysiological activity elicited in narrative speech comprehension reflects components of lexical surprisal and semantic dissimilarity; 2) demonstrating an age-related dissociation between the contribution of surprisal and dissimilarity, with the dissimilarity component being less pronounced in older people. Together, these findings provide convergent evidence that the N400 response reflects contributions from multiple processes relating to prediction at different levels of the linguistic processing hierarchy (Lau et al., 2013; Kuperberg and Jaeger, 2016).

Indeed, the idea that the N400 is affected by prediction at lexical and semantic levels fits with previous N400 literature. N400 responses in younger adults are enhanced for words that are not only unexpected in the context, but also belong to a different semantic category to the expected word (Federmeier and Kutas, 1999). This is consistent with semantic features of upcoming words being predictively preactivated during comprehension. Importantly, the relative contribution of lexical and semantic components to the N400 appears to change with age. Specifically, in older adults there is less sensitivity to an unexpected word's semantic category, especially when the preceding context is highly constraining. These results are consistent with the idea of prediction playing a reduced role at a semantic level in the ageing brain (Federmeier et al., 2002, 2003). However, whereas N400 ERP paradigms propose that different sub processes contribute to the N400 based on how evoked responses vary as a function of sentence-ending, our approach has dissociated lexical surprisal and semantic dissimilarity in neural responses to a continuous stretch of narrative speech. Furthermore, we have uncovered additional evidence that the semantic dissimilarity subprocess, specifically, yields a weaker response in older adulthood. Older adults show remarkably preserved language comprehension skills despite experiencing an overall decline in sensory and cognitive function (Burke and Shafto, 2008). Changes in prediction at a

semantic level may be part of a strategy to successfully comprehend speech with reduced cognitive resources (Manenti et al., 2013). This possibly highlights the putative value of high-level predictions to support speech comprehension in noisy environments, when the input is corrupted and where older adults often struggle to comprehend. Future work, presenting speech at different levels of signal-to-noise ratio could help our understanding of such phenomena.

The notion of preserved predictive language processing in older adults is not without controversy. Eye-tracking, behaviour and cognitive audiology studies have indicated that older adults rely on context-based predictions during spoken language comprehension (Pichora-Fuller et al., 1995; Stine-Morrow et al., 1999; Kliegl et al., 2004; Rayner et al., 2006; DeDe, 2014). In contrast, N400 studies have mostly concluded that older adults make less use of the information available in context (Federmeier and Kutas, 2005; Wlotko et al., 2010). Our study represents the first attempt to model dissociable predictive processes underlying the N400 in older adults using computational language models, with the goal of reconciling these apparently contrasting literatures.

Although Markov (lexical surprisal) models perform well at next word prediction tasks despite being a lot simpler than artificial neural networks, they suffer from the inability to generalize (Bengio et al., 2003, 2006). For example, they are insensitive to the fact that sentences like *I take my coffee with cream and sugar* and *I like my espresso milky and sweet* mean similar things. Thus, they represent a more rigid probabilistic model of language that still works quite well. This model could correspond to a design feature of a lexical system, specialised to use meaningful context to predict upcoming word identity rather than more abstract linguistic information, such as a word's semantic features. This lexical system operates similarly in younger and older adults whose accumulated knowledge of language is more crystalized; less fluid (Stine-Morrow et al., 2008) and reflects the same rigidity as the Markov model. In contrast, ageing results in a lesser engagement in a semantic system that is specialised in the prediction of semantic information.

Hence, our modelling approach suggests that predictive processing is preserved with age, aligning with findings from eye tracking, behaviour, and cognitive audiology literatures. However, the information that is preactivated represents a more rigid commitment to the identity of upcoming words. The preactivation of linguistic information that exists at semantic levels of representation that can generalize across words is much weaker in the brains of older listeners, although subjects with higher verbal fluency scores exhibited stronger preactivation. This could explain why word reading time context effects are preserved with age, as word reading times correlate with lexical surprisal, (Smith and Levy, 2013) and not with semantic dissimilarity when surprisal is factored out (Frank et al., 2017). In contrast, both semantic dissimilarity and lexical surprisal can explain variance in N400 responses when controlling for one another (Frank and Willems, 2017). Based

on the notion that the N400 reflects contributions from multiple distinct linguistic levels, this could explain why weaker N400 responses are consistently observed in older adults (Kutas and Iragui, 1998; Federmeier et al., 2003; Wlotko et al., 2010, 2012).

We wish to further emphasize that the study was undertaken using a short, 12-minute segment of natural continuous speech stimulus. Previous research into the electrophysiological changes in language processing in the ageing brain have leveraged ERP-based experimental protocols that rely on experimenter-configured stimulus sets to enable contrasts between different stimulus conditions (e.g. congruent and incongruent sentence wordings). However, the ERP approach constrains the breadth of linguistic stimuli that can be investigated to the subset of sentences configured into matched experimental pairs. Additionally, the degree of ecological validity of results generated from bespoke ERP setups is unclear because the experimental conditions are rarely experienced in everyday life. By examining electrophysiological responses elicited in audiobook comprehension, we have utilized a stimulus that is actually experienced in the wild, and the participant experiences an uninterrupted prolonged and cohesive discourse, that is likely to be more engaging than listening to disjoint experimental sentences.

In conclusion, we have revealed neural correlates of language prediction relating to distinct measures of lexical surprisal and semantic dissimilarity. We show how one of these forms of prediction becomes less effective with age and patterns with behavioural cognitive measures, enabling us to derive accurate estimates an individual's verbal semantic fluency from their neural data alone. These findings open new possibilities to study language impairment in the elderly and detect the onset of neurodegenerative disorders.

Chapter 6. Top-down effects of linguistic context on the early auditory encoding of natural speech

6.1 Introduction

As reviewed in chapter 2, it is widely believed that the cortical processing of spoken language is underpinned by hierarchical system that processes speech at various stages of abstraction. However, there is ongoing debate over *whether* context-based predictions of upcoming words, made at higher levels, feed-back to affect processing at lower-levels (McClelland and Elman, 1986; Norris et al., 2000) and *how* such neural processes are computationally realised (de Lange et al., 2018). Electrophysiology can significantly contribute to answering such questions by providing direct measures of how the brain responds to predictive speech (Kutas et al., 2011). Its high temporal resolution allows for the investigation of the precise timings of cognitive operations underlying perception. In addition, computational linguistics has given us a battery of methods for quantifying how upcoming words may be predicted from their context (Elman, 1990; Chen and Goodman, 1996; Jurafsky, 1996; Hochreiter and Schmidhuber, 1997; Roark, 2001; Bengio et al., 2003; Buck et al., 2014). Combining these methodologies has the potential to uncover how spoken language perception unfolds in the human brain.

Cortical activity has been shown to encode or ‘track’ the temporal envelope of speech (Ahissar et al., 2001) and is reflective of both the acoustic (Aiken and Picton, 2008) and phonetic processing (Di Liberto et al., 2015; Khalighinejad et al., 2017) of speech. Recent studies have also shown how low frequency cortical encoding to the speech increases as a function of speech intelligibility (Luo and Poeppel, 2007; Ding et al., 2014; Crosse, Di Liberto and Lalor, 2016). This indicates that cortical tracking of the speech envelope isn’t merely a passive following of the acoustic signal but can be influenced by some top down input. It remains unclear, however, what is driving this modulation. It is well known that selective attention modulates the cortical tracking of speech, as seen in studies using the ‘Cocktail Party’ paradigm (Ding and Simon,

2012a; Power et al., 2012; Zion Golumbic et al., 2013; O’Sullivan et al., 2015). Other factors such as listening effort (Peelle, 2018) have also been shown to affect tracking ability. Another possibility is that this increased speech encoding could be partially due to the amount of semantic context and word predictability made available with increasing speech intelligibility. It has been shown that semantic information can improve perceived intelligibility of speech in noise, when comparing coherent sentences with unrelated word lists (Miller et al., 1951), and coherent with incoherent sentences (Davis et al., 2011). Semantic context can also bias the perception of phonemes replaced with noise which has been shown in behavioural (Warren, 1970) and neurophysiological studies (Leonard et al., 2016). Such studies point to a top-down, semantic related effect on the auditory tracking of the speech signal, but whether this occurs at pre-lexical stage remains unclear.

An additional question is precisely how bottom-up sensory inputs combine with top-down prior knowledge to subserve perception. Popular theories propose that this is achieved through hierarchical Bayesian inference whereby the brain infers the causes of its sensory inputs by constructing an internal model of the world that generates predictions which interact with incoming sensory input (Knill and Pouget, 2004). As discussed in section 2.2.1, one prominent mechanistic account is known as predictive coding (Rao and Ballard, 1999; Friston, 2005; Clark, 2013). A core tenet of this theory assumes the involvement of two distinct populations of neurons: representation units, which encode predictions based on prior information, and error units, which encode the discrepancy between bottom-up signals with the top-down predictions (Rao and Ballard, 1999). Given this assumption, one should expect the co-presence of neural signals reflecting computations from these distinct populations during perception. Indeed, while some evidence exists (Egner et al., 2010), there is an overall lack of neuroimaging studies providing direct empirical support for the simultaneous computation of prediction and prediction error. Thus, there is still ongoing debate as to how prior information informs perception (Egner and Summerfield, 2013; Heilbron and Chait, 2017; Friston, 2018).

In this study we sought to first investigate whether higher level semantic information impacts the early auditory encoding of speech. Like previous chapters, we used embedding models to quantify how well a word is expected from its context. We then investigated the effects of prediction and prediction error on same responses to speech. This done using lexico-semantic measures derived from distinct computational language models, introduced in chapter 5.

6.2 Methods

The present study is based on a reanalysis of the dataset collected for the natural speech experiment in Chapter 3. Please refer to Section 3.2 for information on the participants, stimuli and EEG experimental paradigm.

6.2.1 Data Preprocessing

EEG data were bandpass filtered between 1 and 8Hz using a Chebyshev Type II filter (order 54, cutoff 0.5Hz for high pass filtering and 8.5Hz for low pass filtering). Passband attenuation was set to 1dB and stopband attenuation was set to 60dB (high pass) and 80dB (low pass). The data was later filtered in delta and theta bands using 1-4Hz (order 34, cutoff 0.5Hz and 4.5Hz) and 4-8Hz (order 36, cutoff 3.5Hz and 8.5Hz) Chebyshev Type II bandpass filters. These filters had the same passband and stopband attenuation as the 1 to 8Hz filter. After filtering, data were downsampled to 64 Hz (backward modelling) or 128 Hz (forward modelling; see below). To identify channels with excessive noise, the standard deviation of the time series of each channel was compared with that of the surrounding channels. For each trial, a channel was identified as noisy if its standard deviation was more than 2.5 times the mean standard deviation of all other channels or less than the mean standard deviation of all other channels divided by 2.5. Channels contaminated by noise were recalculated by spline interpolating the surrounding clean channels in EEGLAB (Delorme and Makeig, 2004). Finally, the data were re-referenced to the global average of all channels.

6.2.2 Stimulus Characterisation

A speech signal can be represented in several different forms. Each characterisation can be thought to index different aspects of the same speech signal and capture unique features from different levels of the language hierarchy. In our methods we extract a variety of acoustic and linguistic features from each content word in our spoken narrative and use these features as inputs at various points in a 2-stage regression approach. These features can be categorised into 3 main groups based on their role and what stage they are used at in the analysis.

1. The first group of features were used as input to the TRF, mapping stimulus to neural response or as output to the decoder neural response to stimulus. These characteristics are used to assess the encoding of speech at different hierarchical levels.
 - 1a. *Envelope*. The broadband amplitude envelope of the speech signal was calculated as the absolute value of the Hilbert transform.
 - 1b. *Spectrogram*. The speech signal was filtered into 16 different frequency bands between 250Hz and 8kHz according to Greenwood's equations (Greenwood, 1961). After filtering, the amplitude envelope was calculated for each band.

- 1c. *Phonetic Features*: To create the phonetic feature stimulus Prosodylab-Aligner software (Gorman et al., 2011) was used. This automatically partitions each word in the story into phonemes from the American English International Phonetic Alphabet and performs forced-alignment, returning the starting and ending time-points for each phoneme. Each phoneme was then mapped to a corresponding set of 19 phonetic features, which was based on the University of Iowa's phonetics project <http://prosodylab.cs.mcgill.ca/tools/aligner/>.
2. The second group indexes higher-level linguistic features of words and was used in the second stage of the regression analysis
 - 2a. *Semantic similarity* was calculated for each content word in the narrative, as before using word embedding models. Word2vec vectors (Mikolov et al., 2013b; Baroni et al., 2014) were used in the initial analysis. Semantic similarity was calculated as 1 divided by the Euclidean distance between the word's vector and the averaged vector of all the preceding words in the sentence. Thus, small similarity values signify out-of-context words that by extension are unexpected.

Semantic similarity values had a mean of 0.347 and standard deviation of 0.072. To put these values into perspective we estimated distributions of semantic similarity values for sentences that were purposefully constructed to reach the upper and lower bounds of semantic similarity. We took example sentence pairs from an N400 study designed by Federmeier (Federmeier and Kutas, 1999), and calculated semantic similarity values for each word in the second sentence relative to the first. This distribution of values was taken to be the upper bound. We then replaced these sentences with randomly selected sentences from the TIMIT dataset (Garofolo, 1993) to simulate a second distribution, and lower bound, of semantic similarity values. We found that the distribution of semantic similarity values derived from our natural text lies between the upper (mean = 0.367, standard deviation = 0.069) and lower (mean = 0.3, standard deviation = 0.048) distributions of semantic similarity values.

Semantic similarity was later analysed with lexical surprisal. For this model, GloVe (Pennington et al., 2014) vectors were used to calculate semantic dissimilarity as the vectors themselves had been trained on the same corpus as the lexical surprisal model (Common Crawl). Thus, in order to ensure that any dissociable effects were not due to differences in text corpora, this embedding model was chosen. A word's semantic similarity was estimated as the Pearson's correlation between the word's vector and the averaged vector of all the preceding words in the sentence. Semantic similarity values were highly correlated between word2vec and GloVe models (Pearson's $r = 0.75$, $n = 4542$)

- 2b. *Lexical surprisal* was calculated using a Markov model trained on the same corpus as GloVe (common crawl). These models, commonly referred to as n-grams, estimate the conditional probability of the next word in a sequence given the previous n-1 words. We applied a 5-gram model that was produced using interpolated modified Kneser-Ney smoothing (Chen and Goodman, 1996; Buck et al., 2014). Opposite to semantic similarity, lexical surprisal assigns higher values to more unexpected words.
3. Finally, a third group of stimulus characterisations was identified. The purpose of this group was to be included in the 2nd stage regression analysis as “nuisance regressors” to soak up any variance in word reconstruction accuracy related to acoustic changes in the speaker’s voice (detailed discussion below).
 - 3a. *Envelope Variability* was calculated by taking the standard deviation of the speech envelope over the same time window used to measure word reconstruction accuracy. Here, we wished to control for rapid changes in the envelope amplitude as it has been shown that cortical responses monotonically increase with steeper acoustic edges (Oganian and Chang, 2018).
 - 3b. *Relative Pitch* was recently shown to be encoded in EEG (Teoh and Lalor, 2019). It quantifies pitch normalised according to the vocal range of the speaker. Praat software (Boersma and Weenink, 2000) was used to extract a continuous measure of pitch (absolute pitch). The measure was then normalized to zero mean and unit standard deviation (z-units) to obtain relative pitch.
 - 3c. *Resolvability* measures whether the harmonics of a sound can be processed within distinct filters of the cochlea (resolved) or if they interact within the same filter (unresolved). It has previously been shown using fMRI that pitch responses in auditory cortex are predominantly driven by resolved frequency components (Norman-Haignere et al., 2013). Custom written scripts from an acoustic statistics toolbox from the same study were used to extract a continuous measure of harmonic resolvability.

6.2.3 Modelling of the Stimulus-Response Relationship

Our analysis framework consists of 2-stage regression. The first stage models the relationship between audio stimulus and neural response, thus quantifying how well certain features of speech were encoded in the neural signal. This was done using linear regression in which 2 separate models were trained: a backwards model which reconstructs the stimulus from the neural response and a forward model which predicts neural responses from the stimulus. Both approaches offer complementary advantages (Crosse, Di Liberto, Bednar, et al., 2016). The backward modelling approach takes into account the signal from all recorded electrodes

simultaneously and thus can do a better job at separating the speech-related signal from the background EEG. Given that we did not have a good sense *a priori* of how large the effect of semantic information on low-level processing might be, we chose to first exploit the sensitivity of this backward modelling approach to maximize our chances of identifying such an effect. On the other hand, forward modelling gives channel specific predictions and can indicate scalp regions where the stimulus is most prominently represented. It can also more easily incorporate different multivariate speech representations (e.g., spectrograms, phonetic features). As such, we also sought to use the forward modelling approach with a view to identify how different speech features might be differentially affected by semantic information and how any such effects might be distributed on the scalp.

A backwards model, or decoder was trained to reconstruct an estimate of the speech envelope from the neural data (Mesgarani et al., 2009; Ding and Simon, 2012b). The decoder, $g_n(\tau)$, represents the linear mapping from neural response, $r_n(t)$, to stimulus, $s(t)$, expressed by the following equation

$$\hat{s}(t) = \sum_n \sum_{\tau} r_n(t + \tau) g_n(\tau)$$

where $\hat{s}(t)$ is the reconstructed stimulus envelope; n represents the recorded neural channels and τ is a specified number of time lags ranging from -100 to 300ms. Decoder weights are estimated using regularized linear regression, wherein a regularization (ridge) parameter, λ , was tuned to control overfitting. The λ parameters is selected by constructing a range of decoders with different λ values between 0.1 and 1000. Using leave-one-out cross-validation, decoders are trained on 19 out of 20 trials to reconstruct the envelope of the missing trial, which is then compared with the actual envelope to obtain reconstruction accuracy. This is repeated for each trial and the λ value corresponding to the decoder that produced the highest envelope reconstruction accuracy, averaged across trials, was selected as the regularisation parameter for all trials per subject (Figure 6.1; Stage 1).

The forward modelling of stimulus to response was done using the Temporal Response Function (TRF) (Lalor and Foxe, 2010). For N recorded channels the instantaneous neural response $r(t, n)$, sampled at times $t=1 \dots T$ and channel n , consists of a convolution of the stimulus property, $s(t)$, with a channel specific TRF, $w(\tau, n)$. The response can be modelled as

$$r(t, n) = \sum_{\tau} w(\tau, n) s(t - \tau) + \varepsilon(t, n)$$

Time lags also ranged from -100 to 300ms. TRF weights are also estimated using regularized linear regression. However, rather than reconstruct an estimate of the stimulus, forward models

predict EEG responses at each electrode. The λ parameter was also selected using the same cross-validation procedure (Figure 6.1, Stage 1).

Three separate forward TRFs were used based on representing the speech input as its spectrogram (spectrogram-only), a discrete set of time-aligned phonetic features (phonetic features-only) and as a combination of its spectrogram and phonetic features. These models were used to generate three separate sets of EEG predictions that reflect cortical responses to the spectral and phonological properties of speech.

The brain's encoding of individual words, ρ_w , was evaluated in two ways: 1) using a trained decoder to reconstruct the speech envelope of each word in a held-out trial and compare it with the actual word's speech envelope and 2) using a trained TRF to predict the channel-specific EEG response to each word in a held-out trial and compare it with the actual EEG response at each channel. The comparison was done by measuring the Spearman's correlation between the two signals for the first 100ms after word onset.

$$\rho_w = \begin{cases} \text{corr}(s_w, \hat{s}_w) \dots \text{backward model} \\ \text{corr}(r_w, \hat{r}_w) \dots \text{forward model} \end{cases}$$

Thus, encoding of individual word was measured by their reconstruction/prediction accuracy. This was repeated for every trial in the dataset resulting in reconstruction/prediction accuracies for all content words in the experiment. There were, on average, 2 phonemes in every 100ms window.

6.2.4 Modelling the relationship between semantic information and speech encoding

The first goal of the present study was to assess how the abovementioned speech encoding measures of words (envelope reconstruction and response prediction accuracy based on spectral and phonetic features) vary as a function of the semantic information associated with them. To that end, the strength of relationship between stimulus reconstruction or response prediction accuracy and semantic similarity to context was measured using a 2nd stage linear regression model with each content word in the narrative as a datapoint to the model and with associated reconstruction/prediction accuracy and semantic similarity values as dependent and predictor variables, respectively. We used a linear mixed effects (LME) model which models variability due to items and subjects simultaneously (Figure 6.1, Stage 2). The model is described below.

$$r_w \sim \beta_{sem} * sem + \beta_{env} * env + \beta_{frel} * frel + \beta_{resolv} * resolv + (1|subject) + \varepsilon$$

In addition to semantic similarity (*sem*), low level acoustic correlates of pronunciation, measured in the first 100ms after word onset, were included in the model. These were envelope variability

(*env*), average relative pitch (f_{rel}) and average resolvability (*resolv*). The roles of these variables were to act as nuisance regressors to ensure no underlying acoustic related confound existed between our two main measures. It has been shown that semantic information associated with a word will affect how it is spoken (Lieberman, 1963). Therefore, it was important to ensure that any effect of semantic information on cortical tracking was not in fact an indirect one, by proxy of changes in the speaker's voice rather than top-down effects in the listener's brain. We studied the model fit, R^2 , and semantic similarity features weight, β_{sem} , to determine whether there was some relationship between the variables. All variables were normalized to zero mean and unit standard deviation (z-units) before being input to the model. Pairwise correlations for the predictor variables in the 2nd stage regression were calculated using Pearson's correlation and are given in Table 6.1.

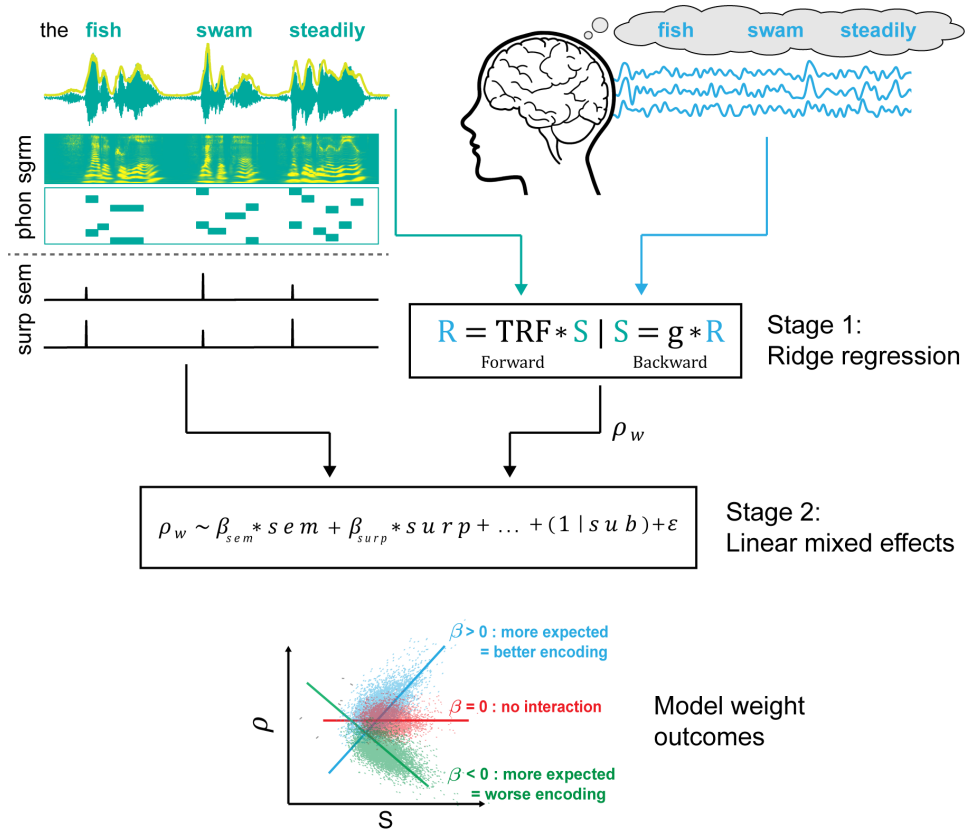


Figure 6.1 Analysis framework. A 2-stage regression model was used to evaluate the impact of lexico-semantic features on the neural encoding of spoken words. Stage 1 uses ridge regression to measure the encoding of individual words. A decoder was used to reconstruct the speech envelope of each word which was then compared with the envelope of the actual spoken word. The temporal response function (TRF) based on the speech spectrogram (sgram) or phonetic features (phon) was used to predict EEG response to individual words. Response prediction were compared with recorded EEG response for each word. Word reconstruction/prediction accuracies (ρ) were passed to the 2nd stage of the model which measured the relationship between semantic similarity (sem) and feature encoding. The linear mixed effects model coefficient β quantified this relationship. Positive or negative β values would suggest that semantic context influences the cortical encoding of words. A β value of zero would suggest no interaction between the measures.

	Semantic Similarity	Envelope Deviation	f_{rel}	Resolvability
Semantic Similarity	1			
Envelope Variability	0.038*	1		
Relative Pitch	0.0011	0.295*	1	
Resolvability	-0.029	0.236*	0.386*	1

Table 6.1 Variable Correlations Pairwise correlations between predictors variable in 2nd stage regression. Asterisks indicate significant correlation after correcting for multiple comparisons.

For our forward modelling approach, EEG prediction accuracies, derived from the combined spectrogram and phonetic features TRF, were used as the dependent variable in the 2nd stage linear mixed effects model, with identical predictor variables as the 2nd stage backwards model. To investigate the effect of semantic context on isolated measures of acoustic and phonological processing, two additional 2nd stage models were constructed. These models were the same as the 2nd stage forward model but included prediction accuracies from the spectrogram-only TRF or the phonetic features-only TRF as additional nuisance regressors for each other. The rationale for including these nuisance regressors was that phonological feature and spectrogram representations of speech share redundant information. So, for example, including EEG prediction accuracies from the spectrogram TRF allowed us to partial out that redundant contribution and identify the effects of semantic information on phonetic feature processing in isolation. The same analysis was carried out to identify effects of semantic context on spectrographic processing in isolation also.

In the second analysis, in order to investigate the precise mechanisms of linguistic prediction, we looked at the impact of distinct lexico-semantic features on the encoding of individual words. This was done using a similar LME model but with lexical-surprisal (*surp*) included as an additional predictor variable.

$$\rho_w \sim \beta_{sem} * sem + \beta_{surp} * surp + \beta_{env} * env + \beta_{f_{rel}} * f_{rel} + \beta_{resolv} * resolv + (1|subject) + \varepsilon$$

Pairwise correlations for predictor variables in the 2nd stage regression (analysis 2) are given in table 6.2.

	Semantic Similarity	Lexical Surprisal	Envelope Deviation	f_{rel}	Resolvability
Semantic Similarity	1				
Lexical Surprisal	-0.12*	1			
Envelope Variability	0.065*	0.159*	1		
Relative Pitch	0.02	0.17*	0.29*	1	
Resolvability	-0.061*	0.04*	0.22*	0.248*	1

Table 6.2 Variable Correlations (Analysis 2) Pairwise correlations between predictors variable in 2nd stage regression (analysis 2). Asterisks indicate significant correlation after correcting for multiple comparisons.

6.2.5 Statistical Analysis

Permutation Testing: To test the significance of our model, we ran repeated permutation tests in which reconstruction or prediction accuracy values were fixed and semantic similarity values were randomly shuffled between words, taking the parameters of the model at each permutation. The testing consisted of 1000 permutations. Parameters of the true model (β and R^2) were deemed significant if they exceeded 95th percentile of the distributions of model parameters based on shuffled semantic similarity values.

Correcting for Multiple Comparisons: Multiple comparison correction was required for analyses that involved statistical testing at multiple scalp electrodes, multiple time windows or both. A cluster-mass non-parametric analysis was conducted (Maris and Oostenveld, 2007) to overcome the multiple comparison problem. This approach includes biophysically motivated constraints that can increase the sensitivity of the statistical test in comparison to a standard Bonferroni correction. Neighbouring electrodes or time windows that show a significant correlation between semantic similarity and reconstruction/prediction accuracy (measured as the T value of the regression coefficient being higher than the critical threshold value; $\alpha = 0.05$) were clustered together. Cluster-level statistics were calculated by summing the T-values in each cluster separately. Permutation testing was then performed on all the data and Monte Carlo p-values were calculated for all significant clusters under the permutation distribution of the maximum cluster-level statistic (Maris and Oostenveld, 2007).

6.3 Results

6.3.1 Early Auditory Encoding of words is influenced by their semantic information

We wished to investigate the relationship between a word's similarity to its preceding context and how well it was encoded in the initial stages of its acoustic-phonetic processing. To measure

the cortical encoding of words we first reconstructed the speech envelope of each word from the recorded neural data using a linear decoder (Figure 6.1, Stage 1). Word reconstruction accuracy was found by comparing the reconstructions to the actual word envelopes. Each content word was also scored on its similarity to preceding sentential context. The resulting sets of semantic similarity values and envelope reconstruction accuracies were used to fit a linear model with the number of datapoints corresponding to the number of spoken content words (Figure 6.1, Stage 2). The model included low level speech features of envelope variability, pitch and resolvability to control for confounding acoustic (or bottom-up) effects on reconstruction accuracy. Table 6.3 gives the feature weights of the model. This shows a significant positive relationship between a word's semantic similarity and its auditory encoding ($\beta = 0.0126$, $p < 5 \times 10^{-11}$). In other words, the more semantically similar a word is to its context, the more faithfully the envelope of that word is reflected in its EEG response. The low-level acoustic predictors, most notably envelope variability, are also correlated with reconstruction accuracy. This is unsurprising given that cortical responses are sensitive to such acoustic measures (Ogania and Chang, 2018). Importantly, however, the relationship between semantic similarity and envelope reconstruction holds despite these measures accounting for a substantial amount of the variance explained in envelope reconstruction accuracy. The proportion of the variance explained by the full model is $R^2 = 0.0171$ (residual degrees of freedom = 103183).

	β	SE	t-Stat	P
Semantic Similarity	0.013	0.0018	6.76	1.3×10^{-11}
Envelope Variability	0.074	0.002	36.52	3.4×10^{-290}
Relative Pitch	-0.015	0.0026	-5.74	9.24×10^{-9}
Resolvability	-0.022	0.002	-10.66	1.4×10^{-22}

Table 6.3 Model weights for 2nd stage regression

Although the p values show the fit and semantic variable coefficient to be highly significant, this could be an artefact of the large sample size (Lin and Lucas, 2013). Therefore, to further test the significance of this relationship, we fixed the word reconstruction accuracy values, but shuffled the values of semantic similarity and recomputed the model fit and β values. This was repeated 1000 times. The null distribution of resulting R^2 values (left) and β values (right) are shown in Figure 6.2A, with the true model values indicated by the vertical dashed line. The average variance explained by the models with randomly permuted semantic similarity is $R^2 = 0.0167$. It is clear from these plots that the true semantic similarity values for each word are significantly related to the accuracy of their decoding.

In the 2nd stage regression, we used a linear mixed effects model which can model variability due to items and subjects simultaneously. To test for the robustness of the effect within subjects,

separate linear models were fit based on individual reconstruction accuracies to calculate R^2 and β values. β values were significantly greater than zero across subjects ($t = 4.2$, $p = 5.4 \times 10^{-4}$; one sample t-test, $N = 19$). Again, we conducted permutation testing, shuffling semantic similarity values for each subject's individual model. This generated null distributions of R^2 and β for every subject. The mean of these null distributions, along with R^2 and β for models with unshuffled values are plotted in Figure 6.2B for each subject. Individual subjects' R^2 and β values were significantly higher than the means of these null distributions ($t = 3.76$, $p = 1.4 \times 10^{-3}$ and $t = 4.2$, $p = 5.4 \times 10^{-6}$, paired t-test, $N=19$). 9 out of 19 subjects had R^2 values that exceeded the 95th percentile of their null distribution ($p = 1.1 \times 10^{-7}$, cumulative binomial probability of achieving $\geq n$ results at $p=0.05$ level) and 9 out of 19 subjects had β values that exceeded the 95th percentile of their null distribution ($p = 1.1 \times 10^{-7}$, cumulative binomial probability of achieving $\geq n$ results at $p=0.05$ level).

The reconstruction accuracies for each word were based on the correlations between the envelope and the reconstructed envelope in the first 100ms of the word's utterance. To study how semantic similarity may affect the tracking of a word as it unfolds in time, we shifted this 100ms window over the word, recalculating reconstruction accuracies at each 15ms increment. Each set of reconstruction accuracies was regressed with the fixed semantic similarity set of values to calculate a model weight β for each time window. Figure 6.2C shows the resulting model weight β for each time window. 95% confidence intervals of β are shown in light blue. Asterisks indicate time windows where β values were significantly greater than 0 after correcting for multiple comparisons. The relationship between word reconstruction accuracy and semantic similarity is maximal and at time window of 0-100ms (window centre = 50ms). It is worth noting that this peak β value corresponds to the one calculated from the model in Figure 6.2A as this model was estimated on the first 100ms of the reconstruction.

Finally, we wished to test if any correlation existed between semantic similarity and potentially confounding measures extracted from the raw EEG signal that might not be related to linguistic processing. We estimated the mean variance across channels in the EEG signal in the first 100ms of each word and fit a linear mixed effects model for this variable with semantic similarity. This returned a slope parameter estimate that was not significant ($\beta=0.043$, $SE = 0.13$, $p = 0.74$). We also looked at the relationship between semantic similarity and channel specific variances in the EEG signal, fitting a LME model to each channel. We found no significant relationship between semantic similarity and channel-specific EEG variance.

6.3.2 Semantic Information Interacts with the Encoding of Speech Features at Different Hierarchical Levels

Cortical tracking can further be assessed by using a forward model (Lalor and Foxe, 2010) to predict the EEG. This method predicts channel specific neural responses from the speech input. The method can be advantageous as it provides an indication of scalp locations where interactions between high-level semantic information and low-level auditory processing are most prominent. Forward modelling can also use different multivariate representations of speech as input to the model. The most commonly used representation for this approach has been the speech envelope. However recent work has shown that spectro-temporal and phonological representations of speech can improve EEG prediction accuracy (Di Liberto et al., 2015). The use of a forward model in tandem with our approach, therefore, gives the added benefit of indicating at which processing stage (spectral or phonological analysis) cortical activity is affected by semantic information. We created a TRF based on representing the speech signal as a combination of its spectrogram and a discrete set of time-aligned phonological features. This gives EEG predictions from which word prediction accuracies were derived. As before, a 2nd stage regression analysis was conducted to measure the strength of the relationship between a word's EEG prediction accuracy and its semantic similarity to context. The model weight of the semantic similarity variable β can now be described as a function of the time window by channel by speech representation.

Figure 6.3A (top row; FS) shows β plotted across all EEG channels for three time-windows (0-100ms, 50-150ms, 100 – 200ms) and for a combined spectrogram and phonetic feature model. Black dots indicate the channels and time windows where beta coefficients were significantly greater than zero, after correcting for multiple comparisons. This revealed that interactions between semantic similarity and early auditory encoding were significant over fronto-central scalp regions.

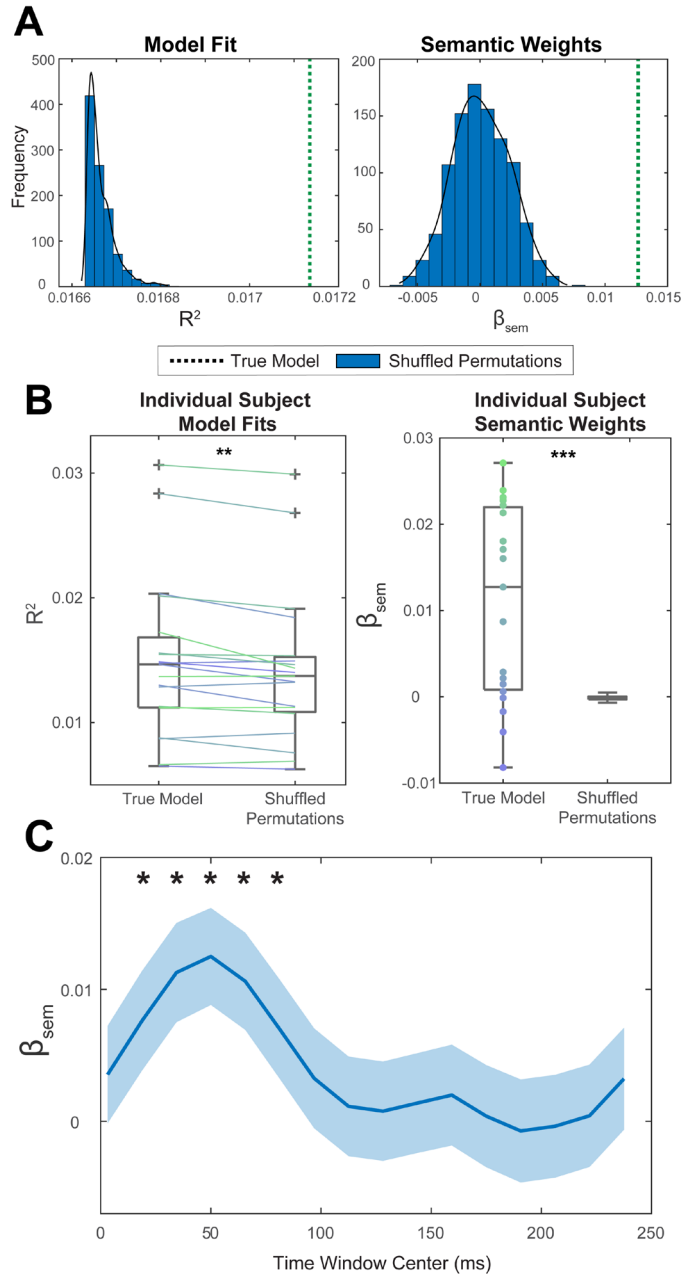


Figure 6.2 Backward Modelling Approach. **(A)** Fitted model parameters, R^2 (left) and β weights for the semantic variable (right) are indicated by the vertical dashed lines. Histograms show the distribution of resulting parameters for models in which the semantic variable was shuffled (1000 permutations). **(B)** Individual subject fitted model parameters, R^2 (left) and β coefficients (right) compared to the mean distribution values for models based on shuffled semantic similarity values ($n=1000$). Each colour represents a different subject. **(C)** 2nd stage linear models were constructed for word envelope reconstruction accuracies for increasingly shifted time windows, all 100ms in length. Plotted are the β coefficients for each fit model as a function of the moving time window centre. Light blue areas indicate the 95% confidence intervals of the β coefficients. Asterisks indicate the time windows where β was significantly greater than 0, after correcting for multiple comparisons.

Next, we wished to determine the effect that semantic similarity on the cortical encoding of speech features in isolation. This was done by running the same 2nd stage model but including prediction accuracies from the spectrogram TRF or phonetic features TRF as reciprocal nuisance

regressors to account for any variance related to processing at the other level. These models can be denoted as either FS-S (phonetic) or FS-F (acoustic). We have conducted similar analyses in the past (Di Liberto et al., 2018) to show that top-down prior information affects the processing of acoustically degraded speech specifically at the level of phonemes. As seen in Figure 6.3A (middle and bottom rows), the isolated phonetic features model shows similar β values to the combined model at earlier stages (0-50ms and 50-150ms windows) but relatively reduced β values in later windows, whereas the isolated spectrogram model shows an overall reduction in β . Figure 6.3B shows β as a function of time-window centre for fronto-central scalp electrodes.

To further test the robustness of this effect at an individual subject level, we constructed subject models through the same forward modelling pipeline but based on individual subject EEG prediction accuracies. We looked at the β values, averaged across fronto-central channels, for the 3 models (isolated spectrogram, isolated phonetic features, combined) at an earlier stage and later stage after word onset. We chose 2 windows that were non-overlapping in time (0-100ms and 100-200ms, respectively) to represent these two stages. Figure 6.3C shows boxplots of scalp mean β values for individual subjects for the 3 models and 2 time-windows. We ran a 2-way ANOVA with factors of model (spectrogram; FS-F, phonetic features; FS-S, combined FS) and time-window (0-100ms, 100-200ms). We found significant main effects for model ($F=11.47$, $p < 0.05$) and time window ($F=10.69$, $p < 0.05$) and a significant interaction effect of model and time window ($F = 3.49$, $p < 0.05$). Due to the significant interaction effect we further compared each model for the two separate time windows using paired t-tests. Significant differences after FDR correction (Benjamini and Hochberg, 1995) are shown in Figure 6.3C (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). This indicates that peaks in β are being driven largely by the effect of semantic information and phonological processing. The significant interaction effect of time windows and model suggests a propagation of information through time from higher linguistic levels of processing to lower ones.

6.3.3 Impact of Semantic Similarity on Theta Band Speech Tracking

The low frequency EEG signal was further filtered into delta and theta bands to investigate the effect of semantic context on tracking in different frequency ranges. This was based on previous studies that have posited different roles in speech processing for the delta and theta bands (Ding et al., 2014).

The filtered EEG signal was put through the same backward and forward modelling pipelines as described above. Figure 6.4A shows β weights of individually fitted 2nd stage models for EEG filtered into delta and theta bands using the backward modelling approach. Results show that β values are significantly greater than zero for both bands ($t = 3.52$, $p = 2.4 \times 10^{-3}$ for delta and $t = 6.69$, $p = 2.8 \times 10^{-6}$ for theta; one sample t-test, $N=19$) as well as a significant increase for β values

derived from theta band cortical tracking compared to delta band ($t = 3.24$, $p = 0.004$, paired t-test, $N=19$). The forward modelling approach confirms these results. β weights for a group-level averaged 2nd stage model were higher for theta band compared to delta band. Topographical plots taken for time windows of 0-100ms and 60 – 160ms, show significant β values over fronto-central electrodes for theta range whereas β values related to delta band show activity over peripheral temporal electrodes.

6.3.4 Dissociable effects of lexico-semantic measures on auditory encoding of speech

Having established that top-down predictions about the semantic features of upcoming words influence the encoding of speech at lower levels, we investigated which mechanistic accounts of prediction our results supported. We observed that the encoding of spoken word features increased as a function of their semantic expectation, which aligns with theories of sharpening. However, measures derived from computational models that are more in line with prediction error could reveal the opposite pattern of effects, where the expected signals are explained away and only the features of unexpected events are encoded at lower linguistic levels. To that end, we added a measure of word-level lexical surprisal as an additional predictor variable to the 2nd stage model. Surprisal is associated with the Kulbach-Liebler divergence of prior and posterior belief distributions (Levy, 2008) and has been formally linked with prediction error in theories of predictive coding (Friston, 2005). We first investigated the 2nd regression weights using the backward modelling approach. The feature weights of this model are given in table 6.4. In addition to a positive relationship between a word's semantic similarity and envelope encoding ($\beta = 0.017$, $p = 1.4 \times 10^{-15}$), there was a significant positive relationship between how surprising a word was relative to its context and how well it was encoded in the brain ($\beta = 0.009$, $p = 1.7 \times 10^{-5}$). This finding was striking given the fact that the lexico-semantic measures reflect opposite patterns of words expectancy, with semantic similarity assigning higher values to more expected words and lexical surprisal assigning higher values to less expected words. This difference is confirmed statistically, with the measures being negatively correlated with each other (Pearson's $r = -0.12$; Table 6.2).

	β	SE	t-Stat	P
Semantic Similarity	0.017	0.002	7.99	1.4×10^{-15}
Lexical Surprisal	0.009	0.002	4.29	1.7×10^{-5}
Envelope Variability	0.063	0.002	27.72	2.4×10^{-168}
Relative Pitch	-0.017	0.003	-5.57	2.5×10^{-8}
Resolvability	-0.023	0.002	-10.1	6.2×10^{-24}

Table 6.4 Model weights for 2nd stage model with both lexical-semantic features.

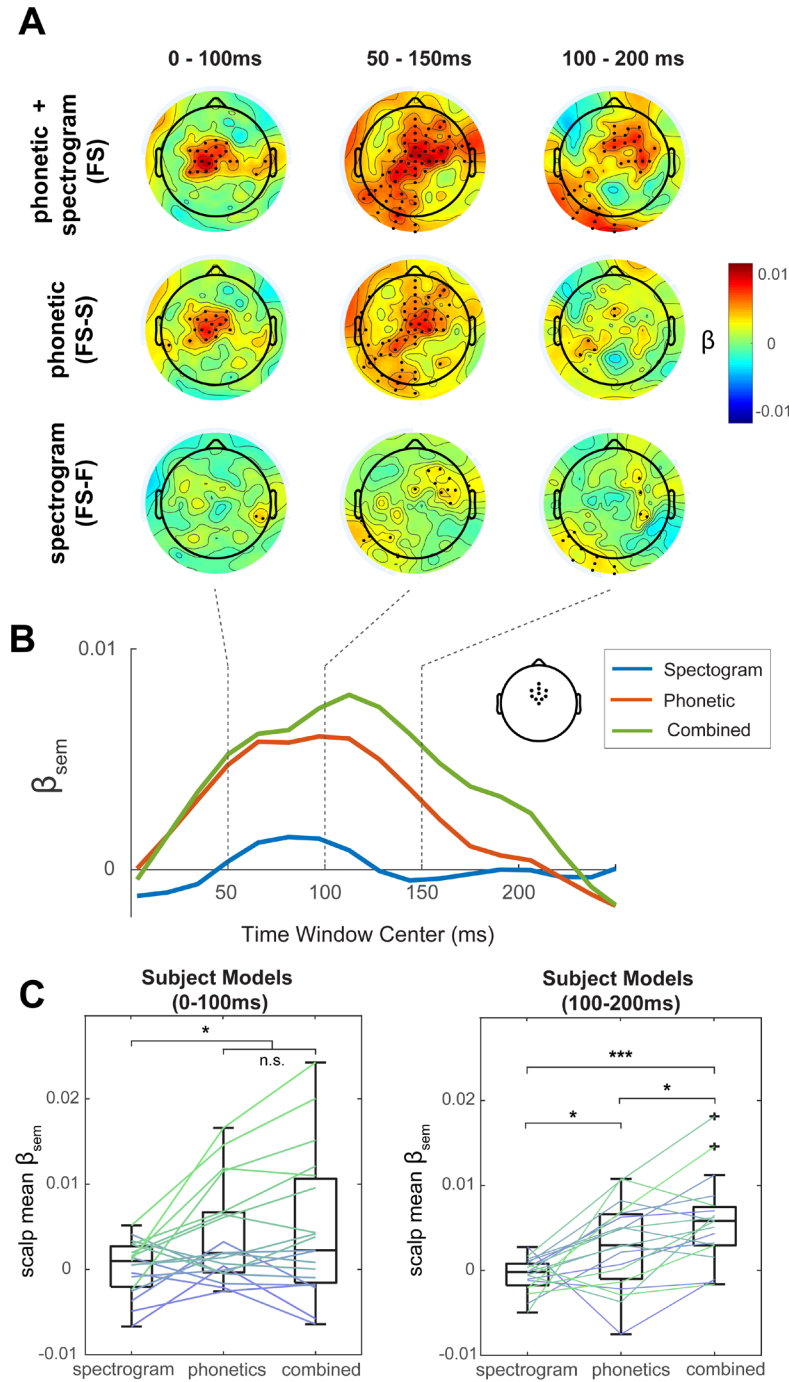


Figure 6.3 Forward Modelling Approach (A) 2nd stage model semantic weights (β_{sem}) plotted across all EEG channels for three time-windows and for TRF inputs of spectrogram (bottom row), phonetic features (middle row) and combined spectrogram and phonetic features (top row). Black dots indicate the channels and time-points where the β_{sem} was significantly greater than zero, after multiple comparison correction. (B) Time-course of model β_{sem} weights for spectrogram (blue), phonetic features (red) and combination (green) models for fronto-central electrodes. The channels selected for this region are indicated in the top right-hand corner of the plot. (C) 2nd stage model semantic weights (β_{sem}) for individual subject models, averaged across fronto-central electrodes and taken at time windows of 0-100ms (left) and 100-200ms (right). Post hoc test revealed significant differences between models in each time window (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ after correction).

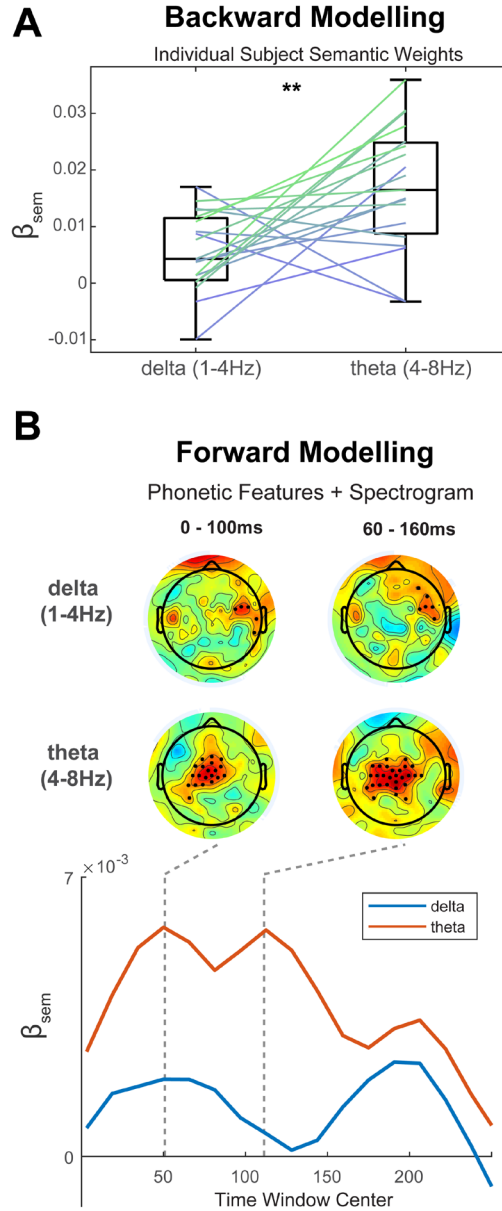


Figure 6.4 Effects of Semantic Similarity on Tracking in Different Frequency Bands. (A), Backward Model. 2nd stage model semantic weights (β_{sem}) for individual subject models show that semantic similarity has a larger effect on theta band (4-8Hz) tracking than delta band (1-4Hz), (** $p < 0.01$). (B), Forward Model. 2nd stage model semantic weights (β_{sem}) for combined phonetic features and spectrogram TRF for delta and theta band filtered EEG. Topographical plots at time windows of 0-100ms and 60-160ms (above) show significant β values for theta band predominantly over fronto-central channels. (Below) Time-course of model β weights for theta (red) and delta (blue) bands average over fronto-central electrodes.

We also ran 2nd stage linear models for each subject individually to obtain single subject beta coefficients. Figure 6.5A show the distribution model features weights of similarity and surprisal. These weights were both significantly greater than zero ($t_{18} = 6.75$, $p = 2.5 \times 10^{-6}$ for similarity and $t_{18} = 4.07$, $p = 7.21 \times 10^{-4}$ for surprisal). We conducted permutation testing, shuffling either semantic dissimilarity or lexical surprisal predictor variables while holding the other predictor

variable constant. The beta coefficients of the shuffled parameter were extracted from each subject's individual model with each permutation. Overlaid histograms in figure 6.5A show the distribution of median values across subjects for each shuffled permutation. The median values for the unshuffled models (boxplot line) exceed the highest values of these distributions.

As before, reconstruction accuracies for each word were based on the correlations between the envelope and the reconstructed envelope in the first 100ms of the word's utterance. To study how lexico-semantic measures affect the tracking of a word as it unfolds in time, we shifted the 100ms window of focus over the word, recalculating reconstruction accuracies at each 15ms increment. Each set of reconstruction accuracies was regressed with the fixed lexico-semantic measures to calculate a model weights for each time window. Figure 6.5B shows the resulting weight values, averages across subjects. Beta weights for both similarity and surprisal show peaks in an early time window of ~0-100ms and were both significant in this window ($p < 0.05$, running t-test, FDR corrected). In addition, beta weights for semantic dissimilarity showed a sustained significant positivity that was significantly greater than lexical surprisal at a later time window of ~100-250ms ($p < 0.05$, running paired t-test, FDR corrected).

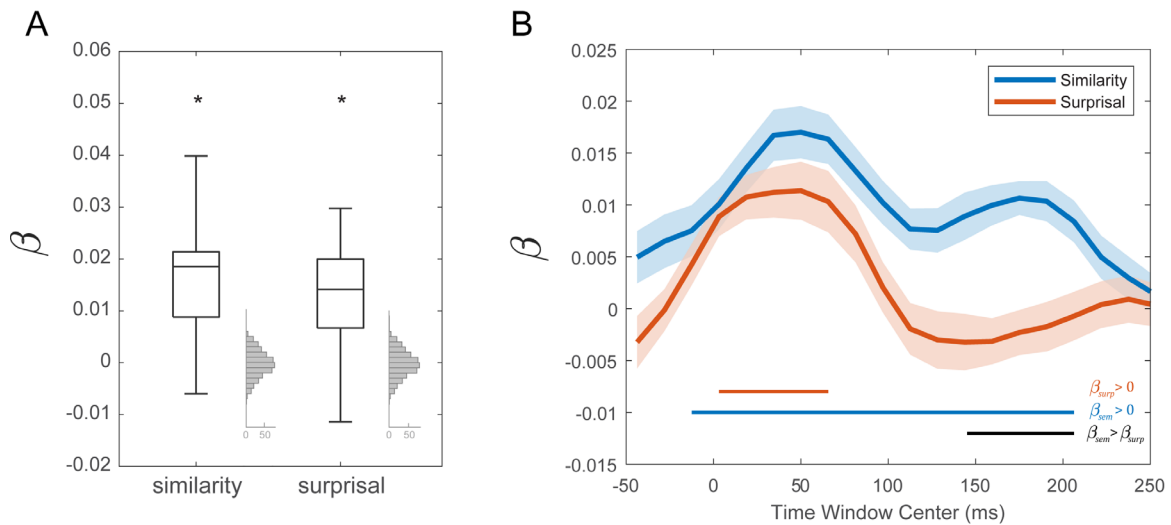


Figure 6.5 Lexical-semantic β coefficients for 2nd stage model. For the 2nd stage model, semantic similarity and lexical surprisal were regressed against word reconstruction accuracies outputted from the 1st stage decoder model. Separate models were estimated for each individual subject (A) Box-and-whisker plots of semantic similarity and lexical surprisal β coefficients of 2nd stage models for each subject. Overlaid histograms show the distribution of median values across subjects for each shuffled permutation. The median values for the unshuffled models (box mid-line) exceed the highest values of the permuted models. (B) Semantic similarity and lexical surprisal β coefficients as function of the time window used for estimating word reconstruction accuracy. Both coefficients were significantly greater than 0 (blue and red horizontal line, FDR corrected) in an earlier windows after word onset. However, β coefficients for semantic similarity showed sustained significant values at later time windows. Semantic similarity β coefficients were significantly greater than lexical surprisal β coefficients in later time windows (black horizontal line, FDR corrected).

We then assessed impact of similarity and surprisal on higher-level linguistic features using the forward model. We focused on the encoding of phonetic features here as it was most strongly affected by semantic information (Figure 6.3). As before, semantic similarity and lexical surprisal measures were regressed to word prediction accuracies, estimated for 2 non-overlapping windows (0-100ms and 100-200ms after word onset) in the 2nd stage model. Figure 6.6A shows the model weights for semantic similarity and surprisal at each EEG channel for both time windows. Black dots indicate the channels and time windows where beta coefficients were significantly greater than zero, after correcting for multiple comparisons. Interactions between both lexico-semantic measures and phonological processing of speech were significant over fronto-central electrodes. Furthermore, the influence of these measures on bottom up processing was dissociated in time, with similarity affecting processing at the earlier time window and surprisal in the later window. This was further supported by single subject beta coefficients, derived using individual 2nd stage linear model and averaged over fronto-central channels. A 2-way anova with factors of time window (0-100ms and 100-200ms) and measure (similarity and surprisal), revealed a significant interaction ($F = 5.93$, $p < 0.02$; figure 3B) and no main effects ($F = 0.01$, $p > 0.05$ and $F = 0.1$, $p > 0.05$, for measure and time-window effects, respectively). Similarity and surprisal weights were compared in each time window separately, post hoc. Similarity beta weights were significantly higher than surprisal beta weights in the earlier time window ($t_{18} = 2.37$, $p = 0.008$, paired t-test) and surprisal beta weights were significantly higher than similarity weight in the later time window ($t_{18} = 2.18$, $p = 0.005$, paired t-test).

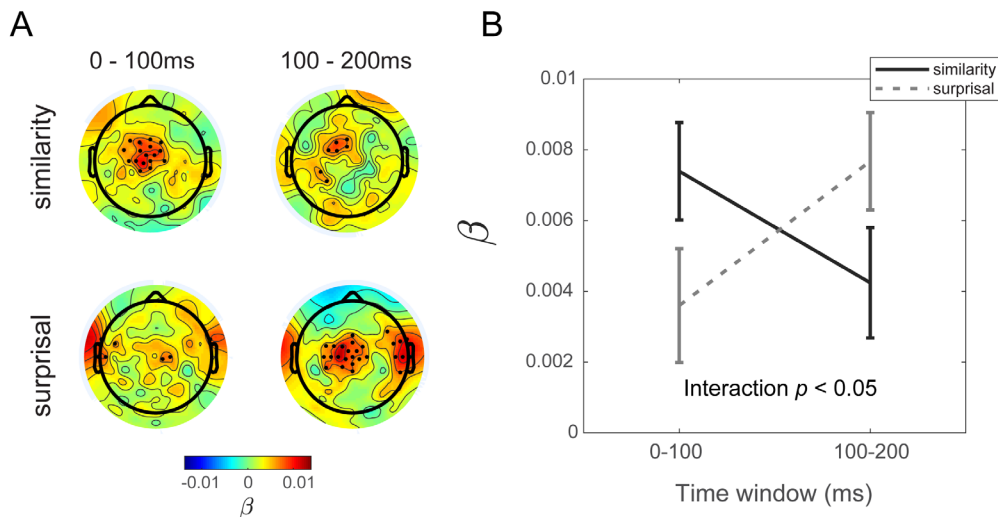


Figure 6.6 Lexical-semantic β coefficients for 2nd stage model based on phonetic feature TRF prediction accuracies. (A) Topographical plots show LME model coefficients for semantic similarity and lexical surprisal as a function EEG channel and time window. Black dots indicate the channels and time points where the β was significantly greater than 0, after multiple-comparison correction. (B) Average weight coefficients ($\pm$ s.e.m.) of individual subject 2nd stage models over frontal central electrodes as a function of time window. A significant interaction between lexico-semantic measure and time-window was found.

6.4 Discussion

In this study we first looked at whether prior knowledge of a word's semantic context impacted the neural processing of its acoustic and phonological features. We found that the more semantically similar a word was to its context, the better it was encoded in the initial stages of its utterance. This was observed using a backwards model, where the reconstruction accuracy of a word's acoustic envelope was positively correlated with its semantic similarity. The result was then mirrored by the forward modelling approach where the prediction of the EEG at the time of the word was higher for more semantically similar words. This suggests a mechanism that links top-down prior information with the early cortical encoding of words.

Previous studies that look at the relationship between cortical activity and top down information tend to involve paradigms where the acoustic information is masked, usually with noise, to vary intelligibility (Luo and Poeppel, 2007; Davis et al., 2011; Sohoglu et al., 2012; Ding and Simon, 2013; Ding et al., 2014). Although much insight has been gained from these studies, there are two main drawbacks to consider. First, changes in the intelligibility of the speech signal inevitably lead to changes in the acoustic properties of the same signal (Peelle et al., 2013). Thus, the observed decrease in delta and theta band tracking as a function of intelligibility could be due primarily to changes in the acoustic signal. Second, the change in a listener's attention for different levels of intelligibility could represent something of a confounding effect when considering the relationship between linguistic information and neural activity. Attention is known to modulate cortical tracking of speech (Zion Golumbic et al., 2013; O'Sullivan et al., 2015) and, therefore, could be the cause of differences in encoding measures rather than the intelligibility or linguistic content per se. Our method circumvents this confound by using natural continuous speech that does not change in its level of intelligibility. Furthermore, by implementing a general linear modelling approach, we can partially control for low-level acoustic factors of the speech signal by including them as predictor variables in the model. Subjects were instructed to engage with the audiobook, all reporting an enjoyment of the story. It is reasonable, therefore, to assume that no systematic, tonic attentional effects were influencing the auditory processing of the speech.

The relationship between semantic similarity and word reconstruction/prediction accuracy was measured by the 2nd stage model fit, R^2 , and coefficients β . Notably, the magnitude of this effect is small. However, the dependent variable used for the regression is the neural signal measured for single spoken words with no averaged repetitions. It is therefore reasonable to expect low model fit and parameter values. Reconstructing the envelope from EEG usually produces accuracies ranging from 0.1 to 0.2 (O'Sullivan et al., 2015; Crosse, Di Liberto, Bednar, et al., 2016), revealing that a high proportion of the EEG signal is unrelated to the stimulus envelope. Permutation tests, whereby the semantic values for words are shuffled, provide a baseline

distribution of our dependent measures. From this, we can see that our R^2 and β values, though small, are significantly higher than would be expected by chance. Another striking feature of the interaction between semantic similarity and envelope tracking is its timing. Figure 6.2C shows that the enhancement of tracking due to context is strongest in the first 100ms of a word's utterance. This implies that this interaction occurs rapidly and at a pre-lexical stage.

The forward modelling approach revealed that interactions between cortical tracking and semantic similarity were significant over fronto-central channels, within the first 100ms of a word's utterance. Distinct aspects of this interaction emerged when different levels of the processing hierarchy were isolated. We showed that the processing of a word's phonetic features was more strongly affected by its similarity to context than the processing of its acoustic features and that early interactions between semantic information and tracking were likely attributable to this phonological processing. We also show that semantic information affects phonological processing at an earlier latency than acoustic processing. This is indicative of a top-down mechanism, where high-level semantic information propagates down through hierarchical levels, first activating representations associated with the phonetic features of the speech and then with the lower-level acoustic features. This finding is in line with studies showing how prior knowledge modulates activity in regions associated with higher level processing, such as IFG, before regions associated with lower-level sensory processing (Sohoglu et al., 2012). We also see that this top-down relationship is more prominent in the theta band filtered signal compared to the same signal filtered into the delta band. Previous studies have shown how theta band is largely implicated in the tracking of speech and how it is sensitive to changes in intelligibility (Luo and Poeppel, 2007; Ding et al., 2014).

Our results support the view that prior knowledge about upcoming words facilitates their perception at lower representational levels (McClelland and Elman, 1986; McClelland et al., 2006). However, there are different ways in which this facilitation can be computationally realised (de Lange et al., 2018). Different models based on Bayesian inference have been developed to explain precisely how top-down information influences bottom up sensory processing. Sharpening theories argue that the encoding of predicted stimulus features are enhanced by feedforward and feedback connections (Lee and Mumford, 2003; Kok et al., 2012), whereas predictive coding argues that feedforward connections carry only the discrepancy (or error) between the predicted signal (embodied in feedback connections) and incoming sensory evidence (Rao and Ballard, 1999; Friston, 2005; Clark, 2013). Adjudicating between these different accounts has been difficult using analyses that compare neural response amplitudes, as both models predict a reduction in neural activity for more expected stimuli (Kersten et al., 2004). Techniques like linear regression, representational similarity analysis (RSA) and multivoxel pattern analysis (MVPA) that more directly measure the patterns encoded in neural activity have

provided a breakthrough in this regard (Haxby, 2001; Kriegeskorte, 2008), although studies using such techniques have yielded apparently contradictory findings (Kok et al., 2012; Blank and Davis, 2016; Blank et al., 2018; Wang et al., 2018; Sohoglu and Davis, 2020). We sought to address this by combining our newly developed framework with lexico-semantic measures of word expectation.

Semantic similarity is derived using word embedding models that capture the latent semantic features of words. It compares a word's semantic features with those that been activated by the context in which it appears. Here, as well as previous chapters, we propose that this measure reflects a neural process in which semantic features of upcoming words are predictively preactivated ahead of their perception (Federmeier and Kutas, 1999; Kuperberg and Jaeger, 2016). The more strongly a word's context preactivates features that match its own, the more predictable that word will be at the level of semantics. Importantly, predictions at these higher levels propagate down the hierarchy to activate internal phonetic and acoustic representations before the arrival of bottom up information (Van Petten et al., 1999a; DeLong et al., 2005). We show enhanced neural encoding of speech features for words with higher semantic similarity to their previous context. Such effects align with findings by Leonard and colleagues. They show that, when a listener is biased by semantic context, they perceive ambiguous sounds as certain phonemes that conform to that bias (Leonard et al., 2016). Crucially, this perception is underpinned by an enhanced neural representation of the perceived phoneme, despite it being omitted from the bottom up signal. This indicates that top down semantic information penetrates perceptual processing at the level of phonemes. Previous studies have observed a similar enhanced neural representation of predictable events (Kok et al., 2012) and have taken this as evidence for the sharpening of predicted signals. On its own, the relationship between semantic similarity and neural encoding would support similar conclusions. However, using a measure of Bayesian surprisal that relates more directly to prediction error expands on this initial finding (Friston, 2005; Levy, 2008).

We added lexical surprisal as an additional predictor variable in the 2nd stage regression. Bayesian surprise reflects the degree of belief updating that a comprehender undergoes as they shift from a prior to posterior distribution of beliefs. As discussed above, it is one way of computationally formalising prediction error (Friston, 2005; Levy, 2008). In contrast to semantic similarity, we found that words that were more surprising (less expected) were more strongly encoded in the neural signal. Previous experiments have observed similar effects, with stronger encoding of more unexpected sensory signals (Gagnepain et al., 2012; Blank and Davis, 2016; Sohoglu and Davis, 2020). These findings, as well as our own, support theories of predictive coding. Crucially though, our analysis revealed the parallel computations of prediction and prediction error.

The co-existence of distinct populations at each cortical level that encode predictions and prediction errors is a core tenet of predictive coding theories which has been questioned due the apparent lack of empirical support (Egner and Summerfield, 2013). Our results indicate the coexistence of these units and show them in operation during the processing of naturalistic continuous speech. The use of recordings with high temporal resolution like EEG allows us to observe the timings of these operations in finer detail and produces results that further align with this view. We observed an interaction with the lexico-semantic measures of prediction and the relative timings of speech feature enhancement. Neural processing of features is first enhanced for more expected words, followed the enhancement of words that are more surprising. This result conforms to the intuition of predictive coding. The brain must first capture predictions about an upcoming signal before the error can be computed (Press et al., 2020). The fact that this was observed at the level of phonetic features and not for the speech envelope may be due to the phonetic feature TRF model capturing more isolated stages of processing, rather than the envelope decoder, which likely captures conflated responses from multiple stages of processing.

The idea that semantic similarity and lexical surprisal map directly to representation and error units, respectively, is by no means conclusive. Indeed, predictions about upcoming words at lexical levels (as indexed with surprisal) could be captured in representation units and top down predictions from semantic levels (indexed with semantic similarity) could lead large responses in error units. However, our results suggest that these measures, on average, reflect response properties in these distinct units. The conceptual nature of the employed language models support this: semantic similarity reflects a more abstract linguistic prediction that does not necessarily commit to a specific lexical item, whereas surprisal, which predict specific word identity, maps more formally to prediction error. Future work, however, using more elaborate models should help to better disentangle contributions from these neural populations. That said, the measures show a clear dissociation in their influence on bottom up sensory processing. Furthermore, the timings dissociations of these processes capture much about the intuition behind predictive coding.

In conclusion, our results support theories for the top-down influence of prior knowledge on bottom up sensory processing and align with the computational accounts of predictive coding. We have shown that such neural processes are pervasive during naturalistic speech comprehension.

Chapter 7. General Discussion

Understanding the influence of semantic context on speech perception is a topic of great importance. Speech is an inherently ambiguous signal which our brains routinely disambiguate by taking full advantage of the rich information available in context. The neural underpinnings of this process are complex and multi-faceted and uncovering these mechanisms will provide key insights into how the brain achieves the tricky task of language comprehension. In this thesis, we have introduced a novel approach for studying natural speech comprehension and have provided interpretable neural correlates that capture the effects of semantic context on speech perception at multiple linguistic levels. Our results support the notion of prediction in language and show that the use of context is pervasive in everyday speech processing.

The first part of this thesis (Chapters 1-2) motivates this research and sets the necessary groundwork for developing a methodology that is both neurobiologically relevant and plausible. Chapter 3 establishes the framework and tests it under a range of conditions where speech understanding is manipulated. The method is further validated in Chapter 4, where the derived neural indices of speech tracking are isolated to the level of semantic representation. We then develop our methodology in Chapter 5 by including measures from different language models that reflect distinct relationships between words and their context. In doing so, we test how context impacts predictive processing at dissociable linguistic levels. Furthermore, we show how the effects of context change with age. In chapter 6, we further investigate the mechanisms underlying prediction. Our results support theories of predictive coding and suggest the co-existence of representation units and error units.

7.1 A Novel Approach to Study Natural Speech Comprehension

In this thesis, we have introduced a novel framework for studying higher-level linguistic processing of natural, continuous speech. Specifically, our initial method measures the processing of a word's semantic features relative to its immediately preceding context. This work was

inspired by the advances in computational language modelling (Bengio et al., 2003; Mikolov et al., 2013b; Pennington et al., 2014) and the extensive literature on the N400 (Kutas and Federmeier, 2011), which shows an approximately linear relationship between a word's semantic fit with preceding context and its evoked response amplitude (Kutas and Hillyard, 1984). Our work reveals that, when listening to natural speech, the ongoing dynamics of cortical activity reflect the semantic processing of words in context in a rapid, time-locked fashion. Furthermore, we show that this measure is robustly sensitive to a subject's ability to understand the speech they are hearing and whether the recognised words follow a coherent, structured context.

Our model of semantic dissimilarity is grounded in theories of how semantic information is stored in the brain. In chapter 2, we reviewed evidence that conceptual knowledge is partly represented in the form of linguistic information (Binder and Desai, 2011; Meteyard et al., 2012; Wang et al., 2020), and that proponents of this view argue that knowledge can be represented as a statistical distribution of word co-occurrences (Landauer and Dumais, 1997). Distributional models such as LSA, word2vec and GloVe that are trained using word co-occurrence information embody this hypothesis. These models accurately capture semantic associations between words (Mikolov et al., 2013a; Fu et al., 2014; Pennington et al., 2014) and have been regularly used to investigate the organisation of semantic memory in the brain (Mitchell et al., 2008; Anderson, Binder, et al., 2019). Given this knowledge, we selected word embeddings to form the basis of our semantic dissimilarity model, employing them to represent the latent semantic features of individual words. Hence, our model of semantic dissimilarity has high neurobiological relevance and is arguably capturing the relationship between words and their context at the level of semantic representation. It is unsurprising, therefore, that it shows such robust sensitivity to speech comprehension.

Our work is part of a consorted effort to parameterize speech beyond its low-level acoustic representation in terms of more abstract higher-level linguistic features, to be used in system identification methods (Di Liberto et al., 2015; Frank and Willems, 2017; Brodbeck, Hong, et al., 2018; Brodbeck, Presacco and Simon, 2018; Heilbron et al., 2019; Weissbart et al., 2019). This has allowed for the possibility to disentangle speech processing mechanisms at different hierarchical levels using the same speech stimuli. As discussed in Chapter 2, the speech envelope has been predominantly used with system identification to study speech perception and comprehension in the brain (Ding and Simon, 2014; Octave and Reichenbach, 2019; Brodbeck and Simon, 2020). This is in part due to the fact that envelope tracking in the brain is highly robust across subjects and can be modulated by speech intelligibility and attention (Power et al., 2012; Ding et al., 2014; O'Sullivan et al., 2015; Vanthornhout et al., 2018). However, although envelope tracking may form the basis of speech comprehension and has a likely causal relationship with the processes underlying successful speech understanding, there clearly isn't a one-to-one correspondence between the two phenomena (Brodbeck and Simon, 2020). This is

indicated clearly in chapter 4 where, despite a decrease in comprehension scores, envelope tracking measures did not differ between speech scrambling conditions. Recent work using a large, age-varying sample of subjects also points out the limitations of envelope tracking as predictor of behavioural success (Tune et al., 2020). Our work highlights the need for high-level feature spaces when modelling speech processing during comprehension. Furthermore, we show that outputs from such higher-level models are better correlated with behavioural measures of speech understanding and offline measures of verbal fluency.

From a neural engineering perspective, our framework could be used for a number of practical applications that attempt to derive objective measures of speech and language processing. Given the central role of spoken language in human life, difficulties with speech understanding can have profound negative consequences for daily functioning. Thankfully, there exist well-developed therapeutic interventions that can mitigate speech and language problems in many clinical groups. Early application of such interventions is a key element in their success, meaning that early diagnosis of problems is critical. One potentially fruitful route towards enabling such early diagnosis is by analysing neurophysiological responses to speech. EEG is a highly suitable device for obtaining such responses due to its relatively low cost and portability. Our methodological framework in combination with this imaging technique could mark a potential breakthrough in acquiring such objective measures in situations where it is unclear to what extent speech has been understood, such as in infants or patients in a reduced state of consciousness. Already, our methods are being used for similar purposes. A recent study used semantic dissimilarity based TRFs (along with other speech features) to decode the language proficiency of 2nd language learners, obtaining a high decoding performance (Di Liberto et al., 2020).

As revealed in chapter 5, output neural measures relating to semantic dissimilarity are correlated with semantic fluency scores for older adults. Although the sample size for this correlation is admittedly low, it represents a positive advance in the ability to acquire cognitive estimates from an individual's neural signal under more naturalistic listening conditions. Given that declines in verbal fluency have been shown to be early indicators of Parkinson's and Alzheimer's disease (Bayles et al., 1993; Pasquier et al., 1995), our methods could potentially be used in wearable neural devices that detect the onset of such disorders.

7.2 Predictive Language Processing

Theories of predictive in language have been a topic of controversy in the literature. One aim of this research was to address this debate using more naturalistic speech material. We have shown that measures reflecting a word's relationship with its context are faithfully tracked by the brain during spoken language comprehension and have interpreted our models of semantic dissimilarity

and lexical surprisal as reflecting the predictive strength between a word and its context which facilitates neural processing. However, one could argue that the same models instead index the effort required to incorporate meaning word into its preceding context, reflecting mechanisms of integration (Brown and Hagoort, 1993; Hagoort et al., 2009). Such accounts would support a feedforward view of language processing (Norris et al., 2000; Norris and McQueen, 2008). We believe our findings in chapters 5 and 6 provide some clarity on this topic.

We show in chapter 6 that lexico-semantic measures affect acoustic and phonetic processing of words in the very early stages of their utterance and that the pattern of interactions suggests a propagation of information from higher to lower levels of processing. We feel that this result provides strong evidence in favour of the top-down influence of higher-order information on bottom up processing, refuting the notion of strictly feedforward mechanisms underlying language comprehension. In addition, we see in chapter 5 that neural correlates of semantic dissimilarity are much weaker in older listeners. These results, viewed from the integration perspective, would suggest that older listeners integrate high level semantic feature information to a lesser extent than younger adults, resulting in clear deficits in their ability to understand clear speech. This of course was contrary to what we observed in the experiment. Rather, we feel that our result aligns better with the idea that our models reflect the predictive preactivation of linguistic features at multiple representational level. This preactivation serves to facilitate the processing of new information. Thus, although arguments can be made for both prediction and integration, we believe our results are more consistent with theories of prediction in language which is supported by a hierarchical neural system with reciprocal feed-forward and feed-back connections (McClelland and Elman, 1986; Kutas et al., 2011; Kuperberg and Jaeger, 2016).

Bayesian surprise has been a popular approach for modelling language prediction in the brain. It has been related to offline normative cloze ratings (Smith and Levy, 2011), reading times (Smith and Levy, 2013) and N400 amplitude (Frank et al., 2015). It has been mathematically related to the Kullback-Leibler divergence (Levy, 2008) which represents the shift from prior to posterior probability distributions (Kullback and Leibler, 1951) and corresponds to prediction error in theories of predictive coding (Friston, 2005). However, there are instances in the literature where surprisal fails to model phenomena observed for responses reflecting prediction in language. For example, surprisal models do not capture reductions in neural responses for sentences like *for breakfast the eggs would eat* (Kuperberg et al., 2003) and syntactic anomalies where order is change (Hagoort and Brown, 2000b; Rabovsky et al., 2018). Therefore, it is likely that surprisal cannot explain everything about prediction in the brain. In this thesis, we have introduced and used measures of semantic (dis)similarity with the goal of explaining neural responses to language beyond what is explainable by surprisal.

In our work, we see dissociations between models of surprisal and (dis)similarity in terms of their impact on the cortical processing of speech. Output measures from both models show a robust relationship to neural responses despite being weakly correlated and being fit simultaneously to the data, thus controlling for one another. This indicates that semantic (dis)similarity is providing additional information about neural processing not captured by surprisal. We have proposed that it relates to the preactivation of semantic features, relating to word meaning. This idea is supported by findings in chapter 5, showing weaker responses for older adults who are known to have less sensitivity to semantic features of upcoming words (Federmeier et al., 2002). It could also explain why we observed significant neural responses for the 2-word scrambling condition in chapter 4, when subjects were still able comprehend the speech signal. Furthermore, we have proposed that semantic (dis)similarity, on average, corresponds more closely to activity from representation units in models of predictive coding. This is because the measure reflects predictions at a higher-level of abstraction that do not explicitly commit to specific lexical items, as is the case with the surprisal measure. Of course, there is some degree of overlap between the two models in that they both reflect how meaningful context influences the selection of upcoming words. Although surprisal does not capture the meaning of individual words, it is trained on word symbol sequences that reflect the structure of meaningful events in the real world. Future work should look more in depth into the commonalities and differences between these models to better uncover the neural processes they might represent.

Some have argued that prediction is merely a lab-based phenomenon, resulting from “prediction-encouraging” experimental set-ups (Huettig and Mani, 2016). Our results argue against this and suggest that prediction is pervasive in everyday speech processing. However, previous studies have shown that comprehenders can unconsciously shift their reliance on predictions to a greater or lesser extent depending previous experience (Lau et al., 2013). This indicates some degree of flexibility in the use of such mechanisms. These findings perhaps relate to recent formulations of predictive coding (Feldman and Friston, 2010) where prediction error is weighted by sensory reliability or precision. When a sensory signal is noisy its precision is said to be low and prediction errors are down-weighted. In this instance, a perceiver must rely more on their top-down knowledge. Conversely, when precision is high, prediction errors are upweighted. Our novel framework introduced in chapter 6 would allow the testing of these hypotheses by varying the precision of sensory stimuli. Measuring β coefficients for surprisal and similarity for mildly degraded speech will allow us to examine the shift in a subject’s reliance on top-down and bottom-up inputs. Furthermore, the same methodological approach could be used to help understand perception in neurodevelopmental disorders such as schizophrenia (Fletcher and Frith, 2009; Corlett et al., 2011) or autism spectrum disorder (Pellicano and Burr, 2012; van de Cruys et al., 2014). Schizophrenia has been linked with a decreased precision which results in

overly strong expectations (Powers et al., 2017; Sterzer et al., 2018). In contrast, autism has been linked with an up-weighting of sensory information, with less guidance from top-down expectations (Pellicano and Burr, 2012). Comparing β weights for both populations with a healthy baseline could provide evidence for whether such theories are borne out in everyday language processing.

7.3 Modelling the N400

The neural correlates derived in this thesis share highly similar characteristics with the N400 ERP component (Kutas and Hillyard 1980, Kutas and Federmeier, 2011). In addition, they are based on computational measures that estimate the relationship between words and their context. Therefore, it is reasonable to conclude that our approach is modelling aspects of the classic N400 response but with more naturalistic stimuli. Given this conclusion, it is interesting that both measures of semantic dissimilarity and lexical surprisal, despite being weakly correlated, produce similar N400 components and can sufficiently explain some variance in EEG responses of younger adults. Furthermore, a clear dissociation is seen between these measures when studying responses in older listeners (chapter 5) and looking at their top-down influence the early auditory encoding speech (chapter 6). This suggests the existence of dissociable processes underlying the N400. Indeed, Frank and colleagues tested similar models on EEG and fMRI and found that, whereas both dissimilarity and surprisal measures resulted in N400 components for EEG, fMRI responses showed anatomically distinct activations for each measure (Frank and Willems, 2017).

The idea of multiple processes contributing to the N400 has received some attention in the literature over the years. It has been proposed that the component represents a cascade of processes reflecting the pre-lexical access and post-lexical integration of words (Baggio and Hagoort, 2011; Wehbe, Vaswani, et al., 2014; M. Nieuwland et al., 2018). However, we believe our results align more with a unified perspective of language processing that goes beyond the traditional views of access and integration (Chomsky, 1957). From this perspective, the N400 represents the input-driven update of an internal probabilistic representation of the event described by a sentence. Context serves to constrain the formation of this representation so that, when more predictable information arrives at the input, the activation state of the internal representation will experience only small changes, resulting in a weaker update magnitude (Kutas and Federmeier, 2011; Rabovsky et al., 2018). Here, we propose that the observed dissociation between our language models relates to how different forms of linguistic information, both semantic and lexical, may shape and constrain the formation of this internal representation before the arrival of new input. Indeed, this relates back to the idea that brain uses multiple aspects of context to form predictions about upcoming speech features at multiple linguistic levels (Kuperberg and Jaeger, 2016). We have argued that our models of lexical surprisal and semantic

dissimilarity reflect predictions at lexical and semantic levels. Crucially, these predictions will feed into the processes underlying the N400, culminating in a response that reflects the conflated activity from different linguistic levels. This does not imply that the N400 and our models reflect the only process that underlies language understanding. Indeed, there are multiple ERP components that relate the language processing (Hagoort and Brown, 2000b). Future work will attempt to uncover the processes underlying such components using different or more refined language models.

Most of the prior work relating the N400 to word predictability has estimated word probability using human judgements rather than statistical models. N400 responses to words show a graded relationship with their cloze probability (Kutas and Hillyard, 1984; Dambacher et al., 2006; Wlotko and Federmeier, 2012), which is estimated from offline norming procedures where individuals are required to complete sentences with missing words (Taylor, 1953). Quantifying word predictability in this way has some disadvantages. First, the measure is costly and time consuming. Estimating cloze probabilities for a novel set of sentences requires a large sample of participants to complete the cloze procedure. In addition, cloze ratings can be estimated for only one word per sentence, resulting in a proportionally small amount of word predictability estimates. Second, the cloze procedure is offline, untimed and rather unnatural, which raises further questions about how well the measure captures prediction in the brain during online comprehension (Smith and Levy, 2011). A final, less obvious, disadvantage is the uncertainty of what cloze actually measures. It is unclear what aspects of context or language knowledge influence human judgements when selecting words to complete the sentences during the cloze task. It is likely the case that humans exploit multiple facets of context (lexical, syntactic, semantic) to inform a missing word selection, which are conflated in a single probability score.

Statistical models are trained with different underlying assumptions about language and what mechanisms are involved in sentence comprehension, and thus the different contributions of context to word predictability can be more readily disentangled and interpreted. As these models become more and more accurate at capturing the true statistics of language, their utility in understanding the functional specificity of the N400 is becoming more apparent. Already, there has been some success in modelling the N400 in this way using language models (Frank et al., 2015; Rabovsky et al., 2018). Furthermore, language models can automatically provide predictability estimates for each word in a stimulus which makes them easily applicable to natural stimuli. Our work further highlights the importance of using statistical models when studying language comprehension.

7.4 Future Work

7.4.1 A framework for modelling wider contextual discourse

Within this body of research, we have established that context impacts the processing of natural speech at multiple linguistics levels ranging from acoustics to semantics. In our approach, we have modelled context at a local level, where words are compared with relatively short window of preceding words. However, humans use all types of linguistic and non-linguistic information within both the local and global context of the speech they are hearing. This could entail knowledge about who is speaking, knowledge of the topic, the setting in which a speaker is heard or general world knowledge. This knowledge, known as pragmatics, is an important factor in how we process speech which has been indicated in some cleverly designed N400 paradigms. Hagoort and colleagues found modulatory N400 effects for words that were perfectly sensical sentence completions but violated truths about real-world knowledge (Hagoort et al., 2004). Similarly, Nieuwland and Van Berkum found that anomalous sentences (“the peanut was in love”) can also modulate N400 responses when they are presented within a context that is known to be fictional (Nieuwland and van Berkum, 2006). These studies reveal that what is often left unsaid can have a profound impact on how words are understood.

While investigating the impact of longer windows of linguistic context and long range dependencies between words is important (Heilbron et al., 2019), an arguably more important direction for our modelling framework would be to attempt to incorporate pragmatic information into our models. The benefit of distributional models like word2vec and GloVe is that they are abstracted from the symbolic forms of words, allowing for abstract pragmatic information to be more easily included into refined models of semantic dissimilarity. A future study to test such models could build on psycholinguistic research conducted by Bransford and Johnson, where subjects listened to speech with and without key contextual cues that were crucial for understanding (Bransford and Johnson, 1972). A picture depicting a scene determined whether a speech excerpt was perfectly understandable or completely non-sensical. Later work showed similar effects with titles as the contextual cues (St George et al., 1999). Modelling responses to speech with semantic dissimilarity under these experimental conditions would be an interesting direction in showing how different aspect of context are important for speech understanding. Furthermore, this work would share similarities to our study conducted in chapter 4, where words were correctly recognised, but the overall meaning was not acquired.

Additional future studies could investigate responses to changes in different aspects of pragmatics such as speaker identity or the genre of speech. Crucially, our novel framework allows for the estimation of electrophysiological responses to continuous speech that reflect its higher-level linguistic processing. Previous work that looked at responses to individual words, did not have

the flexibility to address questions relating to more naturalistic paradigms like the ones described above. Our work has thus unlocked new possibilities for studying speech in more naturalistic settings.

7.4.2 Refining the models

In this thesis, we have introduced a methodology for investigating impact of context on the lexico-semantic processing of natural speech (Chapter 3-5). While this approach has produced interpretable neural markers of comprehension and can significantly explain variance in the recorded electrophysiological signal, it is clear that the method is suboptimal and could stand to benefit from further refinement. Refining our model of semantic dissimilarity could help to improve EEG prediction accuracy and explain even more variance in the behavioural comprehension and verbal fluency scores. Here, we outline some of the improvements that could be made to our model.

Impulses. Our models of lexical surprisal and semantic dissimilarity have represented words as impulses, located at word onset. However, most information about a word is not available to the brain at the very initial burst of acoustic energy representing its onset. Furthermore, word identification is variable and will depend on several factors including its length and word cohort entropy (Grosjean, 1980; Marslen-Wilson, 1987). Attempting to locate each word’s specific perceptual centre or “P-centre” (Morton et al., 1976) could significantly improve model performance. Future studies will investigate the contributing factors to a word’s P-centre such as word length; number of phonemes; number of syllables; and word cohort entropy and incorporate them into the model by shifting impulse location according to such factors. In addition, investigating the optimal time-series representations of surprisal/dissimilarity will be an important consideration for the model. In our model, words are represented using the impulse function. However, using the step, gaussian or Laplacian functions to represent the time series of words could lead to further improvements in the model.

Context. As discussed in the previous section, it is important to incorporate extra linguistic and non-linguistic information in the future models. This would involve studying models with varying context window sizes. In our semantic models, we predominantly used contexts consisting of all preceding words in the same sentence. In future work, this should be extended to test longer windows of words and sentences. Identifying the optimal context length could inform us on the amount of information stored in working memory during natural speech comprehension.

Distributed Representations. In our model of semantic dissimilarity, the latent semantic features of words were represented using distributional vectors (Mikolov et al., 2013a; Pennington et al., 2014). These vectors are believed to capture “linguistic” features of words because they are

trained on text information, i.e. word co-occurrence statistics in a large text corpus. However, one of the criticisms of such models is that they often do not capture the full sensory or experiential information of word meaning. This is because some information about words will often go unstated in language. For example, it is rarely useful to communicate the colour of a banana in language due to its widespread knowledge, despite the colour yellow being its dominant feature (Bruni et al., 2014). As discussed in chapter 2, in addition to linguistic information, word meaning is influenced by non-linguistic perceptual, motor and interoceptive experience (Mezei and Paivio, 1972; Barsalou, 1999; Barsalou et al., 2012). Recent studies have shown that including information of experiential attributes of words improves brain-based decoding (Anderson, Binder, et al., 2019). In addition, representations derived a convolutional neural network, trained on visual information have also improved brain-based decoding (Anderson, Kiela, et al., 2017). Therefore, incorporating non-linguistic information into how our models represent word meaning will be an important consideration for future work.

Grand Average Models. Previous system identification approaches have seen some success in using grand average models (O’Sullivan et al., 2015; Di Liberto and Lalor, 2017). These approaches modify the cross-validation procedure by training and testing models across subjects rather than within subjects. This approach is advantageous because there is a lot more training data for the models. However, it suffers from the variability across subjects’ responses. Given the N400’s remarkably stable latency and topographical distribution, it would be unsurprising if a grand average model would generalise well across subjects and improve EEG prediction performance.

7.5 Summary and Conclusions

Oscar Wilde and the great writers of our time recognised the importance of context and how it shapes our understanding of language. The body of research presented in this thesis investigates the impact of context on the processing of spoken language at the cortical level. It establishes a quantitative framework for modelling contextual language processing in the brain that produces interpretable neural indices that are robustly sensitive to language comprehension. This framework was successfully employed to show that ongoing dynamics of cortical activity reflect the semantic processing of words in context in a rapid, time-locked fashion and how the use of context changes with age. The work supports the notion that a listener draws on multiple aspects of meaning from context as soon as possible and indicates that contextual influence is exerted on processing at lower cortical levels via top-down feedback connections. Furthermore, our results align with theories of predictive language processing in the brain. Finally, this body of research exemplifies how modern computational language models can be used to address questions in neurolinguistics and reveals the close ties between the two areas of research.

Bibliography

- Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H. and Merzenich, M. M. (2001) 'Speech comprehension is correlated with temporal response patterns recorded from auditory cortex', *Proceedings of the National Academy of Sciences*, 98(23), pp. 13367–13372.
- Aiken, S. J. and Picton, T. W. (2008) 'Human cortical responses to the speech envelope', *Ear and Hearing*.
- Aitchison, L. and Lengyel, M. (2017) 'With or without you: predictive coding and Bayesian inference in the brain', *Current Opinion in Neurobiology*, 46, pp. 219–227.
- Altmann, G. T. M. (1999) 'Thematic Role Assignment in Context', *Journal of Memory and Language*, 41(1), pp. 124–145.
- Altmann, G. T. M. and Kamide, Y. (1999) 'Incremental interpretation at verbs: Restricting the domain of subsequent reference', *Cognition*, 73(3), pp. 247–264.
- Anderson, A. J., Lalor, E. C., *et al.* (2019) 'Multiple Regions of a Cortical Network Commonly Encode the Meaning of Words in Multiple Grammatical Positions of Read Sentences', *Cerebral Cortex*, 29(6), pp. 2396–2411.
- Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Aguilar, M., Wang, X., Doko, D. and Raizada, R. D. S. (2017) 'Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation', *Cerebral Cortex*, 27(9), pp. 4379–4395.
- Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Raizada, R. D. S., Lin, F. and Lalor, E. C. (2019) 'An Integrated Neural Decoder of Linguistic and Experiential Meaning', *The Journal of Neuroscience*, 39(45), pp. 8969–8987.
- Anderson, A. J., Kiela, D., Clark, S. and Poesio, M. (2017) 'Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns', *Transactions of the Association for Computational Linguistics*, 5(2008), pp. 17–30.
- Aydelott, J., Jamaluddin, Z. and Nixon Pearce, S. (2015) 'Semantic processing of unattended speech in dichotic listening', *The Journal of the Acoustical Society of America*, 138(2), pp. 964–975.
- Badre, D., Poldrack, R. A., Paré-Blagoev, E. J., Insler, R. Z. and Wagner, A. D. (2005) 'Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex', *Neuron*, 47(6), pp. 907–918.
- Baggio, G. and Hagoort, P. (2011) 'The balance between memory and unification in semantics: A dynamic account of the N400', *Language and Cognitive Processes*, 26(9), pp. 1338–1367.
- Bak, T. H. and Hodges, J. R. (2004) 'The effects of motor neurone disease on language: Further evidence', *Brain and Language*.
- Baroni, M., Dinu, G. and 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)*, pp. 238–247.
- Barsalou, L. W. (1999) 'Perceptual symbol systems', *Behavioral and Brain Sciences*.
- Barsalou, L. W., Santos, A., Simmons, W. K. and Wilson, C. D. (2012) 'Language and simulation in conceptual processing', *Symbols and Embodiment: Debates on Meaning and Cognition*, (1992).

- Baumann, S., Petkov, C. I. and Griffiths, T. D. (2013) 'A unified framework for the organisation of the primate auditory cortex', *Frontiers in Systems Neuroscience*.
- Bayles, K. A., Trosset, M. W., Tomoeda, C. K., Montgomery, E. B. and Wilson, J. (1993) 'Generative naming in Parkinson disease patients', *Journal of Clinical and Experimental Neuropsychology*.
- Bell, A. H., Summerfield, C., Morin, E. L., Malecek, N. J. and Ungerleider, L. G. (2016) 'Encoding of Stimulus Probability in Macaque Inferior Temporal Cortex', *Current Biology*, 26(17), pp. 2280–2290.
- Bengio, Y., Ducharme, R., Vincent, P. and Christian Jauvin (2003) 'A Neural Probabilistic Language Model', *Journal of Machine Learning Research*, 3, pp. 1137–1155.
- Bengio, Y., Schwenk, H., Senécal, J. S., Morin, F. and Gauvain, J. L. (2006) 'Neural probabilistic language models', *Studies in Fuzziness and Soft Computing*.
- Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing', 57(1), pp. 289–300.
- Bentin, S., Kutas, M. and Hillyard, S. A. (1993) 'Electrophysiological evidence for task effects on semantic priming in auditory word processing', *Psychophysiology*, 30(2), pp. 161–169.
- Benton, A. L. and Hamsher, K. (1967) *Multilingual aphasia examination manual*. Iowa City, IA: University of Iowa.
- Van Berkum, J. J. A., Van Den Brink, D., Tesink, C. M. J. Y., Kos, M. and Hagoort, P. (2008) 'The neural integration of speaker and message', *Journal of Cognitive Neuroscience*.
- Bialek, W., Rieke, F., de Ruyter van Steveninck, R. and Warland, D. (1991) 'Reading a neural code', *Science*, 252(5014), pp. 1854–1857.
- Biesmans, W., Das, N., Francart, T. and Bertrand, A. (2017) 'Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.
- Binder, J. R. (2000) 'Human Temporal Lobe Activation by Speech and Nonspeech Sounds', *Cerebral Cortex*, 10(5), pp. 512–528.
- Binder, J. R. and Desai, R. H. (2011) 'The neurobiology of semantic memory', *Trends in Cognitive Sciences*, 15(11), pp. 527–536.
- Binder, J. R., Desai, R. H., Graves, W. W. and Conant, L. L. (2009) 'Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies', *Cerebral Cortex*, 19(12), pp. 2767–2796.
- Blank, H. and Davis, M. H. (2016) 'Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception', *PLoS Biology*, 14(11), pp. 1–32.
- Blank, H., Spangenberg, M. and Davis, M. H. (2018) 'Neural prediction errors distinguish perception and misperception of speech', *Journal of Neuroscience*, 38(27), pp. 6076–6089.
- Block, C. K. and 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), pp. 665–670.
- de Boer, E. and Kuyper, P. (1968) 'Triggered Correlation', *IEEE Transactions on Biomedical Engineering*.
- Boersma, P. and Weenink, D. (2000) *Praat: doing phonetics by computer*, <http://www.fon.hum.uva.nl/praat/>.

- Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017) 'Enriching Word Vectors with Subword Information', *Transactions of the Association for Computational Linguistics*, 5, pp. 135–146.
- Bok, S. T. (1960) 'Histonomy of the cerebral cortex.', *British Journal of Surgery*, 47(204), pp. 454–454.
- Bolla, K. I. *et al.* (2010) 'Category and Letter Fluency in Highly Educated Older Adults Category and Letter Fluency in Highly Educated Older Adults *', (December 2014), pp. 37–41.
- Bonte, M., Parviainen, T., Hytönen, K. and Salmelin, R. (2006) 'Time course of top-down and bottom-up influences on syllable processing in the auditory cortex', *Cerebral Cortex*, 16(1), pp. 115–123.
- Borkowski, J. G., Benton, A. L. and Spreen, O. (1967) 'Word fluency and brain damage', *Neuropsychologia*.
- Boulenger, V., Mechtaouff, L., Thobois, S., Broussolle, E., Jeannerod, M. and Nazir, T. A. (2008) 'Word processing in Parkinson's disease is impaired for action verbs but not for concrete nouns', *Neuropsychologia*.
- Boynton, G. M., Engel, S. A., Glover, G. H. and Heeger, D. J. (1996) 'Linear systems analysis of functional magnetic resonance imaging in human V1', *Journal of Neuroscience*.
- Bransford, J. D. and Johnson, M. K. (1972) 'Contextual prerequisites for understanding: Some investigations of comprehension and recall', *Journal of Verbal Learning and Verbal Behavior*, 11(6), pp. 717–726.
- Broadbent, D. E. (1958) 'Selective listening to speech', in *Perception and Communication*.
- Brodbeck, C., Hong, L. E. and Simon, J. Z. (2018) 'Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech', *Current Biology*, 28(1–8).
- Brodbeck, C., Presacco, A., Anderson, S. and Simon, J. Z. (2018) 'Over-Representation of Speech in Older Adults Originates from Early Response in Higher Order Auditory Cortex', *Acta Acustica united with Acustica*, 104(5), pp. 774–777.
- Brodbeck, C., Presacco, A. and Simon, J. Z. (2018) 'Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension', *NeuroImage*, 172(December 2017), pp. 162–174.
- Brodbeck, C. and Simon, J. Z. (2020) 'Continuous speech processing', *Current Opinion in Physiology*, p. 116544.
- Bronkhorst, A. W. (2000) 'The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions', *Acustica*.
- Brown, C. and Hagoort, P. (1993) 'The processing nature of the N400: Evidence from masked priming', *Journal of Cognitive Neuroscience*.
- Bruni, E., Tran, N. K. and Baroni, M. (2014) 'Multimodal distributional semantics', *Journal of Artificial Intelligence Research*.
- Buck, C., Heafield, K. and Ooyen, B. Van (2014) 'N-gram Counts and Language Models from the Common Crawl', *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 3579–3584.
- Burke, D. M. and Mackay, D. G. (1997) 'Memory, language, and ageing', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 352(1363), pp. 1845–1856.
- Burke, D. M. and Peters, L. (1986) 'Word associations in old age: evidence for consistency in

- semantic encoding during adulthood.’, *Psychology and aging*, 1(4), pp. 283–292.
- Cabeza, R., Anderson, N. D., Locantore, J. K. and McIntosh, A. R. (2002) ‘Aging gracefully: Compensatory brain activity in high-performing older adults’, *NeuroImage*, 17(3), pp. 1394–1402.
- Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M. and Knight, R. T. (2010) ‘Categorical speech representation in human superior temporal gyrus’, *Nature Neuroscience*, 13(11), pp. 1428–1432.
- Chen, S. F. and Goodman, J. (1996) ‘An empirical study of smoothing techniques for language modeling’, in *Proceedings of the 34th annual meeting on Association for Computational Linguistics* -. Morristown, NJ, USA: Association for Computational Linguistics, pp. 310–318.
- Cherry, E. C. (1953) ‘Some Experiments on the Recognition of Speech, with One and with Two Ears’, *Journal of the Acoustical Society of America*.
- de Cheveigné, A. (2010) ‘Time-shift denoising source separation’, *Journal of Neuroscience Methods*, 189(1), pp. 113–120.
- Ciccarelli, G., Nolan, M., Perricone, J., Calamia, P. T., Haro, S., O’Sullivan, J., Mesgarani, N., Quatieri, T. F. and Smalt, C. J. (2019) ‘Comparison of Two-Talker Attention Decoding from EEG with Nonlinear Neural Networks and Linear Methods’, *Scientific Reports*.
- Clark, A. (2013) ‘Whatever next? Predictive brains, situated agents, and the future of cognitive science’, *Behavioral and Brain Sciences*, 36(3), pp. 181–204.
- Coch, D., Sanders, L. D. and Neville, H. J. (2005) ‘An event-related potential study of selective auditory attention in children and adults’, *Journal of Cognitive Neuroscience*.
- Collobert, R. and Weston, J. (2008) ‘A Unified Architecture for Natural Language Processing : Deep Neural Networks with Multitask Learning’, pp. 160–167.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuska, P. (2011) ‘Natural Language Processing (Almost) from Scratch’, *Journal of Machine Learning Research*, 12, pp. 2493–2537.
- Combrisson, E. and Jerbi, K. (2015) ‘Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy’, *Journal of Neuroscience Methods*, 250, pp. 126–136.
- Connolly, J. F., Stewart, S. H. and Phillips, N. A. (1990) ‘The effects of processing requirements on neurophysiological responses to spoken sentences’, *Brain and Language*.
- Conway, A. R. A., Cowan, N. and Bunting, M. F. (2001) ‘The cocktail party phenomenon revisited: The importance of working memory capacity’, *Psychonomic Bulletin & Review*, 8(2), pp. 331–335.
- Corby, J. C. and Kopell, B. S. (1972) ‘Differential Contributions of Blinks and Vertical Eye Movements as Artifacts in EEG Recording’, *Psychophysiology*.
- Corlett, P. R., Honey, G. D., Krystal, J. H. and Fletcher, P. C. (2011) ‘Glutamatergic model psychoses: Prediction error, learning, and inference’, *Neuropsychopharmacology*, 36(1), pp. 294–315.
- Crosse, M. J., Di Liberto, G. M., Bednar, A. and Lalor, E. C. (2016) ‘The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli’, *Frontiers in Human Neuroscience*, 10(November), pp. 1–14.
- Crosse, M. J., Di Liberto, G. M. and Lalor, E. C. (2016) ‘Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal

Temporal Integration', *Journal of Neuroscience*, 36(38), pp. 9888–9895.

van de Cruys, S., Evers, K., van der Hallen, R., van Eylen, L., Boets, B., de-Wit, L. and Wagemans, J. (2014) 'Precise minds in uncertain worlds: Predictive coding in autism', *Psychological Review*, 121(4), pp. 649–675.

Crystal, T. H. and House, A. S. (1990) 'Articulation rate and the duration of syllables and stress groups in connected speech', *The Journal of the Acoustical Society of America*.

Damasio, A. R. (1989) 'The Brain Binds Entities and Events by Multiregional Activation from Convergence Zones', *Neural Computation*, 1(1), pp. 123–132.

Dambacher, M., Kliegl, R., Hofmann, M. and Jacobs, A. M. (2006) 'Frequency and predictability effects on event-related potentials during reading', *Brain Research*, 1084(1), pp. 89–103.

Daube, C., Ince, R. A. A. and Gross, J. (2019) 'Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech', *Current Biology*, 29(12), pp. 1924–1937.e9.

Davis, M. H., Ford, M. A., Kherif, F. and Johnsrude, I. S. (2011) 'Does Semantic Context Benefit Speech Understanding through "Top-Down" Processes? Evidence from Time-resolved Sparse fMRI', *Journal of Cognitive Neuroscience*, 23(12), pp. 3914–3932.

Davis, M. H. and Johnsrude, I. S. (2003) 'Hierarchical processing in spoken language comprehension.', *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 23(8), pp. 3423–3431.

Davis, M. H. and Johnsrude, I. S. (2007) 'Hearing speech sounds: Top-down influences on the interface between audition and speech perception', *Hearing Research*, 229(1–2), pp. 132–147.

Decruy, L., Vanthornhout, J. and Francart, T. (2019) 'Evidence for enhanced neural tracking of the speech envelope underlying age-related speech-in-noise difficulties', *Journal of Neurophysiology*.

DeDe, G. (2014) 'Sentence comprehension in older adults: Evidence for risky processing strategies', *Experimental Aging Research*, 40(4), pp. 436–454.

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L. and Pallier, C. (2015) 'The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees', *Neuron*, 88(1), pp. 2–19.

DeLong, K. A., Groppe, D. M., Urbach, T. P. and Kutas, M. (2012) 'Thinking ahead or not? Natural aging and anticipation during reading', *Brain and Language*, 121(3), pp. 226–239.

DeLong, K. A., Urbach, T. P. and Kutas, M. (2005) 'Probabilistic word pre-activation during language comprehension inferred from electrical brain activity', *Nature Neuroscience*, 8(8), pp. 1117–1121.

Delorme, A. and Makeig, S. (2004) 'EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis', *Journal of Neuroscience Methods*.

Deutsch, J. A. and Deutsch, D. (1963) 'Attention: Some theoretical considerations', *Psychological Review*.

Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm), pp. 4171–4186.

- DeWitt, I. and Rauschecker, J. P. (2012) 'Phoneme and word recognition in the auditory ventral stream', *Proceedings of the National Academy of Sciences*, 109(8), pp. E505–E514.
- Ding, N., Chatterjee, M. and Simon, J. Z. (2014) 'Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure', *NeuroImage*, 88, pp. 41–46.
- Ding, N., Melloni, L., Zhang, H., Tian, X. and Poeppel, D. (2015) 'Cortical tracking of hierarchical linguistic structures in connected speech', *Nature Neuroscience*, 19(1), pp. 158–164.
- Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C. and Poeppel, D. (2017) 'Temporal modulations in speech and music', *Neuroscience and Biobehavioral Reviews*, 81, pp. 181–187.
- Ding, N. and Simon, J. Z. (2012a) 'Emergence of neural encoding of auditory objects while listening to competing speakers', *Proceedings of the National Academy of Sciences*, 109(29), pp. 11854–11859.
- Ding, N. and Simon, J. Z. (2012b) 'Neural coding of continuous speech in auditory cortex during monaural and dichotic listening', *Journal of Neurophysiology*, 107(1), pp. 78–89.
- Ding, N. and Simon, J. Z. (2013) 'Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical Representation of Speech', *Journal of Neuroscience*, 33(13), pp. 5728–5735.
- Ding, N. and Simon, J. Z. (2014) 'Cortical entrainment to continuous speech: functional roles and interpretations', *Frontiers in Human Neuroscience*, 8(May), pp. 1–7.
- Doelling, K. B. and Poeppel, D. (2015) 'Cortical entrainment to music and its modulation by expertise', *Proceedings of the National Academy of Sciences*, 112(45), pp. E6233–E6242.
- Donhauser, P. W. and Baillet, S. (2019) 'Two Distinct Neural Timescales for Predictive Speech Processing', *Neuron*, pp. 1–9.
- Drennan, D. P. and Lalor, E. C. (2019) 'Cortical Tracking of Complex Sound Envelopes: Modeling the Changes in Response with Intensity', *eNeuro*.
- Drullman, R., Festen, J. M. and Plomp, R. (1994a) 'Effect of reducing slow temporal modulations on speech reception', *Journal of the Acoustical Society of America*, 95(5), pp. 2670–2680.
- Drullman, R., Festen, J. M. and Plomp, R. (1994b) 'Effect of temporal envelope smearing on speech reception', *Journal of the Acoustical Society of America*, 95(2), pp. 1053–1064.
- Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C. and Tanenhaus, M. K. (1995) 'Eye movements as a window into real-time spoken language comprehension in natural contexts', *Journal of Psycholinguistic Research*.
- Egner, T., Monti, J. M. and Summerfield, C. (2010) 'Expectation and surprise determine neural population responses in the ventral visual stream', *Journal of Neuroscience*, 30(49), pp. 16601–16608.
- Egner, T. and Summerfield, C. (2013) 'Grounding predictive coding models in empirical neuroscience research', *Behavioral and Brain Sciences*.
- Ehrlich, S. F. and Rayner, K. (1981) 'Contextual effects on word perception and eye movements during reading', *Journal of Verbal Learning and Verbal Behavior*.
- Elman, J. L. (1990) 'Finding structure in time', *Cognitive Science*, 14(2), pp. 179–211.
- Ettinger, A., Feldman, N. H., Resnik, P. and Phillips, C. (2016) 'Modeling N400 amplitude using vector space models of word representation', *Proceedings of the 38th Annual Meeting of*

- the Cognitive Science Society (CogSci 2016), 1, pp. 1445–1450.
- Federmeier, K. D. (2007) ‘Thinking ahead: The role and roots of prediction in language comprehension’, *Psychophysiology*, 44(4), pp. 491–505.
- Federmeier, K. D. and Kutas, M. (1999) ‘A Rose by Any Other Name : Long-Term Memory Structure and Sentence Processing’, *Journal of Memory and Language*, 41, pp. 469–495.
- Federmeier, K. D. and Kutas, M. (2005) ‘Aging in Context: Age-related changes in context use during comprehension’, *Psychophysiology*, 42, pp. 133–141.
- Federmeier, K. D., Kutas, M. and Schul, R. (2010) ‘Age-related and individual differences in the use of prediction during language comprehension’, *Brain and Language*, 115(3), pp. 149–161.
- Federmeier, K. D., McLennan, D. B., de Ochoa, E. and 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(2), pp. 133–146.
- Federmeier, K. D., Schwartz, T. J., Van Petten, C. and Kutas, M. (2003) ‘Sounds, Words, Sentences: Age-Related Changes Across Levels of Language Processing’, *Psychology and Aging*, 18(4), pp. 858–872.
- Feldman, H. and Friston, K. J. (2010) ‘Attention, uncertainty, and free-energy’, *Frontiers in Human Neuroscience*, 4(December), pp. 1–23.
- Ferreira, F. and Chantavarin, S. (2018) ‘Integration and Prediction in Language Processing: A Synthesis of Old and New’, *Current Directions in Psychological Science*, 27(6), pp. 443–448.
- Firth, J. R. (1957) ‘A synopsis of linguistic theory, 1930-1955’, *Studies in Linguistic Analysis*.
- Fischler, I. and 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), pp. 1–20.
- Fletcher, P. C. and Frith, C. D. (2009) ‘Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia’, *Nature Reviews Neuroscience*, 10(1), pp. 48–58.
- Fodor, J. A. (1983) ‘The modularity of mind. Cambridge’, *The MLT*.
- Foltz, P. W., Kintsch, W. and Landauer, T. K. (1998) ‘The measurement of textual coherence with latent semantic analysis’, *Discourse Processes*, 25(2–3), pp. 285–307.
- Frank, S. L., Frank, S. L. and Willems, R. M. (2017) ‘Word Embedding Distance Does not Predict Word Reading Time’, *Language, Cognition and Neuroscience*, (July), pp. 1–12.
- Frank, S. L., Otten, L. J., Galli, G. and Vigliocco, G. (2015) ‘The ERP response to the amount of information conveyed by words in sentences’, *Brain and Language*, 140, pp. 1–11.
- Frank, S. L. and Willems, R. M. (2017) ‘Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension’, *Language, Cognition and Neuroscience*, 32(9), pp. 1192–1203.
- Freeman, W. J., Holmes, M. D., Burke, B. C. and Vanhatalo, S. (2003) ‘Spatial spectra of scalp EEG and EMG from awake humans’, *Clinical Neurophysiology*.
- Friederici, A. D. (2011) ‘The Brain Basis of Language Processing: From Structure to Function’, *Physiological Reviews*, 91(4), pp. 1357–1392.
- Friston, K. (2005) ‘A theory of cortical responses’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456), pp. 815–836.

- Friston, K. (2018) 'Does predictive coding have a future?', *Nature Neuroscience*, 21(8), pp. 1019–1021.
- Fu, R., Guo, J., Qin, B., Che, W., Wang, H. and Liu, T. (2014) 'Learning semantic hierarchies via word embeddings', 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference, 1, pp. 1199–1209.
- Gagnepain, P., Henson, R. N. and Davis, M. H. (2012) 'Temporal predictive codes for spoken words in auditory cortex', *Current Biology*, 22(7), pp. 615–621.
- Ganis, G., Kutas, M. and Sereno, M. I. (1996) 'The Search for "Common Sense": An Electrophysiological Study of the Comprehension of Words and Pictures in Reading', *Journal of Cognitive Neuroscience*.
- Ganong, W. F. (1980) 'Phonetic categorization in auditory word perception.', *Journal of Experimental Psychology: Human Perception and Performance*, 6(1), pp. 110–125.
- Garofolo, J. S. (1993) 'TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1'. Philadelphia: Linguistic Data Consortium.
- Gernsbacher, M. A. (1991) 'Cognitive Processes and Mechanisms in Language Comprehension: The Structure Building Framework', *Psychology of Learning and Motivation - Advances in Research and Theory*.
- Gorman, K., Howell, J. and Wagner, M. (2011) 'Prosodylab-aligner: A tool for forced alignment of laboratory speech', *Canadian Acoustics*.
- Greenwood, D. D. (1961) 'Auditory Masking and the Critical Band', *The Journal of the Acoustical Society of America*.
- Grigor, J., Van Toller, S., Behan, J. and Richardson, A. (1999) 'The effect of odour priming on long latency visual evoked potentials of matching and mismatching objects', *Chemical Senses*.
- Grosjean, F. (1980) 'Spoken word recognition processes and the gating paradigm', *Perception & Psychophysics*.
- Grossman, M., Anderson, C., Khan, A., Avants, B., Elman, L. and McCluskey, L. (2008) 'Impaired action knowledge in amyotrophic lateral sclerosis', *Neurology*.
- Hackett, T. A., Preuss, T. M. and Kaas, J. H. (2001) 'Architectonic identification of the core region in auditory cortex of macaques, chimpanzees, and humans', *Journal of Comparative Neurology*.
- Hagoort, P., Baggio, G. and Willems, R. M. (2009) 'Semantic Unification', *The Cognitive Neurosciences*, pp. 819–836.
- Hagoort, P. and Brown, C. M. (2000a) 'ERP effects of listening to speech: Semantic ERP effects', *Neuropsychologia*, 38(11), pp. 1518–1530.
- Hagoort, P. and Brown, C. M. (2000b) 'ERP effects of listening to speech compared to reading: The P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation', *Neuropsychologia*.
- Hagoort, P., Hald, L., Bastiaansen, M. and Petersson, K. M. (2004) 'Integration of Word Meaning and World Knowledge in Language Comprehension', *Science*, 304(5669), pp. 438–441.
- Hale, J. T. (2001) 'A Probabilistic Earley Parser as a Psycholinguistic Model', *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*.

- Hamilton, L. S. and Huth, A. G. (2018) 'The revolution will not be controlled: natural stimuli in speech neuroscience', *Language, Cognition and Neuroscience*, 0(0), pp. 1–10.
- Harnad, S. (1990) 'The symbol grounding problem', *Physica D: Nonlinear Phenomena*.
- Harris, Z. S. (1954) 'Distributional Structure', *WORD*.
- Haxby, J. V. (2001) 'Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex', *Science*, 293(5539), pp. 2425–2430.
- de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L. and Theunissen, F. E. (2017) 'The hierarchical cortical organization of human speech processing', *Journal of Neuroscience*, 37(27), pp. 6539–6557.
- Heilbron, M. and Chait, M. (2017) 'Great expectations: Is there evidence for predictive coding in auditory cortex?', *Neuroscience*.
- Heilbron, M., Ehinger, B., Hagoort, P. and Lange, F. P. De (2019) 'Tracking Naturalistic Linguistic Predictions with Deep Neural Language Models'.
- Helmholtz, H. (no date) *Handbuch der physiologischen optik*. New York: Dover (English trans.).
- Hemingway, E. (1952) *The Old Man and the Sea*, New York: Scribner. Charles Scribner's Sons.
- Herculano-Houzel, S. (2009) 'The human brain in numbers: A linearly scaled-up primate brain', *Frontiers in Human Neuroscience*.
- Hickok, G. and Poeppel, D. (2004) 'Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language', *Cognition*, 92(1–2), pp. 67–99.
- Hickok, G. and Poeppel, D. (2007) 'The cortical organization of speech processing', *Nature Reviews Neuroscience*, 8(May), pp. 393–402.
- Hink, R. F. and Hillyard, S. A. (1976) 'Auditory evoked potentials during selective listening to dichotic speech messages', *Perception & Psychophysics*.
- Hochreiter, S. and Schmidhuber, J. (1997) 'Long Short-Term Memory', *Neural Computation*.
- Holcomb, P. J. and Neville, H. J. (1990) 'Auditory and Visual Semantic Priming in Lexical Decision: A Comparison Using Event-related Brain Potentials Auditory and Visual Semantic Priming in Lexical Decision: A Comparison Using Event-related Brain Potentials', *Language and Cognitive Processes*, 5(4), pp. 281–312.
- Holdgraf, C. R., De Heer, W. A., Pasley, B. N., Rieger, J., Crone, N., Lin, J. J., Knight, R. T. and Theunissen, F. E. (2016) 'Rapid tuning shifts in human auditory cortex enhance speech intelligibility', *Nature Communications*, 7(May).
- Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T. and Theunissen, F. E. (2017) 'Encoding and Decoding Models in Cognitive Electrophysiology', *Frontiers in Systems Neuroscience*, 11(September).
- Howard, M. F. and Poeppel, D. (2010) 'Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension', *Journal of Neurophysiology*.
- Huetig, F. and Mani, N. (2016) 'Is prediction necessary to understand language? Probably not', *Language, Cognition and Neuroscience*, 31(1), pp. 19–31.
- Humes, L. E., Dubno, J. R., Gordon-Salant, S., Lister, J. J., Cacace, A. T., Cruickshanks, K. J., Gates, G. A., Wilson, R. H. and Wingfield, A. (2012) 'Central presbycusis: A review and

evaluation of the evidence', *Journal of the American Academy of Audiology*.

Humphries, C. J., Binder, J. R., Medler, D. A. and Liebenthal, E. (2006) 'Syntactic and semantic modulation of neural activity during auditory sentence comprehension', *Journal of Cognitive Neuroscience*.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E. and Gallant, J. L. (2016) 'Natural speech reveals the semantic maps that tile human cerebral cortex', *Nature*, 532(7600), pp. 453–458.

Huth, A. G., Lee, T., Nishimoto, S., Bilenko, N. Y., Vu, A. T. and Gallant, J. L. (2016) 'Decoding the Semantic Content of Natural Movies from Human Brain Activity', *Frontiers in Systems Neuroscience*.

Hyvärinen, A. (1999) 'Fast and robust fixed-point algorithms for independent component analysis', *IEEE Transactions on Neural Networks*.

Jackendoff, R. (2003) 'Précis of foundations of language: Brain, meaning, grammar, evolution', *Behavioral and Brain Sciences*.

Jain, S. and Huth, A. G. (2018) 'Incorporating Context into Language Encoding Models for fMRI', *The advances in Neural Information Processing Systems 31, (Nips)*, p. 327601.

Joos, M. (1950) 'Description of Language Design', *Journal of the Acoustical Society of America*.

Jurafsky, D. (1996) 'A probabilistic model of lexical and syntactic access and disambiguation', *Cognitive Science*, 20(2), pp. 137–194.

Kamide, Y., Altmann, G. T. M. and Haywood, S. L. (2003) 'The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements', *Journal of Memory and Language*.

Keitel, A., Gross, J. and Kayser, C. (2018) 'Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features', *PLoS Biology*.

Kersten, D., Mamassian, P. and Yuille, A. (2004) 'Object perception as Bayesian inference', *Annual Review of Psychology*, 55, pp. 271–304.

Khalighinejad, B., Cruzatto da Silva, G. and Mesgarani, N. (2017) 'Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech', *The Journal of Neuroscience*, 37(8), pp. 2176–2185.

Kiela, D. and Clark, S. (2014) 'A Systematic Study of Semantic Vector Space Model Parameters', *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality CVSC EACL*, Gothenburg, Sweden, April 2014, 353, pp. 21–30. Available at: papers3://publication/uuid/A791976B-7D71-41C6-9D47-C12C5436DFCF.

Kintsch, W. and van Dijk, T. A. (1978) 'Toward a model of text comprehension and production', *Psychological Review*.

Kiss, M., Cristescu, T., Fink, M. and Wittmann, M. (2008) 'Auditory language comprehension of temporally reversed speech signals in native and non-native speakers', *Acta Neurobiologiae Experimentalis*.

Kivisaari, S. L., van Vliet, M., Hultén, A., Lindh-Knuutila, T., Faisal, A. and Salmelin, R. (2019) 'Reconstructing meaning from bits of information', *Nature Communications*, 10(1), pp. 1–11.

Kliegl, R., Grabner, E., Rolfs, M. and Engbert, R. (2004) 'Length, frequency, and predictability effects of words on eye movements in reading', *European Journal of Cognitive Psychology*,

16(1–2), pp. 262–284.

Knill, D. C. and Pouget, A. (2004) ‘The Bayesian brain: The role of uncertainty in neural coding and computation’, *Trends in Neurosciences*, 27(12), pp. 712–719.

Kok, P., Jehee, J. F. M. and de Lange, F. P. (2012) ‘Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex’, *Neuron*, 75(2), pp. 265–270.

Kozora, E. and Cullum, C. M. (1995) ‘Generative Naming in Normal Aging: Total Output and Qualitative Changes Using Phonemic and Semantic Constraints’, *The Clinical Neuropsychologist*, 9(4), pp. 313–320.

Kriegeskorte, N. (2008) ‘Representational similarity analysis – connecting the branches of systems neuroscience’, *Frontiers in Systems Neuroscience*, 2(November), pp. 1–28.

Kullback, S. and Leibler, R. A. (1951) ‘On Information and Sufficiency’, *The Annals of Mathematical Statistics*.

Kumar, S., Kaposvari, P. and Vogels, R. (2017) ‘Encoding of predictable and unpredictable stimuli by inferior temporal cortical neurons’, *Journal of Cognitive Neuroscience*.

Kuperberg, G. R. and Jaeger, T. F. (2016) ‘What do we mean by prediction in language comprehension?’, *Language, Cognition and Neuroscience*, 31(1), pp. 32–59.

Kuperberg, G. R., Sitnikova, T., Caplan, D. and Holcomb, P. J. (2003) ‘Electrophysiological distinctions in processing conceptual relationships within simple sentences’, *Cognitive Brain Research*, 17(1), pp. 117–129.

Kutas, M. (1993) ‘In the company of other words: Electrophysiological evidence for single-word and sentence context effects’, *Language and Cognitive Processes*, 8(4), pp. 533–572.

Kutas, M., DeLong, K. A. and Smith, N. J. (2011) ‘A Look around at What Lies Ahead: Prediction and Predictability in Language Processing’, *Predictions in the Brain: Using Our Past to Generate a Future*, pp. 190–207.

Kutas, M. and Federmeier, K. D. (2000) ‘Electrophysiology reveals semantic memory use in language comprehension’, *Trends in Cognitive Sciences*, 4(12), pp. 463–470.

Kutas, M. and Federmeier, K. D. (2007) ‘Event-related brain potential (ERP) studies of sentence processing’, *The Oxford Handbook of Psycholinguistics*, pp. 385–406.

Kutas, M. and 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(1), pp. 621–647.

Kutas, M. and Hillyard, S. A. (1980a) ‘Event-related brain potentials to semantically inappropriate and surprisingly large words’, *Biological Psychology*.

Kutas, M. and Hillyard, S. A. (1980b) ‘Reading senseless sentences: Brain potentials reflect semantic incongruity’, *Science*.

Kutas, M. and Hillyard, S. A. (1983) ‘Event-related brain potentials to grammatical errors and semantic anomalies’, *Memory & Cognition*.

Kutas, M. and Hillyard, S. A. (1984) ‘Brain Potentials During Reading reflect word expectancy and semantic association’, *Nature*, 307, pp. 161–163.

Kutas, M. and Iragui, V. (1998) ‘The N400 in a semantic categorization task across 6 decades’, *Electroencephalography and Clinical Neurophysiology - Evoked Potentials*, 108(5), pp. 456–471.

Kutas, M., Neville, H. J. and Holcomb, P. J. (1987) ‘A preliminary comparison of the N400

response to semantic anomalies during reading, listening and signing.’, *Electroencephalography and clinical neurophysiology. Supplement*.

Lachter, J., Forster, K. I. and Ruthruff, E. (2004) ‘Forty-five years after broadbent (1958): Still no identification without attention’, *Psychological Review*.

Lalor, E. C. and Foxe, J. J. (2010) ‘Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution’, *European Journal of Neuroscience*, 31(1), pp. 189–193.

Lalor, E. C., Power, A. J., Reilly, R. B. and Foxe, J. J. (2009) ‘Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli’, *Journal of Neurophysiology*, 102(1), pp. 349–359.

Landauer, T. K. and Dumais, S. T. (1997) ‘A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge’, *Psychological Review*.

de Lange, F. P., Heilbron, M. and Kok, P. (2018) ‘How Do Expectations Shape Perception?’, *Trends in Cognitive Sciences*, 22(9), pp. 764–779.

Laszlo, S. and Federmeier, K. D. (2011) ‘The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects’, *Psychophysiology*, 48(2), pp. 176–186.

Lau, E. F., Holcomb, P. J. and Kuperberg, G. R. (2013) ‘Dissociating N400 Effects of Prediction from Association in Single-word Contexts’, *Journal of Cognitive Neuroscience*, 25(3), pp. 484–502.

Lau, E. F., Phillips, C. and Poeppel, D. (2008) ‘A cortical network for semantics: (De)constructing the N400’, *Nature Reviews Neuroscience*, 9(12), pp. 920–933.

Lee, T. S. and Mumford, D. (2003) ‘Hierarchical Bayesian inference in the visual cortex’, *Journal of the Optical Society of America A*, 20(7), p. 1434.

Chomsky, N. (1957) ‘Syntactic Structures’, Mouton.

Leonard, M. K., Baud, M. O., Sjerps, M. J. and Chang, E. F. (2016) ‘Perceptual restoration of masked speech in human cortex’, *Nature Communications*, 7, pp. 1–9.

Lerner, Y., Honey, C. J., Silbert, L. J. and Hasson, U. (2011) ‘Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story’, *Journal of Neuroscience*, 31(8), pp. 2906–2915.

Levy, R. (2008) ‘Expectation-based syntactic comprehension’, *Cognition*.

Lieberman, A. M., Cooper, F. S., Shankweiler, D. P. and Studdert-Kennedy, M. (1967) ‘Perception of the speech code’, *Psychological Review*.

Di Liberto, G. M., Crosse, M. J. and Lalor, E. C. (2018) ‘Cortical Measures of Phoneme-Level Speech Encoding Correlate with the Perceived Clarity of Natural Speech’, *Eneuro*, 5(2), p. ENEURO.0084-18.2018.

Di Liberto, G. M. and Lalor, E. C. (2017) ‘Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research’, *Hearing Research*, 348, pp. 70–77.

Di Liberto, G. M., Nie, J., Yeaton, J., Khalighinejad, B., Shamma, S. and Mesgarani, N. (2020) ‘Neural representation of linguistic feature hierarchy reflects second-language proficiency’, *bioRxiv*, p. 2020.06.15.142554.

- Di Liberto, G. M., O'Sullivan, J. A. and Lalor, E. C. (2015) 'Low-frequency cortical entrainment to speech reflects phoneme-level processing', *Current Biology*, 25(19), pp. 2457–2465.
- Di Liberto, G. M., Wong, D., Melnik, G. A. and de Cheveigné, A. (2019) 'Low-frequency cortical responses to natural speech reflect probabilistic phonotactics', *NeuroImage*, 196(March), pp. 237–247.
- Lieberman, P. (1963) 'Some Effects of Semantic and Grammatical Context on the Production and Perception of Speech', *Language and Speech*, 6(3), pp. 172–187.
- Lin, M. and Lucas, H. C. (2013) 'Too Big to Fail : Large Samples and the p -Value Problem', *Information Systems Research*, 7047(August 2016), pp. 1–12.
- Lotto, A. and Holt, L. (2000) 'The Illusion of the Phoneme', *Department of Psychology*, (March 2003), p. 166.
- Luck, S. J. (2014) 'An Introduction to the Event-Related Potential Techniques', MIT Press, Cambridge, MA.
- Lund, K. and Burgess, C. (1996) 'Producing high-dimensional semantic spaces from lexical co-occurrence', *Behavior Research Methods, Instruments, and Computers*, 28(2), pp. 203–208.
- Luo, H. and Poeppel, D. (2007) 'Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex', *Neuron*, 54(6), pp. 1001–1010.
- Maris, E. and Oostenveld, R. (2007) 'Nonparametric statistical testing of EEG- and MEG-data', *Journal of Neuroscience Methods*, 164(1), pp. 177–190.
- Marmarelis, V. Z. (2004) *Nonlinear dynamic modeling of physiological systems*, Nonlinear Dynamic Modeling of Physiological Systems.
- Marslen-Wilson, W. D. (1973) 'Linguistic Structure and Speech Shadowing at Very Short Latencies', *Nature*, 244(5417), pp. 522–523.
- Marslen-Wilson, W. D. (1975) 'Sentence Perception as an Interactive Parallel Process', *Science*, 189(4198), pp. 226–228.
- Marslen-Wilson, W. D. (1987) 'Functional parallelism in spoken word-recognition', *Cognition*, 25(1–2), pp. 71–102.
- McClelland, J. L. and Elman, J. L. (1986) 'The TRACE model of speech perception', *Cognitive Psychology*, 18(1), pp. 1–86.
- McClelland, J. L., Mirman, D. and Holt, L. L. (2006) 'Are there interactive processes in speech perception?', *Trends in Cognitive Sciences*, 10(8), pp. 363–369.
- Mesgarani, N. and Chang, E. F. (2012) 'Selective cortical representation of attended speaker in multi-talker speech perception', *Nature*, 485(7397), pp. 233–236.
- Mesgarani, N., Cheung, C., Johnson, K. and Chang, E. F. (2014) 'Phonetic feature encoding in Human Superior Temporal Gyrus', *Science*, 343(January), pp. 1–6.
- Mesgarani, N., David, S. V., Fritz, J. B. and Shamma, S. A. (2008) 'Phoneme representation and classification in primary auditory cortex', *The Journal of the Acoustical Society of America*, 123(2), pp. 899–909.
- Meteyard, L., Cuadrado, S. R., Bahrami, B. and Vigliocco, G. (2012) 'Coming of age: A review of embodiment and the neuroscience of semantics', *Cortex*, 48(7), pp. 788–804.
- Mezei, A. and Paivio, A. (1972) 'Imagery and Verbal Processes', *Leonardo*.

- Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013a) ‘Distributed Representations of Words and Phrases and Their Compositionality’, *Proc. Advances in Neural Information Processing Systems*.
- Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013b) ‘Efficient Estimation of Word Representations in Vector Space’, pp. 1–12.
- Miller, G. A., Heise, G. A. and Lichten, W. (1951) ‘The Intelligibility of Speech as a Function of the Context of the Test Materials’, *Journal of Experimental Psychology*, 41(5), pp. 329–335.
- Mitchell, J. and Lapata, M. (2010) ‘Composition in Distributional Models of Semantics’, *Cognitive Science*, 34(8), pp. 1388–1429.
- Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A. and Just, M. A. (2008) ‘Predicting human brain activity associated with the meanings of nouns’, *Science*, 320(5880), pp. 1191–1195.
- Mollica, F., Siegelman, M., Diachek, E., Piantadosi, S. T., Mineroff, Z., Futrell, R. and Fedorenko, E. (2018) ‘High local mutual information drives the response in the human language network’.
- Monsalve, I. F., Frank, S. L. and Vigliocco, G. (2012) ‘Lexical surprisal as a general predictor of reading time’, *Eacl 2012*, p. 398. Available at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.230.9321&rep=rep1&type=pdf#page=422%5Cnpapers2://publication/uuid/B7889F7D-5127-4BA3-9589-9EE2E415CAC0>.
- Morton, J., Marcus, S. and Frankish, C. (1976) ‘Perceptual centers (P-centers)’, *Psychological Review*, 83(5), pp. 405–408.
- Mumford, D. (1992) ‘On the computational architecture of the neocortex’, *Biological Cybernetics*, 66(3), pp. 241–251.
- Nager, W., Dethlefsen, C. and Münte, T. F. (2008) ‘Attention to human speakers in a virtual auditory environment: Brain potential evidence’, *Brain Research*.
- Nieuwland, M. *et al.* (2018) ‘Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials’, p. 267815.
- Nieuwland, M. S. *et al.* (2018) ‘Large-scale replication study reveals a limit on probabilistic prediction in language comprehension’, *eLife*, 7, pp. 1–24.
- Nieuwland, M. S. and van Berkum, J. J. A. (2006) ‘When Peanuts Fall in Love’, *Journal of Cognitive Neuroscience*, 18(7), pp. 1098–1111.
- Norman-Haignere, S., Kanwisher, N. and McDermott, J. H. (2013) ‘Cortical Pitch Regions in Humans Respond Primarily to Resolved Harmonics and Are Located in Specific Tonotopic Regions of Anterior Auditory Cortex’, *Journal of Neuroscience*, 33(50), pp. 19451–19469.
- Norris, D. and McQueen, J. M. (2008) ‘Shortlist B: A Bayesian Model of Continuous Speech Recognition’, *Psychological Review*, 115(2), pp. 357–395.
- Norris, D., McQueen, J. M. and Cutler, A. (2000) ‘Merging information in speech processing: Feedback is never necessary’, *Behavioral and Brain Sciences*, 23(3), pp. 299–370.
- Norris, D., McQueen, J. M. and Cutler, A. (2016) ‘Prediction, Bayesian inference and feedback in speech recognition’, *Language, Cognition and Neuroscience*.
- O’Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A. and Lalor, E. C. (2015) ‘Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG’, *Cerebral Cortex*, 25(7), pp. 1697–1706.

Obleser, J. (2014) 'Putting the Listening Brain in Context', *Linguistics and Language Compass*, 8(12), pp. 646–658.

Octave, E. and Reichenbach, T. (2019) 'Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise', *Journal of Neuroscience*.

Oganian, Y. and Chang, E. F. (2018) 'A speech envelope landmark for syllable encoding in human superior temporal gyrus', *bioRxiv*, pp. 1134–1140.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I. H., Saberi, K., Serences, J. T. and Hickok, G. (2010) 'Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech', *Cerebral Cortex*, 20(10), pp. 2486–2495.

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K. (2016) 'WaveNet: A Generative Model for Raw Audio', pp. 1–15. Available at: <http://arxiv.org/abs/1609.03499>.

Overath, T., McDermott, J. H., Zarate, J. M. and Poeppel, D. (2015) 'The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts', *Nature Neuroscience*, 18(6), pp. 903–911.

Parsons, T. (1987) *Voice and Speech Processing*. McGraw-Hill.

Pasquier, F., Lebert, F., Grymonprez, L. and Petit, H. (1995) 'Verbal fluency in dementia of frontal lobe type and dementia of Alzheimer type', *Journal of Neurology Neurosurgery and Psychiatry*.

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S. and Griffiths, T. D. (2002) 'The processing of temporal pitch and melody information in auditory cortex', *Neuron*.

Payne, B. R. and Federmeier, K. D. (2017) 'Event-related brain potentials reveal age-related changes in parafoveal-foveal integration during sentence processing', *Neuropsychologia*, 106(October 2016), pp. 358–370.

Payne, B. R. and Silcox, J. W. (2019) *Aging, context processing, and comprehension*. 1st edn, *Psychology of Learning and Motivation - Advances in Research and Theory*. 1st edn. Elsevier Inc.

Peelle, J. E. (2018) 'Listening Effort', *Ear and Hearing*, 39(2), pp. 204–214.

Peelle, J. E. (2019) 'Language and Aging', *The Oxford Handbook of Neurolinguistics*.

Peelle, J. E. and Davis, M. H. (2012) 'Neural oscillations carry speech rhythm through to comprehension', *Frontiers in Psychology*, 3(SEP), pp. 1–17.

Peelle, J. E., Gross, J. and Davis, M. H. (2013) 'Phase-locked responses to speech in human auditory cortex are enhanced during comprehension', *Cerebral Cortex*, 23(6), pp. 1378–1387.

Peelle, J. E., Troiani, V., Wingfield, A. and Grossman, M. (2010) 'Neural processing during older adults' comprehension of spoken sentences: Age differences in resource allocation and connectivity', *Cerebral Cortex*, 20(4), pp. 773–782.

Pellicano, E. and Burr, D. (2012) 'When the world becomes "too real": A Bayesian explanation of autistic perception', *Trends in Cognitive Sciences*, 16(10), pp. 504–510.

Peña, M. and Melloni, L. (2012) 'Brain oscillations during spoken sentence processing', *Journal of Cognitive Neuroscience*.

Pennington, J., Socher, R. and Manning, C. (2014) 'Glove: Global Vectors for Word Representation', in *Proceedings of the 2014 Conference on Empirical Methods in Natural*

- Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1532–1543.
- Van Petten, C., Coulson, S., Rubin, S., Plante, E. and Parks, M. (1999a) ‘Time Course of Word Identification and Semantic Integration in Spoken Language’, *Journal of Experimental Psychology: Learning Memory and Cognition*, 25(2), pp. 394–417.
- Van Petten, C., Coulson, S., Rubin, S., Plante, E. and Parks, M. (1999b) ‘Time Course of Word Identification and Semantic Integration in Spoken Language’, *Journal of Experimental Psychology: Learning Memory and Cognition*, 25(2), pp. 394–417.
- Pichora-Fuller, M. K., Schneider, B. A. and Daneman, M. (1995) ‘How young and old adults listen to and remember speech in noise’, *The Journal of the Acoustical Society of America*, 97(1), pp. 593–608.
- Picton, T. W. (2013) ‘Hearing in time: Evoked potential studies of temporal processing’, *Ear and Hearing*, 34(4), pp. 385–401.
- Plomp, R. and Mimpen, A. M. (1979) ‘Improving the reliability of testing the speech reception threshold for sentences’, *International Journal of Audiology*.
- Poeppel, D., Emmorey, K., Hickok, G. and Pylkkänen, L. (2012) ‘Towards a new neurobiology of language’, *Journal of Neuroscience*, 32(41), pp. 14125–14131.
- Poeppel, D., Idsardi, W. J. and Van Wassenhove, V. (2008) ‘Speech perception at the interface of neurobiology and linguistics’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493), pp. 1071–1086.
- Poullisse, C., Wheeldon, L. and Segaert, K. (2019) ‘Evidence Against Preserved Syntactic Comprehension in Healthy Aging’, *Journal of Experimental Psychology: Learning Memory and Cognition*.
- Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B. and Lalor, E. C. (2012) ‘At what time is the cocktail party? A late locus of selective attention to natural speech’, *European Journal of Neuroscience*, 35(9), pp. 1497–1503.
- Powers, A. R., Mathys, C. and Corlett, P. R. (2017) ‘Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors’, *Science*, 357(6351), pp. 596–600.
- Presacco, A., Simon, J. Z. and Anderson, S. (2016) ‘Evidence of degraded representation of speech in noise, in the aging midbrain and cortex’, *Journal of Neurophysiology*, 116(5), pp. 2346–2355.
- Press, C., Kok, P. and Yon, D. (2020) ‘The Perceptual Prediction Paradox’, *Trends in Cognitive Sciences*, 24(1), pp. 13–24.
- Price, C. J. (2010) ‘The anatomy of language: A review of 100 fMRI studies published in 2009’, *Annals of the New York Academy of Sciences*.
- Pynte, J., New, B. and Kennedy, A. (2009) ‘On-line contextual influences during reading normal text: The role of nouns, verbs and adjectives’, *Vision Research*.
- Rabovsky, M., Hansen, S. S. and McClelland, J. L. (2018) ‘Modelling the N400 brain potential as change in a probabilistic representation of meaning’, *Nature Human Behaviour*, 2(9), pp. 693–705.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019) ‘Language Models are Unsupervised Multitask Learners Alec’, *ArXiv Preprint*.
- Ralph, M. A. L., Jefferies, E., Patterson, K. and Rogers, T. T. (2016) ‘The neural and

- computational bases of semantic cognition', *Nature Reviews Neuroscience*, 18(1), pp. 42–55.
- Rao, R. P. N. and Ballard, D. H. (1999) 'Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects.', *Nature neuroscience*, 2 VN-re(1), pp. 79–87.
- Rauschecker, J. P. and Scott, S. K. (2009) 'Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing', *Nature Neuroscience*, 12(6), pp. 718–724.
- Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C. and Pollatsek, A. (2006) 'The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers', *Psychology and Aging*, 21(3), pp. 448–465.
- Roark, B. (2001) 'Probabilistic top-down parsing and language modeling', *Computational Linguistics*.
- Rodd, J. M., Davis, M. H. and Johnsrude, I. S. (2005) 'The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity', *Cerebral Cortex*, 15(8), pp. 1261–1269.
- Rugg, M. D. (1984) 'Event-related potentials and the phonological processing of words and non-words', *Neuropsychologia*.
- Rugg, M. D. (1987) 'Dissociation of Semantic Priming, Word and Non-Word Repetition Effects by Event-Related Potentials', *The Quarterly Journal of Experimental Psychology Section A*.
- Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) 'Learning representations by back-propagating errors', *Nature*, 323(6088), pp. 533–536.
- Saberi, K. and Perrott, D. R. (1999) 'Cognitive restoration of reversed speech', *Nature*.
- Sacks, H., Schegloff, E. A. and Jefferson, G. (1974) 'A Simplest Systematics for the Organization of Turn-Taking for Conversation', *Language*.
- Saenz, M. and Langers, D. R. M. (2014) 'Tonotopic mapping of human auditory cortex', *Hearing Research*.
- Salmelin, R. (2007) 'Clinical neurophysiology of language: The MEG approach', *Clinical Neurophysiology*, 118(2), pp. 237–254.
- Salthouse, T. A. (1994) 'The Aging of Working Memory', *Neuropsychology*, 8(4), pp. 535–543.
- Salthouse, T. A. (1996) 'The processing-speed theory of adult age differences in cognition.', *Psychological Review*, 103(3), pp. 403–428.
- Scharenborg, O., Norris, D., ten Bosch, L. and McQueen, J. M. (2005) 'How should a speech recognizer work?', *Cognitive Science*.
- Schnupp, J., Nelken, I. and King, A. J. (2019) *Auditory Neuroscience*, Auditory Neuroscience.
- Schrimpf, M., Blank, I., Tuckute, G., Kauf, C. and Hosseini, E. A. (2019) 'Artificial Neural Networks Accurately Predict Language Processing in the Brain'.
- Scott, S. K. and Johnsrude, I. S. (2003) 'The neuroanatomical and functional organization of speech perception', *Trends in Neurosciences*, 26(2), pp. 100–107.
- Searle, J. R. (1980) 'Minds, brains, and programs', *Behavioral and Brain Sciences*.
- Shannon, C. E. (1951) 'Prediction and Entropy of Printed English'.
- Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J. and Ekelid, M. (1995) 'Speech

recognition with primarily temporal cues', *Science*.

Simpson, G. B. (1991) *Understanding Word and Sentence*. Elsevier.

Sitnikova, T., Holcomb, P. J., Kiyonaga, K. A. and Kuperberg, G. R. (2008) 'Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events', *Journal of Cognitive Neuroscience*.

Siuly, S., Li, Y. and Zhang, Y. (2016) 'Electroencephalogram (EEG) and Its Background', in, pp. 3–21.

Smith, N. J. and Levy, R. (2011) 'Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing', *Proceedings of the 33rd Annual Meeting of the Cognitive Science Conference*, pp. 1637–1642.

Smith, N. J. and Levy, R. (2013) 'The effect of word predictability on reading time is logarithmic', *Cognition*, 128(3), pp. 302–319.

Snowden, J. S. (1989) 'Semantic dementia: a form of circumscribed cerebral atrophy. *Behavioural Neurology*, 2, 167–182', *Behavioural Neurology*.

Sohoglu, E. and Davis, M. H. (2020) 'Rapid computations of spectrotemporal prediction error support perception of degraded speech'.

Sohoglu, E., Peelle, J. E., Carlyon, R. P. and Davis, M. H. (2012) 'Predictive Top-Down Integration of Prior Knowledge during Speech Perception', *Journal of Neuroscience*, 32(25), pp. 8443–8453.

St George, M., Kutas, M., Martinez, A. and Sereno, M. I. (1999) 'Semantic integration in reading: Engagement of the right hemisphere during discourse processing', *Brain*, 122(7), pp. 1317–1325.

Staub, A. (2015) 'The Effect of Lexical Predictability on Eye Movements in Reading: Critical Review and Theoretical Interpretation', *Language and Linguistics Compass*, 9(8), pp. 311–327.

Sterzer, P. *et al.* (2018) 'The Predictive Coding Account of Psychosis', *Biological Psychiatry*, 84(9), pp. 634–643.

Stine-Morrow, E. A. L., Miller, L. M. S., Gagne, D. D. and Hertzog, C. (2008) 'Self-Regulated Reading in Adulthood', *Psychology and Aging*.

Stine-Morrow, E. A. L., Soederberg Miller, L. M. and Nevin, J. A. (1999) 'The effects of context and feedback on age differences in spoken word recognition', *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*.

Strauss, A., Kotz, S. A. and Obleser, J. (2013) 'Narrowed Expectancies under Degraded Speech: Revisiting the N400', *Journal of Cognitive Neuroscience*, 25(8), pp. 1383–1395.

Sumby, W. H. and Pollack, I. (1954) 'Visual Contribution to Speech Intelligibility in Noise', *The Journal of the Acoustical Society of America*.

Summerfield, C. and De Lange, F. P. (2014) 'Expectation in perceptual decision making: Neural and computational mechanisms', *Nature Reviews Neuroscience*, 15(11), pp. 745–756.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. and Sedivy, J. C. (1995) 'Integration of visual and linguistic information in spoken language comprehension', *Science*.

Taylor, W. L. (1953) "'Cloze Procedure': A New Tool for Measuring Readability", *Journalism Bulletin*.

Teoh, E. S. and Lalor, E. C. (2019) 'EEG decoding of the target speaker in a cocktail party scenario: Considerations regarding dynamic switching of talker location', *Journal of Neural*

Engineering.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K. and Farah, M. J. (1997) 'Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation', *Proceedings of the National Academy of Sciences*, 94(26), pp. 14792–14797.

Thompson-Schill, S. L., D'Esposito, M. and Kan, I. P. (1999) 'Effects of repetition and competition on activity in left prefrontal cortex during word generation', *Neuron*, 23(3), pp. 513–522.

Todorovic, A. and de Lange, F. P. (2012) 'Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields', *Journal of Neuroscience*, 32(39), pp. 13389–13395.

TREISMAN, A. M. (1964) 'VERBAL CUES, LANGUAGE, AND MEANING IN SELECTIVE ATTENTION.', *The American journal of psychology*.

Tune, S., Alavash, M., Fiedler, L. and Obleser, J. (2020) 'Neural attention filters do not predict behavioral success in a large cohort of aging listeners', *bioRxiv*, p. 2020.05.20.105874.

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z. and Francart, T. (2018) 'Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope', *JARO - Journal of the Association for Research in Otolaryngology*, 19(2), pp. 181–191.

Verschueren, E., Somers, B. and Francart, T. (2019) 'Neural envelope tracking as a measure of speech understanding in cochlear implant users', *Hearing Research*.

Wacongne, C., Labyt, E., Van Wassenhove, V., Bekinschtein, T., Naccache, L. and Dehaene, S. (2011) 'Evidence for a hierarchy of predictions and prediction errors in human cortex', *Proceedings of the National Academy of Sciences of the United States of America*, 108(51), pp. 20754–20759.

Wagner, A. D., Paré-Blagoev, E. J., Clark, J. and Poldrack, R. A. (2001) 'Recovering Meaning', *Neuron*, 31(2), pp. 329–338.

Wang, L., Kuperberg, G. R. and Jensen, O. (2018) 'Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity', *eLife*, pp. 1–24.

Wang, X., Men, W., Gao, J., Caramazza, A. and Bi, Y. (2020) 'Two Forms of Knowledge Representations in the Human Brain', *Neuron*, 107(2), pp. 383–393.e5.

Warren, R. M. (1970) 'Perceptual restoration of missing speech sounds', *Science*, 167(3917), pp. 392–393.

Waters, G. S. and Caplan, D. (2003) 'The reliability and stability of verbal working memory measures', *Behavior Research Methods, Instruments, and Computers*.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A. and Mitchell, T. M. (2014) 'Simultaneously uncovering the patterns of brain regions involved in different story reading Subprocesses', *PLoS ONE*, 9(11), pp. 1–19.

Wehbe, L., Vaswani, A., Knight, K. and Mitchell, T. M. (2014) 'Aligning context-based statistical models of language with brain activity during reading', *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 118(2), pp. 237–254.

Weissbart, H., Kandylaki, K. D. and Reichenbach, T. (2019) 'Cortical Tracking of Surprisal during Continuous Speech Comprehension', pp. 1–12.

Wernicke, C., Cohen, R. and Wartofsky, M. (1874) 'Boston studies in the philosophy of science'.

- White, E. B. (1951) *Charlotte's Web*. Harper and Borthers.
- Wicha, Nicole Y Y, Bates, E. A., Moreno, E. M. and Kutas, M. (2003) 'Potato not Pope: human brain potentials to gender expectation and agreement in Spanish spoken sentences.', *Neuroscience letters*, 346(3), pp. 165–8.
- Wicha, Nicole Y.Y., Moreno, E. M. and 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*, 39(3), pp. 483–508.
- Wicha, N. Y. Y., Moreno, E. M. and 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), pp. 1272–1288.
- Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P. and Van Den Bosch, A. (2016) 'Prediction during Natural Language Comprehension', *Cerebral Cortex*, 26(6), pp. 2506–2516.
- Wingfield, A. and Grossman, M. (2006) 'Language and the Aging Brain: Patterns of Neural Compensation Revealed by Functional Brain Imaging', *Journal of Neurophysiology*, 96(6), pp. 2830–2839.
- Wlotko, E. W. and Federmeier, K. D. (2012) 'Age-related changes in the impact of contextual strength on multiple aspects of sentence comprehension', *Psychophysiology*, 49(6), pp. 770–785.
- Wlotko, E. W., Federmeier, K. D. and Kutas, M. (2012) 'To predict or not to predict: Age-related differences in the use of sentential context', *Psychology and Aging*, 27(4), pp. 975–988.
- Wlotko, E. W., Lee, C.-L. and Federmeier, K. D. (2010) 'Language of the Aging Brain: Event-Related Potential Studies of Comprehension in Older Adults', *Language and Linguistics Compass*, 4(8), pp. 623–638.
- Wong, D., Fuglsang, S., Hjortkjær, J., Ceolini, E., Slaney, M. and Cheveigné, A. de (2018) 'A Comparison of Temporal Response Function Estimation Methods for Auditory Attention Decoding', p. 281345.
- Woods, D. L., Hillyard, S. A. and Hansen, J. C. (1984) 'Event-related brain potentials reveal similar attentional mechanisms during selective listening and shadowing', *Journal of Experimental Psychology: Human Perception and Performance*.
- Wu, M. C.-K., David, S. V. and Gallant, J. L. (2006) 'Complete Functional Characterization of Sensory Neurons By System Identification', *Annual Review of Neuroscience*, 29(1), pp. 477–505.
- Yan, S., Kuperberg, G. R. and Jaeger, T. F. (2017) 'Prediction (or not) during language processing. A commentary on Nieuwland et al. (2017) and DeLong et al. (2005) Shaorong', (2005), pp. 1–14.
- Yi, H. G., Leonard, M. K. and Chang, E. F. (2019) 'The Encoding of Speech Sounds in the Superior Temporal Gyrus', *Neuron*, 102(6), pp. 1096–1110.
- Zan, P., Presacco, A., Anderson, S. and Simon, J. Z. (2019) 'Exaggerated Cortical Representation of Speech in Older Listeners : Mutual Information Analysis', *biorxiv*.
- Zatorre, R. J., Evans, A. C., Meyer, E. and Gjedde, A. (1992) 'Lateralization of Phonetic and Pitch Discrimination in Speech', *Science (New York, N.Y.)*, 256(5058), pp. 846–849.
- Zion Golumbic, E. M. *et al.* (2013) 'Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”', *Neuron*, 77(5), pp. 980–991.

Zou, J. *et al.* (2019) ‘Auditory and language contributions to neural encoding of speech features in noisy environments’, *NeuroImage*.

Appendix I. Comparing Embedding Models for Neural Correlates of Semantic Dissimilarity

In chapters 3 and 6, semantic dissimilarity was estimated using word2vec (Mikolov et al., 2013a). However, this estimation method was developed throughout the course of our research by using alternative word vector representations from different models, such as GloVe (Pennington et al., 2014). In a short analysis, we wished to test which embedding model would better capture the semantic processing of words in context. Semantic dissimilarity vectors were estimated using word2vec and GloVe and were regressed simultaneously to the hour-long natural speech EEG dataset (presented in chapter 3). We used the same method described in chapter 5 (see section 5.2 for details), however, instead of comparing lexical surprisal and semantic dissimilarity vectors, we compared GloVe dissimilarity vectors with word2vec dissimilarity vectors. Estimated dissimilarity values were strongly correlated across the different models (Pearson's $r = 0.75$ $p < 0.05$, $n = 4542$). Hence, we expected that one model would provide a good fit to the data and produce reliable EEG prediction accuracies, whereas the other model would perform poorly.

Figure IA shows the trained TRF weights at Pz, averaged across subjects, for word2vec and GloVe dissimilarity features. TRF weights for the GloVe dissimilarity feature showed a prominent N400 component whereas weights for the word2vec dissimilarity feature were weaker. Figure IB shows the prediction accuracy of each feature above a null baseline when controlling for the effect of the other feature (see 5.2 for details). EEG prediction accuracy for the GloVe dissimilarity feature was significantly greater than the null baseline ($t_{18} = 6.4$, $p = 5.4 \times 10^{-6}$, one-sample t-test). EEG prediction accuracy for the word2vec dissimilarity feature was not significantly greater than the null-baseline ($t_{18} = 1.6$, $p = 0.128$, one-sample t-test). EEG prediction accuracy was significantly greater for GloVe compared to word2vec ($t_{18} = 3.75$, $p = 0.0014$, paired t-test).

Although the word2vec dissimilarity model performs well on its own, as indicated in chapter 3, when controlling for the effect of semantic dissimilarity estimated from GloVe vectors, it fails sufficiently explain additional variance relating to subjects' neural responses. This result is unsurprising, as GloVe has outperformed word2vec when tested in word analogy tasks in NLP (Pennington et al., 2014). These results prompted us to use GloVe vectors over word2vec vectors when estimating semantic dissimilarity in chapters 4 and 5 of the thesis.

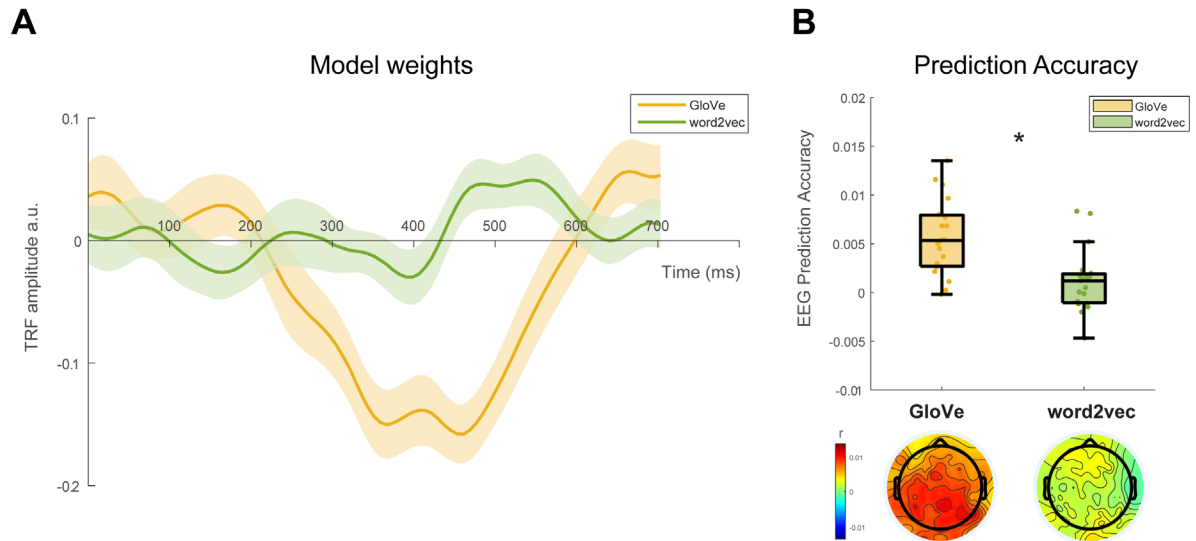


Figure I. 1 TRF weights and EEG prediction accuracies for semantic dissimilarity derived using GloVe and word2vec

(A) GloVe (yellow) and word2vec (green) semantic dissimilarity TRF weights at Pz, averaged across subjects. Shaded areas show s.e.m. across subjects. TRF weights for the GloVe dissimilarity feature showed a prominent N400 component, whereas weights for word2vec dissimilarity showed no such component. **(B)** EEG prediction accuracies for both models. Boxplot (above) of subject prediction accuracy at Pz. Dots indicate individual subjects. Topographical distribution (below) of EEG prediction accuracy, averaged across subjects, at each electrode.