UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Language Representation in Human Cerebral Cortex

Permalink
https://escholarship.org/uc/item/2mj111cm

Author
Gong, Xue

Publication Date
2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Language Representation in Human Cerebral Cortex

by

Xue Lily Gong

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Neuroscience

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Frédéric E. Theunissen, Chair
Professor Jack L. Gallant
Professor Keith A. Johnson
Professor Michael Deweese

Spring 2024

Language Representation in Human Cerebral Cortex

Copyright 2024
by
Xue Lily Gong

Abstract

Language Representation in Human Cerebral Cortex

by

Xue Lily Gong

Doctor of Philosophy in Neuroscience

University of California, Berkeley

Professor Frédéric E. Theunissen, Chair

To comprehend language, the human brain transforms sound pressure waveforms (speech) and visual patterns (reading) into meaningful words through a set of dynamic and intermediate representations. This capability is particularly evident in bilinguals who can extract the same meaning from different spoken and written languages despite the distinctions in acoustic and visual inputs. However, our understanding of these complex brain functions remains limited. This dissertation presents three functional magnetic resonance imaging (fMRI) experiments that explore how semantic information is extracted from acoustic and visual patterns in the cerebral cortex of monolingual English speakers and Mandarin-English bilingual speakers. Using naturalistic stimuli and voxelwise encoding models, these experiments investigate the cerebral cortical representation of visual, acoustic, articulatory, phonemic, orthographic, semantic, and syntactic components of language. The first experiment (Chapter 2) reveals that human brains segment continuous speech into diphones to facilitate semantic comprehension, identifying two putative brain regions where phonemic processing transitions to semantic representation. The second experiment (Chapter 3) maps the cerebral cortical representation of visual, orthographic, and semantic aspects during reading, demonstrating that language processing involves simultaneous interaction across multiple information levels, rather than a series of independent steps. The third experiment (Chapter 4) compares these processes between English and Mandarin, showing while visual and semantic representations are similar across languages, orthographic processing varies significantly.

Contents

Contents **i**

List of Figures **iii**

List of Tables **vi**

1 Introduction **1**

2 Phonemic segmentation of narrative speech in human cerebral cortex **6**
2.1 Abstract 6
2.2 Introduction 6
2.3 Results 8
2.4 Discussion 21
2.5 Materials and Methods 26

3 Cerebral cortical representation of natural language processing in English **37**
3.1 Abstract 37
3.2 Introduction 37
3.3 Results 39
3.4 Discussion 49
3.5 Materials and Methods 52

4 Cerebral cortical representation of natural language processing in Mandarin-English bilinguals **60**
4.1 Abstract 60
4.2 Introduction 60
4.3 Results 62
4.4 Discussion 73
4.5 Materials and Methods 76

5 Conclusion **84**

Bibliography **86**

A Supplements for "Phonemic segmentation" 97
A.1 Supplementary Figures 98
A.2 Supplementary Tables 115
A.3 Supplementary Methods 122
A.4 Supplementary Notes 123

B Supplements for "Representation of English reading" 125
B.1 Supplementary Figures 126

C Supplements for "Representation of Bilingual reading" 146
C.1 Supplementary Figures 147

List of Figures

2.1 Feature spaces used for voxelwise modeling. 9
2.2 Auditory processing: prediction performance of the acoustic baseline voxelwise model (VM). 11
2.3 Variance partitions for phonemic processing. 13
2.4 Phonemic processing and segmentation. 14
2.5 Contribution of distinct diphone categories to phonemic VM. 17
2.6 Phonemic to semantic processing. 18
2.7 Contribution to VM performance of phonemic vs semantic features in transition areas within the lateral temporal cortex (LTC) and inferior prefrontal cortex (IPFC). 20

3.1 Experimental procedure and voxelwise modeling (VM) 40
3.2 Joint model prediction accuracy across the cerebral cortical surface. 42
3.3 Variance partitioning. 44
3.4 Visual feature tuning. 47
3.5 Orthographic feature tuning. 48

4.1 Participant language demographic information and experimental procedure. . . 63
4.2 Explainable variance and joint model prediction accuracy of English versus Mandarin. 65
4.3 Variance partitioning. 68
4.4 Across-language prediction accuracy. 70
4.5 Visual feature tuning. 72

A.1 Voxelwise Modeling. 99
A.2 Variance Partitioning. 100
A.3 Validation by simulations: effect of voxel sensitivity. 101
A.4 Validation by simulation: sensitivity to longer combinations of phonemes. 102
A.5 Prediction performance of auditory processing on flat maps for the remaining 10 subjects. 103
A.6 Prediction performance of phonemic processing on flat maps for the remaining 10 subjects. 104
A.7 Phonemic processing: hemispheric analysis. 105

A.8 Prediction performance of distinct diphone categories for 11 subjects. 106
A.9 Modeling on diphoneme statistical properties. 107
A.10 Modeling on vowel and consonants. 108
A.11 Prediction performance of phonemic-semantic cortical maps for the remaining 10 subjects. 109
A.12 Semantic vs phonemic contributions: hemispheric analysis. 110
A.13 Flatmap of individual subject for prediction performance of phonemic versus semantic based VM. 111
A.14 Anatomical location of the center of mass for individual subjects. 112
A.15 Prediction Performance from five sessions of data 113
A.16 Signal-to-noise ratio of BOLD signal for each subject. 114

B.1 Joint prediction performance for individual participants. 127
B.2 Split prediction performance of visual features for individual participants. . . . 128
B.3 Split prediction performance of orthographic features for individual participants. 129
B.4 Split prediction performance of semantic features for individual participants. . 130
B.5 Split prediction performance of statistics features for individual participants. . 131
B.6 Split prediction performance of phonemes features for individual participants. . 132
B.7 Split prediction performance of syntax features for individual participants. . . . 133
B.8 Prediction accuracy is shown for visual (orange), orthographic (green) and semantic (blue) feature spaces on the flattened cortical surface of individual participants. 134
B.9 Unique variance of visual partition for individual participants. 135
B.10 Unique variance of orthographic partition for individual participants. 136
B.11 Unique variance of semantic partition for individual participants. 137
B.12 Unique variance of visual-orthographic partition for individual participants. . . . 138
B.13 Unique variance of orthographic-semantic partition for individual participants. . 139
B.14 Unique variance of visual-semantic partition for individual participants. 140
B.15 Unique variance of visual-orthographic-semantic partition for individual participants. 141
B.16 Visual features direction tuning for individual participants. 142
B.17 Visual features position tuning for individual participants. 143
B.18 Visual features spatial frequency tuning for individual participants. 144
B.19 Visual features temporal frequency tuning for individual participants. 145

C.1 Explainable variance of English versus Mandarin. 148
C.2 Joint prediction performance for individual participants per language per language. 149
C.3 Split prediction performance of visual features for individual participants per language. 150
C.4 Split prediction performance of orthographic features for individual participants per language. 151
C.5 Split prediction performance of semantic features for individual participants per language. 152

C.6 Prediction accuracy is shown for visual (red), orthographic (green) and semantic (blue) feature spaces on the flattened cortical surface of individual participants for English (left) and Mandarin (right). 153
C.7 Grouped-averaged prediction accuracy for the all seven partitions in English and Mandarin are shown on the flattened cortical surface of the standard fsAverage brain with a 2D colormap. 154
C.8 Across-language prediction accuracy of visual features for individual participants per language. 155
C.9 Across-language prediction accuracy of orthographic features for individual participants per language. 156
C.10 Across-language prediction accuracy of semantic features for individual participants per language. 157
C.11 Visual features direction tuning for individual participants per language. 158
C.12 Visual features spatial frequency tuning for individual participants per language. 159
C.13 Visual features temporal frequency tuning for individual participants per language. 160

List of Tables

A.1 Table of prediction performance of phonemic processing in each ROI for each subject and statistics based on subject averages. 116
A.2 Table of voxel counts and corresponding statistics of phonemic processing in each ROI for each subject based on subject averages with logistic regression. 117
A.3 Table of prediction performance of each diphone category for each subject. . . . 118
A.4 Table of prediction performance performance of Phonemic-Semantic processing in each ROI for each subject and statistics based on subject averages. 119
A.5 Table of voxel counts and corresponding statistics of Phonemic-Semantic processing in each ROI for each subject based on subject averages with logistic regression. 120
A.6 Table of definition of diphone phontactic probability features 121

Acknowledgments

As I reflect on the journey of my graduate studies, I am overwhelmed with gratitude for the myriad of people who supported me through this challenging yet rewarding phase. The completion of my thesis would not have been possible without their unwavering support, and it is with a deep sense of appreciation that I acknowledge their contributions.

First and foremost, I extend my heartfelt thanks to my advisor, Frédéric Theunissen. His guidance and insights have been pivotal to my research. He is both a mentor and role model whose qualities I aspire to emulate. I am also grateful to the members of Theunissen lab for their collaborative spirit and invaluable contributions: Cong Du, her diligent, meticulous work that, though time-consuming and often behind the scenes, has been crucial to this these; Kevin Yu, whose expertise spans programming, visual design, and a consistent readiness to assist, improved my work to another level; and both Bill Wood and Logan Thomas for taking the time to listen to all my questions, concerns, and uncertainties, and for offering their sincere and helpful advice.

My co-advisor, Jack Gallant has been equally instrumental in my academic journey. Despite his stern exterior, his continuous encouragement and caring support have been invaluable, motivating me throughout my studies. The close collaboration with the bilingual teams, including Catherine Chen, Fatma Deniz, and Christine Tseng, has been crucial to this thesis work. I am grateful to Cathy for her expertise, Fatma for her mentoring, and Christine for their scientific insights as well as the enjoyable exchanges about our shared hobbies. Special thanks are due to Sara Popham for her mentorship from my first year to this day, guiding me through both academic and personal growth. I am also thankful to Tom du Pré la Tour for his pattern-like coding which I followed eagerly, and the creation of the Himalaya and pymoten packages, which has been a cornerstone of this thesis work. Thanks to Matteo Visconti di Oleggio Castello, an expert in neuroimaging, for teaching me all the intricacies of fMRI and always approachable for questions. Further appreciation goes to Michele Winter, who has provided me with tremendous support academically, personally, and emotionally. Sharing the highs and lows of graduate school with her has transformed miserable moments into cherished diamonds that will shine brightly in my life. My other lab members, Tianjiao Zhang, Emily Meschke, Jen Holmberg, Amanda LeBel, Cheol Jun Cho, and Alicia Zeng, have also been pillars of support and camaraderie throughout my graduate studies.

I am grateful to my qualifying exam committee, including Keith Johnson, who cultivated my knowledge in phonetics and linguistics, and Michael DeWeese, for his invaluable feedback.

My collaborators in Alex Huth's lab, Deniz Lab, Julie Elie, and Will Liberti, have contributed to my research with their unique perspectives and expertise, for which I am profoundly thankful.

A special thanks to the HWNI faculty and administrative staff, especially Canada Groskreutz and Lelena Avila, who were a beacon of light for graduate students like myself lost in the darkness of bureaucracy.

Lastly, I am thankful for my pre-graduate advisors—Rolf Nelson, Jeremy Wolfe, Randall Davis, and Dana Penney. Their guidance helped pave my path to Berkeley, setting the foundation for my graduate studies.

To all mentioned and others who contributed in various ways, your support has left an indelible mark on both my professional development and personal growth. Thank you for being part of my journey.

Chapter 1

Introduction

Language is a fundamental aspect of human interaction and communication, requiring the transformation of acoustic sound pressure waves and visual patterns into meaning through dynamic intermediate representations. This capability is markedly pronounced in bilinguals who can effortlessly extract the same semantic content from diverse acoustic and visual cues of different languages. However, the underlying neural mechanisms of these processes are not fully understood. This dissertation explores the cerebral cortical processes involved in extracting semantic information from both acoustic and visual inputs in monolingual English speakers and Mandarin-English bilingual speakers. Employing functional magnetic resonance imaging (fMRI) and voxelwise encoding models, this work investigates how the brain represents various components of language, including visual, acoustic, articulatory, phonemic, orthographic, semantic, and syntactic features.

From Sound to Meaning

The cerebral cortex is responsible for transforming continuous speech sound waves into coherent semantic meaning. This process begins with the identification of spectro-temporal features in the primary auditory cortex (AC, Hullett et al., 2016) and secondary auditory areas in dorsal superior temporal gyrus (dSTG, Venezia et al., 2019. These features are then combined into phonemic units in the mid-superior temporal sulcus (mSTS, de Heer et al., 2017; Mesgarani et al., 2014), and subsequently organized into syllables and words in the middle temporal gyrus/inferior temporal sulcus (pMTG, pITS, B. T. Carter et al., 2019; Humphries et al., 2005). This ventral pathway exhibits a dorsal-to-ventral gradient with an increasing semantic complexity (Hagoort & Indefrey, 2014), and interfaces with the broadly distributed conceptual network (Binder et al., 2009; Huth et al., 2016).

However, this functional localization provides limited insights into how "higher-level" representations (e.g. semantic) emerge from "lower-level" representation (phonemic), and vice-versa how "high-level" responses influence "low-level" responses. To bridge this gap, we propose a segmentation and identification (S/I) chain model. This model suggests that

the human brain initially segments continuous acoustic sound waves into dynamic, coarse segments using a speech optimized modulation filter bank. These filters are hypothesized to be located in the posterior superior temporal gyrus (pSTG), a region recently identified as a parallel primary-like auditory region specialized for speech (Hamilton et al., 2021). The model then refines these segments into finer spectro-temporal segments based on detecting rapid changes in the identified spectro-temporal features. These finer temporal segments, or acoustic chunks, are obtained without any linguistic input, corresponding to the acoustic units processed by young children prior to language acquisition (Räsänen et al., 2018). However, in fluent speakers, these acoustic chunks are encoded as specific phonemic units in the identification step. The process continues with the segmentation and identification of syllables and words, increasingly relying on top-down linguistic input. Syllables are segmented from a stream of phonemic units using speech amplitude envelope information and language-dependent phonotactic rules, and identified as individual units. Words are subsequently segmented from these syllable streams based on stress patterns, additional phonotactic rules for word boundaries, and "veto" information from the lexicon that eliminates nonsensical syllable combinations. This process integrates both acoustic and linguistic inputs, accomplishing the word recognition though acquired lexical knowledge.

To empirically test this hypothesis, we first investigated phonemic segmentation, which is the first stage in this model requiring top-down for phonemic awareness. In natural speech, single phonemes don't occur in isolation and are known to exhibit coarticulation effects, where their acoustic properties change even in simple combinations (Öhman, 1966). We use the term "phonemic unit" rather than "phoneme" to reflect ongoing discussions in phonology about the nature of these perceptual units in prelexical speech processing. Understanding these phonemic units from a neural perspective is crucial for unraveling the sound-to-phoneme processing steps in the human brain(Goldinger & Azuma, 2003), as explored in the experiments detailed in Chapter 2.

From Graph to Meaning

Apart from speech processing, the cerebral cortex is also responsible for the unique capability of human beings to transform written texts into semantic meaning. This process requires recruiting the existing cortical ventral pathway of object recognition (Dehaene & Cohen, 2011). This pathway begins in the primary visual cortex where fundamental features, such as edges, motions, and colors, are extracts (Hubel & Wiesel, 1962). These features are then combined into more complex forms like shapes and textures in higher cortical areas including V2, V3, and V4 (reviewed in Bracci and Op de Beeck, 2023). Subsequently, these complex features are integrated in the inferotemporal cortex, which specializes in recognizing complex objects like faces and body parts (reviewed in DiCarlo et al., 2012; Grill-Spector and Weiner, 2014).

However, the specifics of how the brain recruits this ventral pathway of general object recognition for reading comprehension remain less understood. What visual features are

important for written word recognition, is still being explored. Research highlights that the visual word form area (VWFA) in the left ventral occipitotemporal cortex is crucial for the invariant recognition of written words (Cohen et al., 2000). Further complexities arise in how orthographic information is transformed into semantic and syntactic understanding. Previous studies show that recognized words are mapped onto phonological representations in the Wernicke's area (Booth et al., 2002) to be matched with their lexical semantic meaning (Price, 2012) and processed syntactically in Broca's area (Friederici, 2011; Friederici et al., 2017). However, the specific phonological and orthographic representations during reading remain unclear. Chapter 3 explores how visual, orthographic and semantic features are represented across the cerebral cortex.

From Different Graphs to Same Meaning

The human brain is not only capable of transforming acoustic sound pressure waves and visual patterns into coherent meaning, but also capable of extracting the same semantic content from languages with significantly different acoustic and visual stimuli. This remarkable multilingual ability prompts critical questions: Do the same cortical regions process multiple languages? Are the brain responses similar across languages? What specific visual, orthographic, and semantic features are represented in each language?

Despite extensive research, the findings on how bilingual brains accomplish these tasks remain inconclusive, largely due to the variability in experimental designs and participant demographics, such as native language, age of acquisition, and language proficiency (reviewed in Comstock, 2024). For visual processing, research has focused on visual attention span (VAS), which is the ability to process multiple orthographic units simultaneously and is critical for fluent reading (Bosse et al., 2007). However, investigations into the specific visual features essential for comprehending scripts and words across different languages are limited. For orthographic processing, while distinct brain activation patterns in different cortical areas associated with different writing systems have been documented (Shen & Del Tufo, 2022; Zhan et al., 2023), the specific orthographic features encoded in various languages are not well understood. For semantic processing, while some studies show a significant overlap in how semantic information is processed across languages, others point to unique language-specific semantic representations (Sulpizio et al., 2020; Zhang et al., 2023).

To bridge these gaps, Chapter 4 presents an fMRI experiment in which Mandarin-English bilingual participants read engaging natural narrative stories for two hours in each language. With voxelwise encoding models, this study aims to elucidate how visual, orthographic, and semantic information is represented across languages.

Limitations in current data analysis methods

The mixed results often seen in neuroimaging studies of the language representations in the human cerebral cortex may be attributed to several factors. Foremost, the cerebral cortex operates as a hierarchically organized processing system, likely involving complex and mixed representations of different features. Traditionally, neuroimaging studies assume that each brain area is dedicated to one specific computational step or representation level, isolating these steps from the broader language processing system. This method provides limited insight into the relationships between representations at different levels in the language processing hierarchy.

Another critical limitation is the use of unnatural, isolated stimuli in studying speech and reading comprehension. Neuroimaging experiments typically use isolated or segmented sounds, words or sentences with limited context, which are markedly different from the continuous and dynamic nature of natural language. It is well documented that factors, such as attention and context, significantly impact language comprehension (Mesgarani & Chang, 2012; Peelle et al., 2013; C. Perfetti, 2007). Given the inherent nonlinearities in language processing, the extent to which traditional neuroimaging studies accurately explain brain responses during natural language comprehension remains questionable. Additionally, most studies focus on languages in isolation, failing to consider the dynamic interplay and integration of multiple languages in multilingual individuals (Chee et al., 2004).

To address these limitations, our study employs large-scale, data-driven methods to analyze bilingual brain representations using naturalistic stimuli. This approach may provide a more comprehensive and systematic view of the cerebral cortical representation of different languages across different levels of information, by incorporating naturalistic language contexts and considering the multilingual experience.

Chapter Outline

Chapter 2, entitled *Phonemic segmentation of narrative speech in human cerebral cortex* was previously published as (Gong et al., 2023) and included here with modifications. It describes the results from an fMRI experiment in which eleven human participants listen to two hours of engaging autobiographical stories. It reveals that human brains segment continuous speech into diphones to facilitate semantic comprehension. This chapter identifies two putative brain regions where the transitions from phonemic processing to semantic meaning representation occurs.

Chapter 3, entitled *Cerebral cortical representation of reading in English*, presents the results from an fMRI experiment in which nine human participants read two hours of engaging natural narrative stories. It details a cerebral cortical map highlighting the visual, orthographic and semantic representation of English reading comprehension. The findings suggest that the cerebral cortical representation of reading is not a linear process but involves com-

plex, simultaneous processing of multiple information levels, challenging traditional views that focus on isolated feature processing.

Chapter 4, entitled *Cerebral cortical representation of reading in Mandarin-English bilinguals*, describes the results from an fMRI experiment where six Mandarin-English bilingual participants read engaging narrative stories for two hours in each language. It compares the cerebral cortical representations of reading between English and Mandarin in bilingual brains. The results show that similar cerebral cortical regions are involved in reading comprehension across languages, indicating a shared language processing network. However, while the visual and semantic information are represented in similar cortical regions, orthographic information is represented in different cortical areas with different features encoded across languages. Additionally, the visual properties critical for reading also differ across languages. These differences highlight adaptive capabilities of the bilingual brains in different languages.

Chapter 2

Phonemic segmentation of narrative speech in human cerebral cortex

This chapter is reprinted from *Gong, X. L., Huth, A. G., Deniz, F., Johnson, K., Gallant, J. L., & Theunissen, F. E. (2023). Phonemic segmentation of narrative speech in human cerebral cortex. Nature communications, 14(1), 4309* with permission.

2.1 Abstract

Speech processing requires extracting meaning from acoustic patterns using a set of intermediate representations based on a dynamic segmentation of the speech stream. Using whole brain mapping obtained in fMRI, we investigate the locus of cortical phonemic processing not only for single phonemes but also for short combinations made of diphones and triphones. We find that phonemic processing areas are much larger than previously described: they include not only the classical areas in the dorsal superior temporal gyrus but also a larger region in the lateral temporal cortex where diphone features are best represented. These identified phonemic regions overlap with the lexical retrieval region, but we show that short word retrieval is not sufficient to explain the observed responses to diphones. Behavioral studies have shown that phonemic processing and lexical retrieval are intertwined. Here, we also have identified candidate regions within the speech cortical network where this joint processing occurs.

2.2 Introduction

The human cerebral cortex performs complex computations to transform a continuous speech sound pressure waveform into a language based message. These transformations include the detection of information-bearing spectro-temporal features in the speech sound, the

combination of these acoustic features into phonemic units, syllables, words and meaningful clauses. At a coarse level, the location of cortical regions underlying the transformation from sound to words has been well described (Bornkessel-Schlesewsky et al., 2015; de Heer et al., 2017; DeWitt & Rauschecker, 2012). However, the neural basis of the phonemic processing steps where phonemic subunits are combined into meaningful larger units for the identification of syllables and words remains poorly understood. In order to explore phonemic processing in the cerebral cortex, we measured human fMRI blood-oxygen level-dependent (BOLD) response to narrative stories (Deniz et al., 2019; Huth et al., 2016). First, we contrasted the predictive power of linearized encoding models based on acoustic features, phonemic features and semantic features in order to delineate the cortical regions that were sensitive to phonemic information. Second, to explore the nature of phonemic segmentation within these identified phonemic cortical areas, we then contrasted the predictive power of nested phonemic encoding models that used single phonemes, diphones and triphones as regressors.

Our approach implicitly combines two complementary analyses that have been used successfully in prior neurolinguistic research. The hierarchical processing in the sound to meaning transformation of the speech signal requires both abstraction at different levels and segmentation at different time scales. On the one hand, research based on fMRI experiments has principally leveraged differences in responses in predictive power for different levels of abstraction. For example, by using words versus non-word matched speech sounds as stimuli, researchers have analyzed BOLD responses to delineate cortical word-specific regions from phonemic regions within the temporal lobe (Al-Zubaidi et al., 2022; Boets et al., 2013; de Heer et al., 2017; Kell et al., 2018). Similarly, the predictive power of statistical linear models using spectral features versus articulatory features has been used to distinguish primary auditory cortical regions from speech specific cortical regions within the superior temporal gyrus (de Heer et al., 2017; Kell et al., 2018).

On the other hand, research based on EEG, MEG and ECoG experiments have principally leveraged the entrainment observed at different temporal scales between the speech waveform and the neural signals measured with those techniques (Brodbeck & Simon, 2020; N. Ding et al., 2016; Gross et al., 2013). For example, neural correlates of word segmentation have been localized to cortical regions found along the left inferior and middle frontal gyri (McNealy et al., 2006). At lower frequencies (< 2 Hz), it has also been shown that the time scales of phrase and sentence processing can be reflected in the neural activity found in middle and posterior superior temporal gyrus and inferior frontal gyrus (N. Ding et al., 2016). At higher frequencies, corresponding to the syllable rate, the robust phenomenon of cortical entrainment to the speech envelope has been linked to lower level processing of speech sounds. It is well accepted, however, that the functional properties of this cortical speech entrainment as measured in EEG needs to be further analyzed in order to distinguish its role in speech or linguistic processing from a more general role in acoustic processing (N. Ding & Simon, 2014).

Here we assessed the predictive power of linear models based on phonemic units combined at different levels of *temporal* granularity and contrasted those models to those based on lower

level (spectral) and higher level (word embeddings) of *abstraction*. This combined approach allowed us to clearly delineate phonemic cortical regions from acoustic or word/semantic regions and simultaneously examine the granularity of the segmentation for phonemic units. Although it appears counter-intuitive to be able to detect the fast rates of phonemic segmentation in the slow fMRI BOLD signal, we will show that this analysis is possible as long as individual voxels show sufficient sensitivity for the identity of specific phonetic units.

Lastly, the locus of the transition from phonemic processing to lexical retrieval has also been debated as both anterior and lateral posterior regions of the temporal lobe as well as inferior prefrontal cortex (reviewed in Poeppel et al., 2012). The investigation of the temporal granularity of phonemes in combination with the representation of speech processing at different levels of abstractions allowed us to more rigorously assess the unique phonemic cortical region and thus revisit the locus of the phonemic to lexical retrieval transition. For this purpose, a detailed quantitative analysis of the differences in predictive power between linearized models based on phonemic features versus semantic features was performed.

2.3 Results

The goals of this study were, first, to localize the phonemic brain regions; second, to investigate the nature of the phonemic segmentation; and third, to determine the putative brain regions where the transformation from phonemic processing to lexical semantic meaning representation occurs.

To achieve these goals, we collected fMRI BOLD data while 11 participants listened to more than two hours of spoken narrative stories from *The Moth Radio Hour* (de Heer et al., 2017; Deniz et al., 2019; Huth et al., 2016). Following the logic of the voxelwise encoding model (VM) framework, acoustic, phonemic, and semantic features were first extracted from the speech stimuli (Figure 2.1). These features were then used in regularized linear regression (la Tour et al., 2022) to predict the time varying BOLD signals in each voxel independently. The predictive power of these voxelwise models was then estimated by cross-validation, using a separate data set reserved for this purpose (Materials and Methods, Figure A.1 and Figure A.2). By comparing the predictive power obtained using different feature spaces, one can obtain a functional map of phonemic processing.

Responses to acoustic features and presence of speech sounds.

In order to localize the cortical regions sensitive to phonemic processing, we first removed the fraction of the BOLD response that could be explained by the mere presence versus absence of speech sounds. For this purpose, we obtained the BOLD predictions from a joint model, *Acoustic Baseline* VM, consisting of the time-varying power spectrum and phoneme count features (Figure 2.1). The phoneme count feature quantifies the presence versus absence of speech sounds without being linearly sensitive to the sound intensity or the exact spectrum of particular utterances. Within the *Acoustic Baseline* VM, we examined

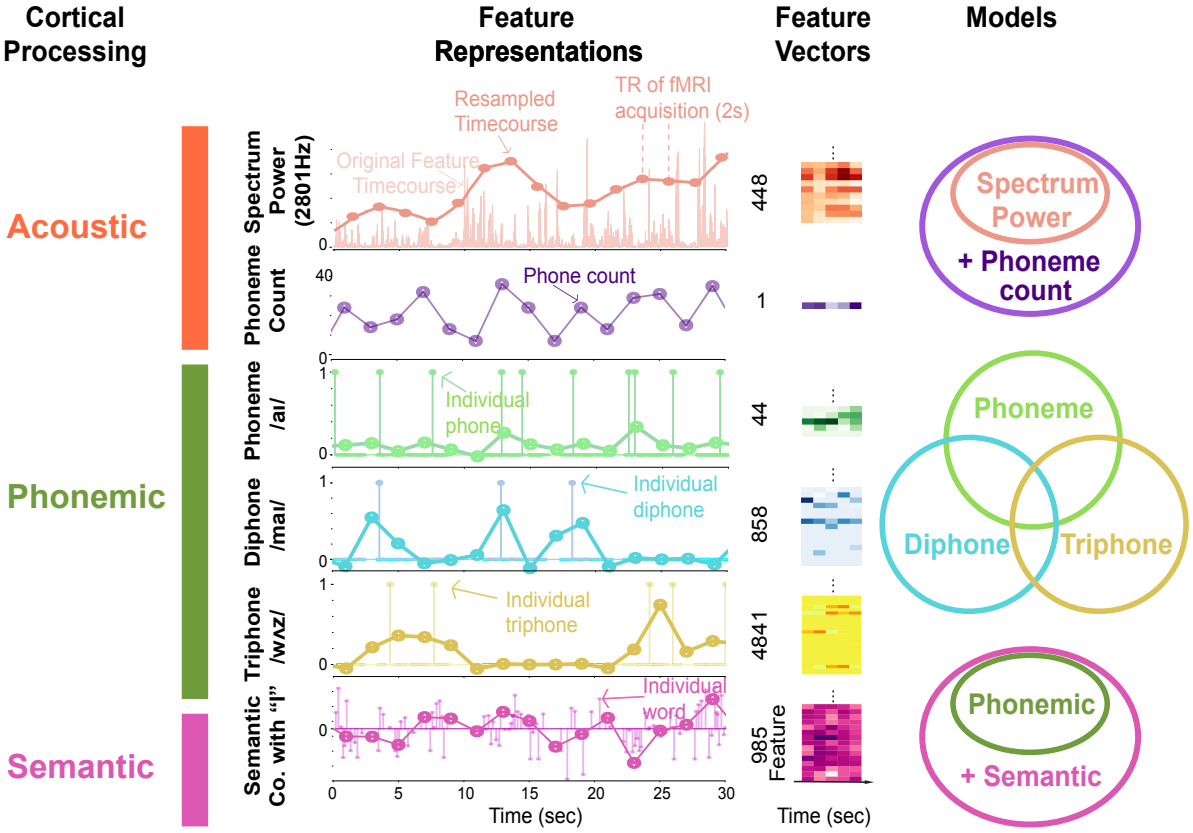


Figure 2.1: **Feature spaces used for voxelwise modeling.** To identify the cortical representation of features important for speech comprehension, the sound waveforms from the stimuli were first transformed into six different feature spaces. Acoustic features were extracted by creating separate spectral power and phoneme count feature spaces. These two acoustic feature spaces describe brain activity that can be explained by the mere presence or absence of speech sounds. Phoneme-related features were extracted by creating separate feature spaces that reflect single phonemes, diphones and triphones. Semantic features were extracted by creating a feature space based on a 985-dimensional word embedding space (Huth et al., 2016). **Cortical Processing:** Six distinct feature spaces were used to investigate the Acoustic, Phonemic and Semantic processing in the speech cortical network. **Feature Representations:** Illustration of the time course of a single feature from each feature space: the spectrum of sound signal in the frequency band centered at 2801Hz, the phoneme (/ai/), the diphone (/m.ai/), the triphone (/w.ah.s/), and the semantic co-occurrence with "I". These signals are low-pass filtered to generate the continuous values discretized at the TR (bold colored lines). **Feature Vectors:** Illustration of matrices of each feature space for 5 TRs and a subset of the features in each feature space. **Models:** Venn diagrams illustrating the features used in nested voxelwise models (VMs). The baseline model used sound power spectrum and the phoneme count as features. After subtracting the predictions from the Baseline VM, nested models using phonemic and semantic features (green and pink circles at bottom) were fitted to localize phonemic regions and phonemic to semantic cortical boundaries. Nested models using single phonemes, diphones and triphones were fitted to assess the granularity of the phonemic segmentation in phonemic regions.

the anatomical location of voxels sensitive to time-varying power spectrum, and phoneme count individually (Figure 2.2; see Figure A.5 for per subject data). The time-varying spectrum best predicts the BOLD response of voxels located in the most medial part of the auditory cortex (AC; identified functionally using localizers as described in the Materials and Methods). This medial region surrounds the Heschl's gyrus (cyan line in Figure 2.2) where the primary auditory cortex (PAC) is found. Phoneme count best predicts the BOLD response of voxels in the more lateral regions of AC including voxels along the superior temporal gyrus (STG). Significant predictions are also observed in the more lateral ventral speech premotor area (sPMv) and inferior prefrontal cortex (IPFC) including Broca's area, mostly in the left hemisphere of this subject. These phoneme count auditory voxels could simply correspond to general auditory regions that are more sound level invariant than what can be predicted by linear weights on the time varying power spectrum. Alternatively, these phoneme count voxels could also correspond to an area of auditory processing specialized for speech. Since the goals of our study were not to functionally parcelate these auditory regions, we did not perform any additional analyses or experiments to disambiguate these two alternatives. In order to delineate the cortical regions sensitive to phonemic identity and not simply to their presence, we subtracted the predictions from this *Acoustic Baseline* VM from the BOLD response to obtain a BOLD response residual. All subsequent predictions refer to predictions for this residual response, which we will refer to as Y_{res}.

Phonemic voxels and phonemic segmentation.

To investigate phonemic processing, we defined phonemic voxels as the voxels that responded to the identity of speech phonemes and/or phoneme combinations beyond what could be explained by the mere presence or absence of speech sounds. The response to the identity of speech phonemes was assessed by significant predictions of the BOLD responses based on the full *Phonemic* VM. The full *Phonemic* VM consists of single phoneme, diphone and triphone features. Here we will first describe the anatomical location of the phonemic voxels. Second, we examine phonemic segmentation by quantifying the variance in brain activity that could be explained by models based on each of these three phoneme-related features and the joint of these features.

Anatomically, phonemic voxels were found in large regions of the temporal, parietal and prefrontal cortex (Figure 2.3a: Phonemic Processing). In the temporal cortex, phonemic voxels were found in the superior temporal gyrus (STG) and had increasing predictive power in the superior temporal sulcus (STS) and the adjoining lateral area covering a large section of the lateral temporal cortex (LTC). These temporal cortex phonemic areas were found on both hemispheres and also had large anterior to posterior coverage (see Figure A.7 and Laterality of the phonemic representation and segmentation. for a detailed analysis of hemispheric differences). By contrast, predictive power was very low in the primary auditory cortical areas (PAC). The temporal phonemic area also extended more posteriorly into the lateral and posterior parietal cortex (LPC). Phonemic areas were also found bilaterally along the medial parietal cortex (MPC). Finally, the inferior and superior prefrontal cortex (IPFC,

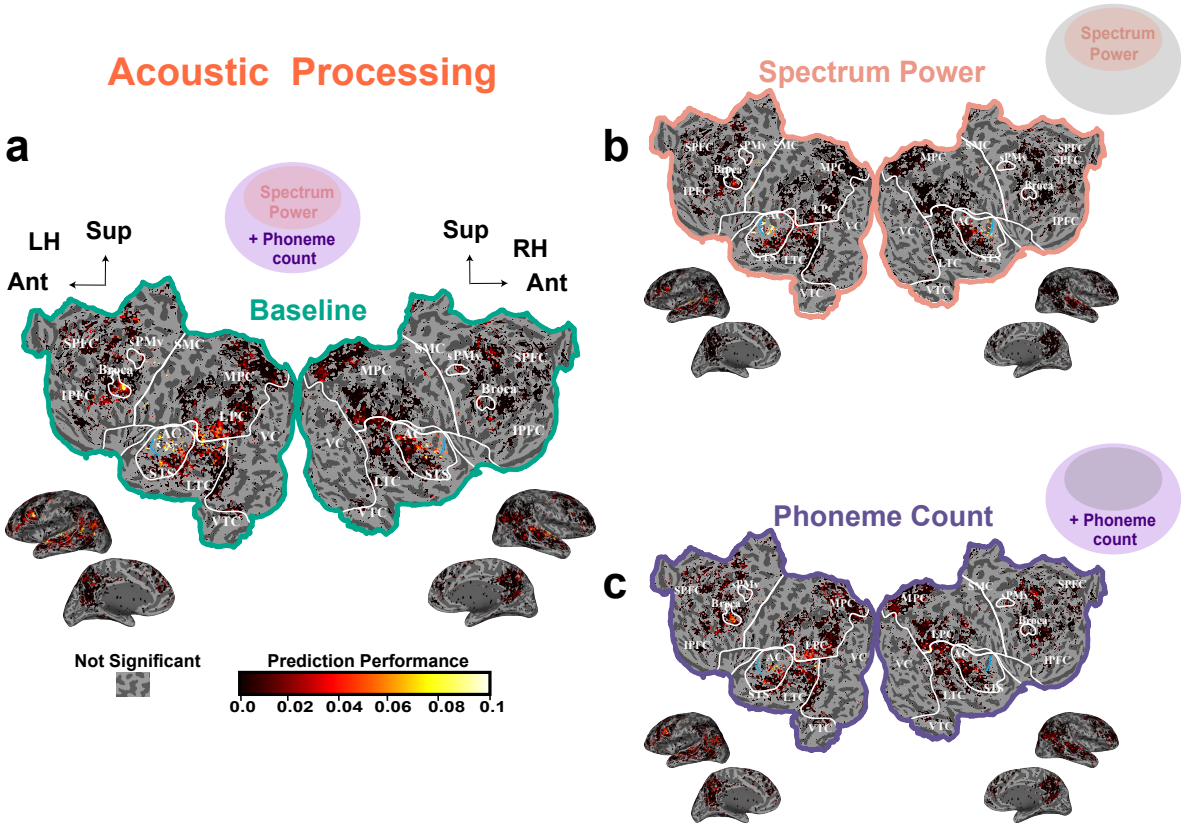


Figure 2.2: **Auditory processing: prediction performance of the acoustic baseline voxelwise model (VM).** To identify acoustic processing across the cerebral cortex, a joint acoustic baseline model consisting of spectrum power and phoneme count feature spaces was constructed. The prediction performance obtained from one typical subject (S5) for the joint acoustic baseline VM, and the models based on the time-varying spectrum feature only, and the phoneme count feature only after excluding the time-varying spectrum are shown. **Panel a** shows that the joint baseline model produces accurate predictions of brain activity in PAC, LTC, sPMv and IPFC. To determine whether these representations were best modeled using spectrum power, or phoneme count features, variance partitioning was used to identify how much of the variance in brain activity could be explained by models based on each of these two features. **Panel b** shows that spectrum power features best explain response variance in PAC. **Panel c** shows that phoneme count features best explain response variance in STG, sPMv and IPFC. Color shows the value of cross-validated prediction performance R^2 for all statistically significant voxels (Permutation test with FDR correction). White lines separate occipital, temporal, parietal and frontal cortex. White continuous lines encircle regions of interest acquired from separate localizer scans. The cyan line within AC locates Heschl's gyrus. Anatomical regions labeled are: SPFC: superior prefrontal cortex, IPFC: inferior prefrontal cortex, MPC: medial parietal cortex, LPC: lateral parietal cortex, STS: superior temporal sulcus, LTC: lateral temporal cortex, MTC: medial temporal cortex, VC: visual cortex; Additional areas defined with localizers: AC: auditory cortex, Broca, sPMv: ventral speech premotor area.

SPFC) contained clusters of phonemic voxels, including in areas just anterior and inferior to Broca's area and anterior to sPMv.

To visualize the unique and joint contribution of single phonemes, diphones and triphones to this cortical phonemic representation, we used variance partitioning. The single phoneme features alone significantly predicted the BOLD response of a small number of voxels located along the superior temporal sulcus (STS) of both hemispheres (Figure 2.3b: Phoneme). By contrast, the diphone features (Figure 2.3b: Diphone) significantly predicted the BOLD response in LTC, LPC, MPC, IPFC and SPFC. The unique contribution of the triphone features, and the joint contribution of each pair of phonemeic features and all three phoneme features are relatively minor and scattered in voxels found in all phonemic cortical regions (Figure 2.3b: Triphone, Phoneme + Diphone, Diphone + Triphone, Phoneme + Triphone, and All). Thus, cortical phonemic segmentation appears to occur principally at the diphone level. Single phoneme representation is limited to more "primary" speech processing areas along the STS. There is no clear distinctive triphone region beyond what was already identified as phonemic processing areas using the full *Phonemic* VM.

To quantify these observations, we estimated the unique contribution of each phonemic feature to the full *Phonemic* VM predictions in the entire cortex, in four ROIs partitioning the large phonemic areas found in the temporal cortex (PAC, STG, STS, LTC), and in Broca's area in the prefrontal cortex.

On average across the entire cortex, the unique contribution of the diphones is superior to the prediction obtained for the unique contribution of single phonemes and to the triphones (mixed-effect statistical models with subject as random factor: likelihood ratio test with null statistical model that ignores phonemic feature, $\chi^2(2) = 10355.60, p < 2.2 \times 10^{-16}$). The effect size is large: the averaged d-prime between mean unique contribution of the diphone feature relative to the unique contribution of single phoneme prediction and the triphone features is $d' = 13.114 + -0.203$ ($2SE$; obtained from mixed effect coefficients; Table A.1 and Table A.2 for per subject data and analyses).

Within the temporal cortex, the prediction performance of the full *Phonemic* VM shows a gradient: it is lowest in primary AC, highest in STG and STS, and lower again in LTC (Figure 2.4). Furthermore, the diphone contribution varied relative to the single phoneme and triphone contributions as a function of anatomical location (Figure 2.4). A linear mixed-effect statistical model analysis was used to predict the unique contribution to the prediction performance with the three phonemic features, the four ROIs in TC and their interaction as fixed effects and subject as random effect. This full statistical model has 12 coefficients (1 for intercept, 3 for ROIs, 2 for features and 6 for the interaction term). Prediction performance varied significantly across ROIs (likelihood ratio test with the nested statistical model that does not include ROIs and the interaction: $\chi^2(9) = 625.68, p < 2.2 \times 10^{-16}$) and across phonemic features (likelihood ratio test with the nested statistical model that does not include feature spaces and the interaction $\chi^2(8) = 4828.82, p < 2.2 \times 10^{-16}$). The unique contribution is higher for the diphone than for the single phoneme and triphone features in STG, STS, and LTC. Moreover, there was a significant interaction between cortical regions and feature space (likelihood ratio test with the nested statistical model that does not include

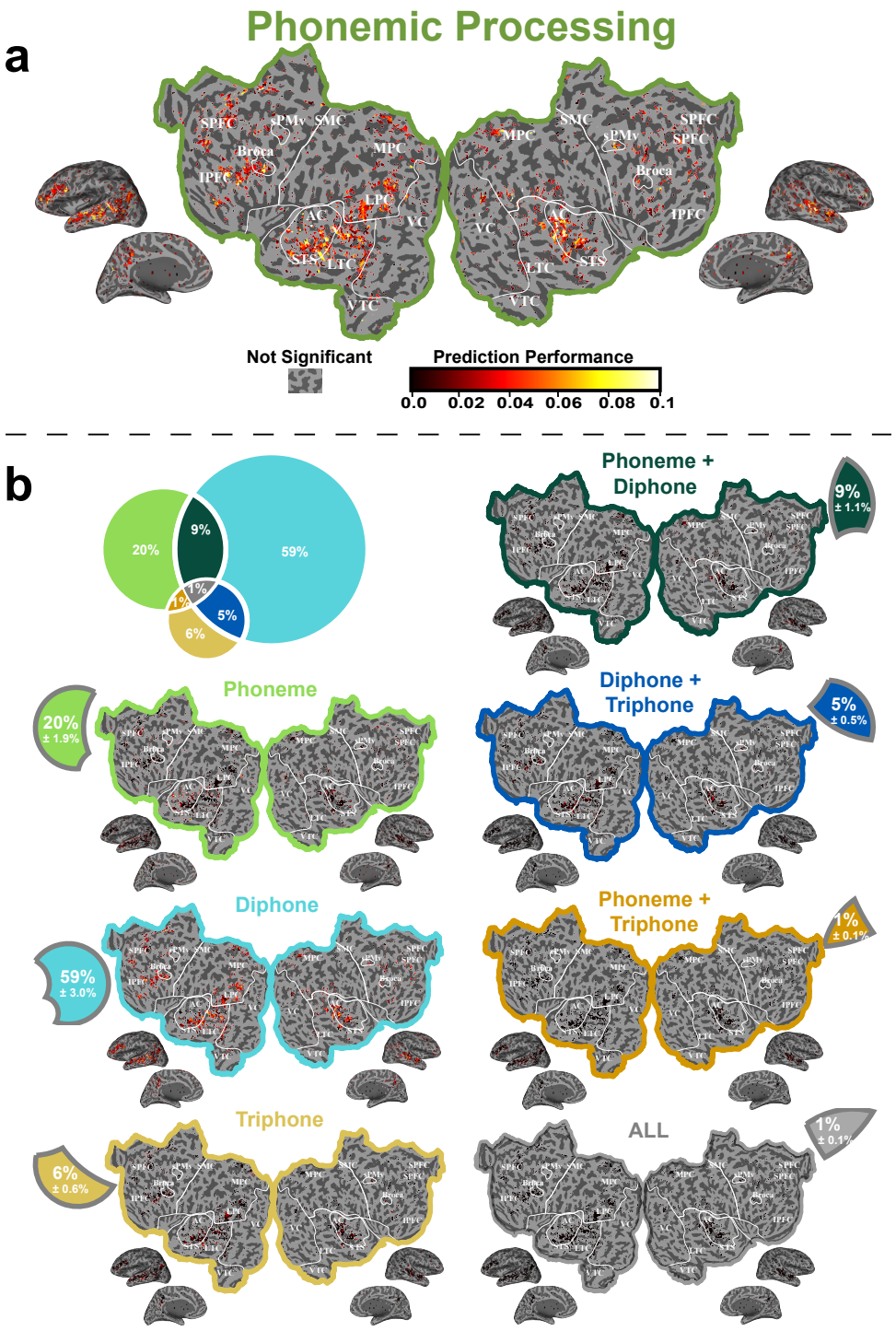


Figure 2.3: **Variance partitions for phonemic processing.** To identify phonemic representations across the cerebral cortex, a joint phonemic model consisting of single phoneme-, diphone- and triphone-based feature spaces was constructed. **Panel a** shows that the joint phonemic model produces accurate predictions of brain activity in LTC, LPC, MPC, IPFC and SPFC. To determine whether these representations were best modeled using single phoneme-, diphone-, or triphone-based features or the joint of these features, variance partitioning was used to identify how much of the variance in brain activity could be explained by models based on each of these three phoneme-related features and their joint pairs. **Panel b** shows that single phoneme features best explain response variance along STS. Diphone features best explain response variance in LTC, LPC, MPC, IPFC and SPFC. Triphone features and the joint of each pair of these phonemic features produce poor predictions in most voxels. Data used to generate this figure has been provided in source data. Cortical regions referred are: SPFC: superior prefrontal cortex, IPFC: inferior prefrontal cortex, MPC: medial parietal cortex, LPC: lateral parietal cortex, STS: superior temporal sulcus, LTC: lateral temporal cortex, MTC: medial temporal cortex, VC: visual cortex, sPMv: ventral speech premotor area.

the interaction $\chi^2(6) = 522.32, p < 2.2 \times 10^{-16}$). This interaction can be described by the effect size used to quantify the differences in the unique contribution between the diphone and the single phoneme and triphone features in each ROI within the temporal cortex. The effect size is small in PAC ($\bar{d'} = 0.440 + -0.968(2SE)$), increases in the STG ($\bar{d'} = 12.838 + -0.838(2SE)$), and in the STS ($\bar{d'} = 12.509 + -0.327(2SE)$), and then decreases in the LTC ($\bar{d'} = 9.060 + -0.471(2SE)$). All $\bar{d'}$ and SE are obtained from mixed effect coefficients. Thus, in all phonemic regions of the temporal cortex, the unique contribution of the diphone features is the highest and the effect size of this difference is larger in STG and STS.

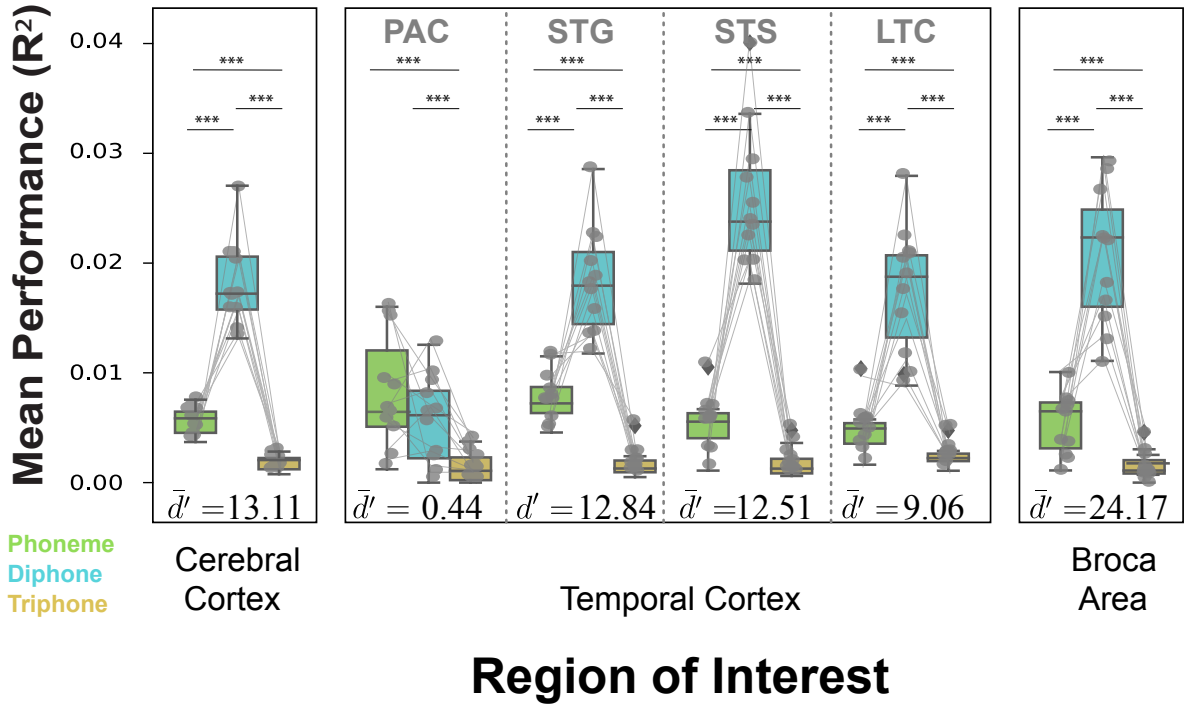


Figure 2.4: **Phonemic processing and segmentation.** To quantify phonemic representations across the cerebral cortex, the average prediction accuracy by ROI (PAC: primary auditory cortex, STG: superior temporal gyrus, STS: superior temporal sulcus, LTC: lateral temporal cortex) for the unique contribution of single phoneme, diphones and triphones was shown in the box plots. It indicates that phonemic segmentation throughout the cortex, and specifically in regions in the temporal cortex and prefrontal cortex with high phonemic representation, occurs at the level of diphones, rather than single phonemes or triphones. The statistics are derived from the performance of significant phonemic voxels for each ROI (n = 201 to 12,411 voxels) across 11 independent subjects (*p< 0.05, **p< 0.01 and ***p< 0.001). The effect size of difference is estimated by the average Cohen's d-prime ($\bar{d'}$) of diphone-single and diphone-triphone comparisons. All tests are two sided and corrected for multiple comparisons. The exact p-values can be found in the Phonemic voxels and phonemic segmentation. section. For box plots, the horizontal lines show the first quartile, median, and the third quartile of the data. Whiskers extend to 1.5 times of interquartile range, and more extreme points are marked as outliers in diamonds. Data used to generate this figure has been provided in source data.

In Broca's area, the diphone features also significantly predict BOLD response better than single phoneme or triphone features (likelihood ratio test with the nested statistical

model that does not include feature spaces, $\chi^2(2) = 443.64, p < 2.2 \times 10^{-16}$; Effect size is $\bar{d'} = 24.168 + -0.203(2SE)$).

In summary, phonemic segmentation throughout the cortex, and specifically in regions in the temporal cortex and prefrontal cortex with high phonemic representation, occurs at the level of diphones, rather than single phonemes or triphones.

Finally, we also investigated to what extent these results could be affected by the fMRI methodology and, in particular, by a potential mismatch between the rate of individual phonemic units and the temporal resolution of the BOLD signal (Figure 2.1). The rate of single phonemes in natural speech stimuli is rapid and varies from 10 to 40 phonemes per TR (i.e. up to 20 Hz). However, the rate of particular single phonemes, diphones and triphones is much slower. Thus, as long as a putative phonemic voxel is sufficiently sensitive for phonemic identity, one should be able to predict its BOLD signal from the presence and absence of a small set of particular phonemes in the voxel's "phonemic receptive field". The requirement of "sufficiently sensitive" depends on the signal to noise ratio of the BOLD signal, and on the length of the signal acquired during the experiment. To determine this sensitivity threshold, we performed a series of simulations where we systematically varied the sensitivity of model voxels (Materials and Methods). Using two hours of narrated speech data, the VM approach used on our data set would distinguish phonemic identity voxels from phonemic count voxels if the model voxel was sensitive for fewer than 10 particular single phonemes and fewer than 100 particular diphones (Figure A.3). The same simulations showed that one could also distinguish voxels that were sensitive to single phonemes from those that were sensitive to diphones or triphones but not to longer combinations (Figure A.4).

Diphone segmentation for identification, expectations or lexical retrieval?

Models capturing the diphone identity yielded, among phonemic features, the highest predictive power in large parts of the cortex implicated in speech processing. Does this phonemic segmentation at the level of the diphone reflect sensitivity for the identity of particular diphones or does it reflect other speech related processing such as expectations for specific diphone sequences or lexical retrieval?

A large body of previous work has shown the importance of predictions in speech processing (Di Liberto et al., 2019; Heilbron et al., 2022; Leonard et al., 2015). In order to further explore if the cortical BOLD activities truly encode the content of diphones or merely the statistical properties of diphones, we built one diphone statistics model consisting of 8 phonological statistical features extracted from Irvine Phonotactic Online Dictionary (IPhoD;Vaden et al., 2009). These features describe the phonotactic probability of diphones (Table A.6). As shown in Figure A.9 and Figure A.10, although transition probabilities across diphones yielded some predictive power in the phonemic regions, the prediction strength of VM models based solely on these transitions probabilities was much smaller than VM models that explicitly represented the diphone identity.

Furthermore, since many diphones are also short words (e.g. diphone "M.AY" could represent the word "my"), it is important to elucidate whether the unique explanatory power from diphone features could be a reflection of responses to high-frequency short words. In addition, the beginning of words are considered to be linguistically important in the process of lexical activation during spoken-word recognition (W. Marslen-Wilson & Zwitserlood, 1989). In order to test these potential effects, we divided the diphone features into three categories (Figure 2.5 upper panel): short words (e.g. "M.AY"), word beginnings (e.g "AE.N" in word "and"), and diphone residues not belonging to either previous groups (e.g. "N.D" in word "undid"). We then estimated the contribution to the prediction of a VM phonemic model based on diphones (but also including words made of single phonemes) coming from each of these three diphone categories. We then normalized these raw contribution values by the proportion of occurrences in each group to effectively obtain an average prediction measure per diphone for each category.

As shown in Figure 2.5, there is a significant difference among short words, word beginnings, and diphone residues in terms of their normalized contribution to the prediction of the diphone features (mixed-effect statistical models with subject as random factor: likelihood ratio test with the nested statistical model that does not include diphone categories: $\chi^2(2) = 52.75, p = 3.5 \times 10^{-12}$). The diphone's dominance in explaining cortical BOLD can be in part, but not solely, explained by responses to short words. The contribution to the prediction of the short words is significantly higher than the beginning of words ($F(2, 10) = 7.44, p = 6.6 \times 10^{-5}$) and diphone residuals ($F(2, 10) = 8.62, p = 1.8 \times 10^{-5}$). It should also be noted that we did not observe distinct cortical subregions within cortical areas with phonemic representations where the responses to short words were systematically higher than those to word beginnings of other diphones (Figure A.8 and Table A.3).

Phonemic versus semantic cortical representations

The cortical areas with significant phonemic representations described above were large and overlapped significantly with cortical areas that have been previously assigned to lexical retrieval and semantic processing (de Heer et al., 2017; Deniz et al., 2019; Huth et al., 2016). To determine whether these regions represent phonemic information, semantic information, or both, we compared prediction accuracy of *Phonemic* VM using single phoneme, diphone and triphone feature spaces to the accuracy of *Semantic* VM using semantic feature space (see Figure A.12 for the hemispheric analysis). The prediction accuracy of *Semantic* VM is obtained from variance partitioning from subtracting the prediction accuracy of *Phonemic* VM from that of *Phonemic-Semantic* VM using all three phonemic feature spaces and the semantic feature space (Figure 2.1, Figure 2.6a). As expected, phonemic and semantic cortical representations overlap extensively (Figure 2.6bd). However, the relative contribution of the phonemic and semantic features can vary across cortical regions.

First, we quantified the phonemic and semantic features' contributions averaged across the entire cortex. These average values serve as a baseline to assess to what extent a particular cortical region of interest is more phonemic or more semantic. We found that the

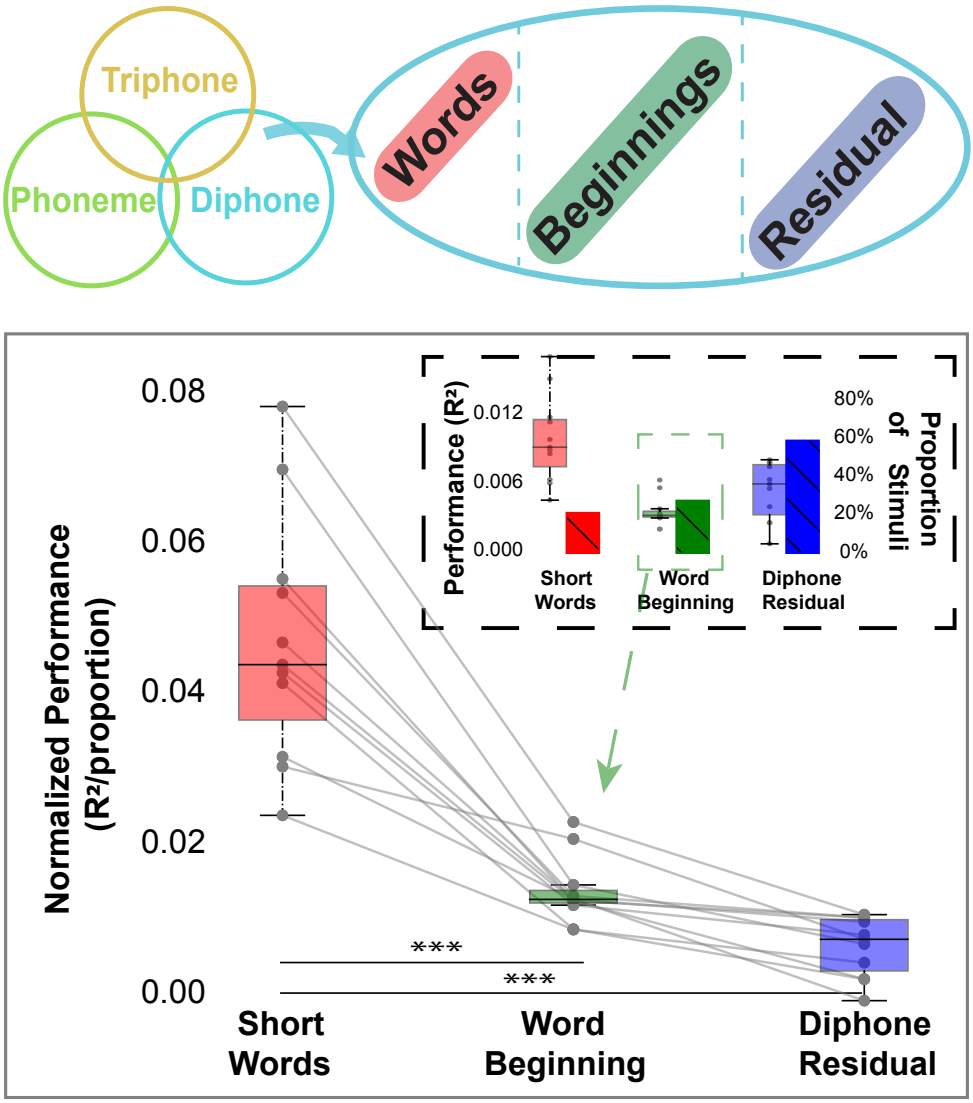


Figure 2.5: **Contribution of distinct diphone categories to phonemic VM.** In order to elucidate whether the unique explanatory power from diphone features could be a reflection of responses to high-frequency short words, **top panel** shows that diphones are split into three categories: short words, word beginnings and diphones that do not belong to these two categories called residual. **Bottom panel** shows that the average normalized prediction performance of short words is significantly higher than word beginnings and the prediction performance of the diphone residual. The inset shows the fraction explained before normalization (lighter box) and the proportion of diphones in each group (solid bar). The statistics have been derived from all the significantly diphone voxels (n = 15,396 to 16,001 voxels) examined across 11 independent subjects. All tests are two sided and corrected for multiple comparisons (*p< 0.05, **p< 0.01 and ***p< 0.001). The exact p-values can be found in the Diphone segmentation for identification, expectations or lexical retrieval? section. The box plots are defined the same way as in Figure 4. Data used to generate this figure has been provided in source data.

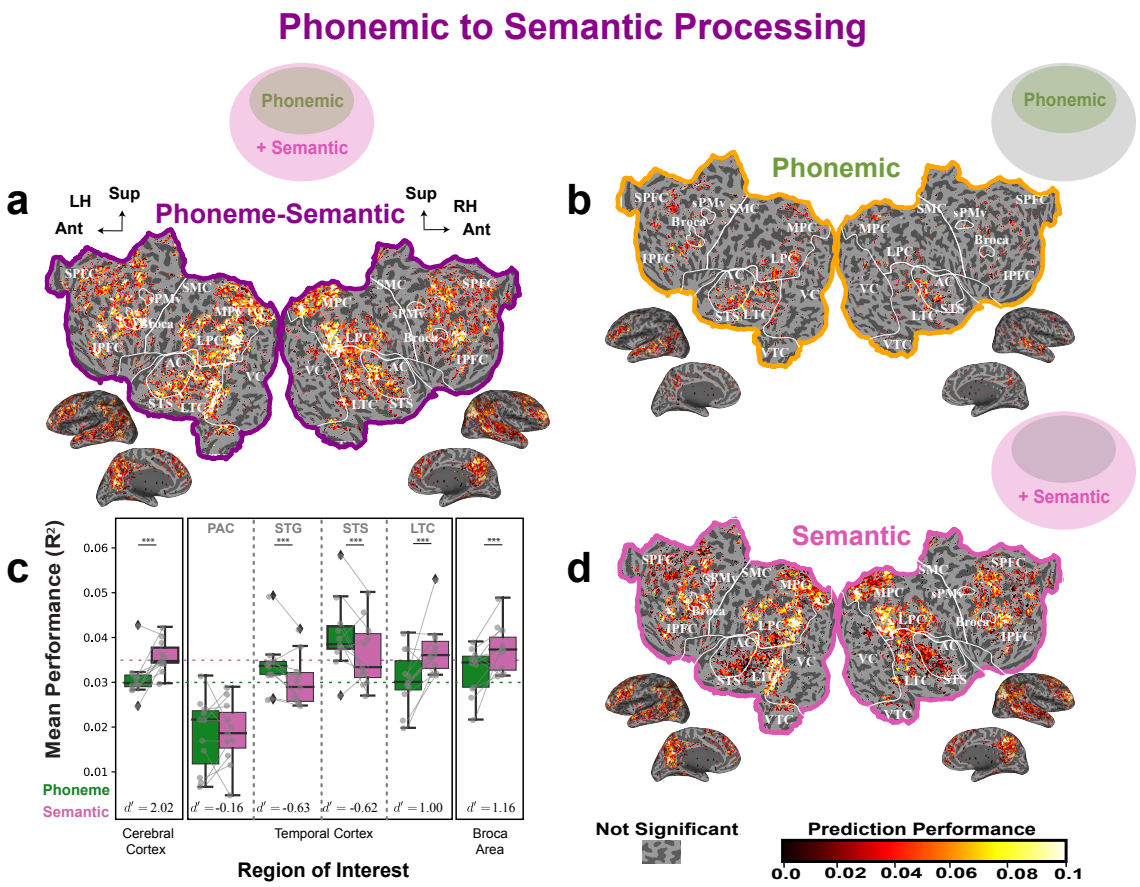


Figure 2.6: **Phonemic to semantic processing.** To explore the cortical phonemic versus semantic processing, a joint phonemic-semantic model consisting of single phoneme, diphone, triphone and semantic feature spaces was constructed. Variance partitioning was then used to obtain the cortical representation of phonemic and semantic models. The flat maps show the prediction performances for these models for one subject. **Panel a** shows that the phonemic-semantic model produces accurate predictions in LTC, VTC, LPC, MPC, SPFC and IPFC. **Panel b** shows that the phonemic model produces accurate predictions in LTC, LPC, MPC, SPFC and IPFC. **Panel d** shows that the semantic feature produces accurate predictions in VTC, LPC, MPC, SPFC and IPFC. **Panel c** show the mean and spread of the raw prediction performance across all subjects in ROIs in the temporal cortex, Broca's area and the entire cerebral cortex. It reveals that the phonemic model produces significantly higher prediction accuracy than the semantic feature in STG and STS. In contrast, the semantic feature produces significantly higher prediction accuracy than the phonemic model in other regions of cortex, Broca's area and LTC (*p$<$ 0.05, **p$<$ 0.01 and ***p$<$ 0.001). The statistics are derived from the performance of significant phoneme-semantic voxels for each ROI (n $= 2,884$ to $380,229$ voxels) across 11 independent subjects. All tests are two sided and corrected for multiple comparisons. The exact p-values can be found in the Phonemic versus semantic cortical representations section. The box plots are defined the same way as in Figure 4. Cortical regions referred are: PAC: primary auditory cortex, STG: superior temporal gyrus, STS: superior temporal sulcus, LTC: lateral temporal cortex, SPFC: superior prefrontal cortex, IPFC: inferior prefrontal cortex, MPC: medial parietal cortex, LPC: lateral parietal cortex, VTC: ventral temporal cortex, MTC: medial temporal cortex, VC: visual cortex, sPMv: ventral speech premotor area. Data used to generate this figure has been provided in source data.

cortex is significantly more involved in semantic processing than phonemic processing (linear mixed-effect statistical model with subject as random effect and feature as a fixed effect;the likelihood ratio test with the nested statistical model that does not include feature spaces, $\chi^2(1) = 712.23, p < 2.2 \times 10^{-16}$). The effect size at the level of the whole cerebral cortex is large: the d-prime obtained from the differences in the additive contribution of the semantic feature and the predictive power of the phonemic feature is $d' = 2.018 + -0.151(2SE)$ (see Figure A.11, Table A.4 and Table A.5 for per subject data and analyses).

Next, we examined the additive contribution of the semantic features in the four ROIs of the temporal cortex defined above. The predictive power of the *Phonemic-Semantic* VM is lowest in PAC, increases in STG and peaks in STS and LTC (Figure 2.6c). Moreover, a linear mixed-effect statistical model with features (two levels: Phonemic and Semantic), ROIs (four levels: PAC, STG, STS, and LTC) and their interactions (3 coefficients) as fixed effects, subjects as the random effect, and the contributions to prediction performance as the response variable, shows that the prediction performance of semantic features is significantly different from that of the phonemic features across ROIs (likelihood ratio test with the nested statistical model that does not include ROIs and the interaction: $\chi^2(6) = 561.87, p < 2.2 \times 10^{-16}$) and across features (likelihood ratio test with the nested statistical model that does not include the interaction: $\chi^2(3) = 80.63, p < 2.2 \times 10^{-16}$). Phonemic features explain a significantly larger portion of response variance than semantic features in STG ($\chi^2(1) = 19.12, p = 8.1 \times 10^{-6}$) and STS ($\chi^2(1) = 51.75, p = 6.3 \times 10^{-13}$). In contrast, semantic features explain significantly more response variance than phonemic features in LTC ($\chi^2(1) = 36.65, p = 1.4 \times 10^{-9}$). These ROI analyses suggest a medial-lateral gradient of phonemic versus semantic representation within the temporal cortex. The most medial regions near the auditory cortex ($d = 0.155 + 0.513$) better represent non-phonemic acoustic features, more lateral voxels in STG ($d = 0.626 + 0.280$) and STS ($d = 0.617 + 0.171$) better represent phonemic features, and the most lateral voxels in LTC ($d = 0.999 + 0.329$) best represent semantic features. We will further delineate the boundary between the phonemic regions and the semantic regions in the last section of the results.

Within Broca's area in the prefrontal cortex, the additive contributions for the semantic features is significantly higher than the phonemic feature ($\chi^2(1) = 15.24, p = 9.5 \times 10^{-5}$). However, the effect size ($d' = 1.158 + -0.593(2SE)$) is smaller than the average observed across all cortical regions.

Phoneme to semantic transitions.

The analyses described above revealed a gradient of phonemic to semantic cortical representation in the temporal cortex extending from its medial and superior region to its more lateral and posterior region. In addition, voxels in the inferior prefrontal cortex (IPFC) appear to be better predicted by phonemic features in their inferior portion and better by semantic features in their superior portion (Figure 2.6, Figure A.13). We thus performed a more detailed anatomical analysis to examine the transition from phonemic to semantic representations in LTC and IPFC.

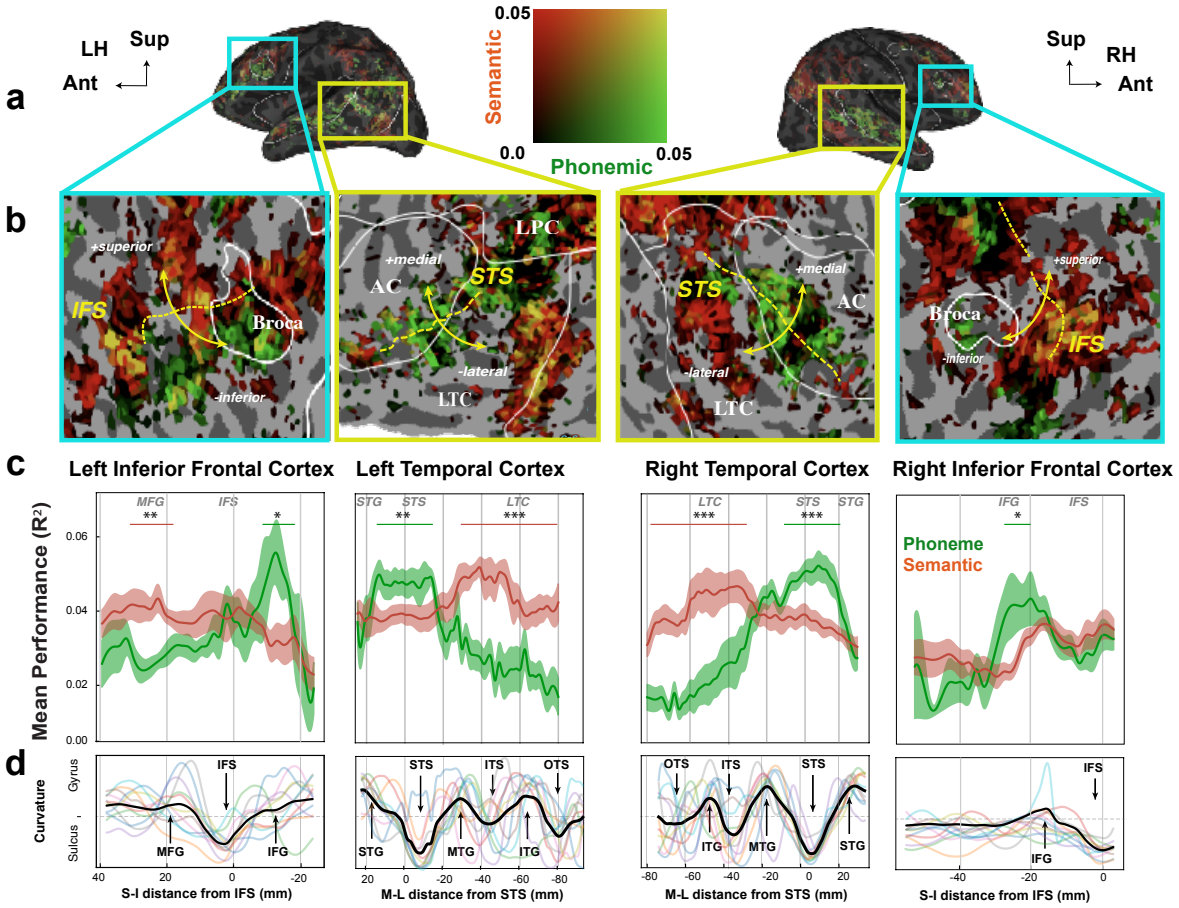


Figure 2.7: **Contribution to VM performance of phonemic vs semantic features in transition areas within the lateral temporal cortex (LTC) and inferior prefrontal cortex (IPFC).** To explore the gradient of semantic and phonemic representation in LTC and IPFC, voxels in LTC and IPFC were assigned a medial-lateral coordinate (M-L distance) based on their distance from the bottom of STS, and an inferior-superior coordinate (S-I distance) based on their distance from the bottom of IFS, respectively. The voxels were binned to 55 discrete bins in the LTC, and 30 bins in IPFC, aggregated across subjects. **Panel a** shows the performance of phonemic and semantic features on the lateral view of an inflated brain. **Panel b** shows the zoomed-in views of panel a in the superior temporal cortex and inferior prefrontal cortex. **Panel c** shows the average prediction performances across all subjects along those two axes. It reveals that the prediction accuracy of phonemic features is significantly higher around STS and inferior to IFS in both hemispheres, while the prediction accuracy of semantic features is significantly higher around ITS in both hemispheres and MFG of left hemisphere (*p< 0.05, **p< 0.01 and ***p< 0.001). It suggests that phonemic to semantic transition happens in a medial-lateral gradient in the LTC and an inferior-superior gradient in the IPFC. Shaded areas show SE estimated across subjects. **Panel d** shows the average cortical curvature in each bin for each subject (colored lines) and aggregated across subjects (bold black line), with positive curvature indicating convexity (as on a gyrus) and negative curvature indicating concavity (as in a sulcus). All tests are two sided and corrected for multiple comparisons. The exact p-values can be found in the Phoneme to semantic transitions. section. Cortical regions referred are: AC: auditory cortex, STG: superior temporal gyrus, STS: superior temporal sulcus, LTC: lateral temporal cortex, SPFC: superior prefrontal cortex, IPFC: inferior prefrontal cortex, MPC: medial parietal cortex, LPC: lateral parietal cortex, VTC: ventral temporal cortex, MTC: medial temporal cortex, VC: visual cortex, sPMv: ventral speech premotor area, IFS: inferior frontal sulcus. Data used to generate this figure has been provided in source data.

We projected the prediction performance of the phonemic and semantic based VMs for all the significant voxels onto a "medial-lateral" axis in LTC (perpendicular to STS), and "inferior-superior" axis IPFC (perpendicular to IFS) (Figure 2.7b, yellow lines). For each voxel in the corresponding region, we computed the distance to the nearest point along the bottom of the STS and the IFS for the LTC and the IPFC respectively. The average prediction performances along those two axes are shown in Figure 2.7c: positive values are more medial (i.e. towards the STG), and negative values are more lateral (i.e. towards the MTG). In LTC the contribution of the phonemic features to the overall prediction was highest around the STS (paired t-tests: -14 to 7mm around STS of left hemisphere: $t(10) = 3.43, p = 6.5 \times 10^{-3}$; -12 to 20mm around STS of right hemisphere: $t(10) = 5.36, p = 3.2 \times 10^{-4}$), while the additive contribution of the semantic features was higher around the Inferior Temporal Sulcus(ITS) (-30mm lateral to STS of left hemisphere: $t(10) = -8.59, p = 6.0 \times 10^{-6}$; -40mm lateral to STS of right hemisphere: $t(10) = -10.87, p = 7.3 \times 10^{-7}$). To coarsely quantify these effects, we estimated the physical location of the "center of mass" of the spatial density of the variance explained by phonemic versus semantic features for each subject (Figure A.14). We found that in LTC, the center of mass of the phonemic feature contribution is consistently and significantly more medial than the center of mass of semantic feature contribution (Figure A.14, paired t-test: left hemisphere: $t(10) = 5.88, p = 1.1 \times 10^{-4}$, 11/11 subjects; right hemisphere: $t(10) = 9.28, p = 1.6 \times 10^{-6}$, 11/11 subjects). In IPFC, the phonemic contribution is significantly higher inferior to the IFS (-23 to -10mm inferior to IFS of left hemisphere: $t(10) = 3.04, p = 6.5 \times 10^{-3}$; -27 to -23mm inferior to IFS of right hemisphere: $t(10) = 3.58, p = 2.3 \times 10^{-2}$), while the additive contribution of the semantic features was higher superior to IFS in the left hemisphere only (20 to 30mm superior to IFS: $t(10) = -4.70, p = 3.3 \times 10^{-3}$)(Figure 2.7c). In summary, these analyses indicate that phonemic to semantic transition happens in a medial-lateral gradient in the LTC and an inferior-superior gradient in the IPFC.

2.4 Discussion

We examined the anatomical location and granularity of phonemic representations in the human cerebral cortex. To clearly delineate the phonemic cortical regions, we also examined the anatomical locations of acoustic and semantic representations and their overlap with the regions identified as being involved in phonemic processing. For these analyses, we trained voxelwise encoding models (VMs) to predict cortical fMRI signals elicited by natural narrative stories. The VMs used a hierarchy of feature spaces that reflect the transformation of speech sound to meanings: power spectrum, phoneme count, identity of phonemes, diphones, triphones and semantic embeddings.

We found phonemic cortical regions are primarily found not only in the superior and lateral temporal cortex (STG, STS, LTC), but also in the ventral and dorsal regions of the inferior parietal lobule (IPL) and in the inferior prefrontal cortex (IPFC). Within all phonemic areas, the segmentation at the level of the diphone yielded the most accurate predictions

of the measured fMRI signal. The phonemic regions were distinct from primary auditory regions, but overlapped with cortical regions where word and semantic representations were found. Comparing the phonemic and semantic based predictions revealed a medial-lateral gradient in the temporal cortex: phonemic information was better represented in more medial regions (STG, STS), while semantic information was better represented in more lateral regions (LTC). An analogous inferior-superior gradient appeared in IPFC: phonemic information was better represented in more inferior regions of the IPFC, while semantic information was better represented in more superior regions.

One of the surprising results from this study is the large cortical areas in the lateral and posterior region of the temporal cortex involved in phonemic processing. In prior research, phonemic processing had primarily been assigned to small regions in the mid STG. In our prior work, we had used a phonemic feature space based on articulatory features describing single phonemes and found that the predictive power of this articulation based phoneme model was limited to small regions in mid STG (de Heer et al., 2017). Similar conclusions were also reached in a meta-analysis of fMRI studies (DeWitt & Rauschecker, 2012) as well as with ECoG recordings showing that electrodes placed on top of STG demonstrated tuning responses for groups of phonemes (Mesgarani et al., 2014). Recently, both intracranial recordings and stimulation studies indicated human mid STG might be a parallel higher level auditory area specialized for low-level speech processing (Hamilton et al., 2021). Our study confirms these previous findings implicating the mid STG in phonemic processing: the VMs based on phoneme counts and the identity of single phonemes had the largest explanatory power in that region.

However, when diphones and triphones are added as distinct features in VMs, the cortical region in the temporal cortex implicated in phonemic processing is much larger. It extends laterally from STG to STS and much of the LTC. This large phonemic area in the LTC further extends into the ventral and posterior area of the IPL without a clear break along the temporoparietal junction. Since diphones and triphones include many common short words, the phonemic regions defined here might also be cortical regions involved in lexical retrieval. In another word, these regions could be involved in the segmentation of the speech signals into known words. In order to disambiguate lexical processing from phonemic processing, we estimated the contribution short words made to our VM (Figure 2.5). Indeed, we found that approximately half of the explanatory power originating from the diphone features could be attributed to short words. In addition, the semantic feature space based on word embedding is by necessity correlated with a lexical space. We showed that there was a large number of phonemic voxels where the additional predictive power obtained from the semantic feature space was significant. Thus, we propose that the large cortical regions where the Phonemic VM yielded good predictions of the BOLD response are cortical areas involved both in phonemic and lexical processing.

In the more lateral regions of the temporal cortex, the semantic features yielded additional predictive power that was greater in magnitude than the one obtained by phonemic features only, suggesting that this lateral region constitutes a cortical network implicated in the phonemic/lexical to semantic transition. Both this network in the lateral temporal cortex

(R. Carter, 2019) and the adjacent region in the IPL (Binkofski et al., 2016) have been implicated in lexical retrieval and speech comprehension in multiple studies. We also found discrete locations within IPFC around the IFS where the phonemic model yields better predictions than the semantic model. These locations might also be involved in lexical retrieval potentially under a different linguistic context than the one found in lateral temporal cortex (Petrides, 2016). For example, it has been suggested that these regions in the IPFC are implicated in lexical retrieval with higher linguistic load such as with increase word cohort competition (Zhuang et al., 2011).

One key finding from this study is that, within the phonemic/lexical regions in the LTC, IPL and IPFC, segmentation occurs principally at the level of the diphone. This finding is consistent with basic phonetic theory. In natural speech, single phonemes don't occur in isolation and it is well known that single phonemes are acoustically different when produced in combinations, a phenomenon known as coarticulation (Öhman, 1966). In other words, the acoustic features characterizing single phonemes vary depending on the phonological context. This acoustic variability prevents a simple transformation from temporal-spectral acoustical features to the identification of a single phoneme. This acoustical variability is reduced when segmentation occurs at the level of diphones. Since the number of diphones used in languages remains relatively small (e.g. around 860 in English) compared to memory capacity of the human brain, it is reasonable to postulate that the transformation between an acoustic based neural representation and a phonemic based neural representation of a speech stream could therefore rely principally on a direct operation on spectral-temporal features to extract the diphone identity. As caveats, first, we have not exhaustively explored segmentation based on other speech features. For examples a segmentation based on the speech signal envelope (e.g. syllables, envelope peaks, etc) have shown to have large explanatory power for recovering the neural signals in the STG obtained in ECoG recordings (Oganian & Chang, 2019). Second, we have only investigated models with a static segmentation. More sophisticated models could also include an adaptive segmentation model based both on acoustic features and linguistic features as proposed in the adaptive resonance theory (Grossberg, 2003). Third, the dominance of diphones may also reflect the engagement of predictive speech processing mechanisms which become more important in noisy situations (Trecca et al., 2019). In particular, under the noisy fMRI scanning environment, diphones could be more reliably represented and resistant to noise than single phonemes. On the other hand, the predictive power of VM that were solely based on the statistical properties of diphone sequences and ignored diphone identities was significantly smaller (see Figure A.9 and Figure A.10). Finally, our conclusions are based on measures of predictions of nested models of phonemic features. Based on a systematic variance partitioning of single phonemes versus diphones and triphones, we are confident that the unique variance explained by diphones over single phonemes is substantial while the unique variance explained by triphones over diphones is small. However, we don't provide direct evidence for diphone specificity, for example in the form of finding a cortical map that is systematically organized along diphone structure but not triphone structure. Additional explorations on the exact nature of the cortical representation for phonemic segments is needed.

As mentioned above, the over-representation of diphones in phonemic brain regions might also be related to word segmentation, a key process in lexical retrieval. Many common short words are diphones. To begin to assess the extent of putative word processing within the regions identified here as phonemic, we sorted diphones into three groups: short words, word beginnings and other. We found that the diphones that were short words were better represented than word beginnings and the other diphones. This over-representation might thus reflect the role of STG, STS and/or LTC in word segmentation and lexical retrieval. However, within the phonemic/lexical regions, we did not find subregions that were principally phonemic or principally lexical. These results are consistent with lexical retrieval mechanisms as revealed and established in behavioral studies that the process of word recognition entails gradual integration across multiple phonemes (Grosjean, 1996; W. D. Marslen-Wilson, 1987). Thus phonemic and lexical processing appear to be intertwined both mechanistically and in cortical functional mapping.

Finally, we also assessed the extent with which our findings on the granularity of phonemic segmentation could be a consequence of the spatial and temporal resolution of the fMRI signal. The low temporal resolution of fMRI due to sluggish hemodynamic response function (HRF) signal has long been considered as a challenge to study fast brain processing like speech segmentation (Ghuman & Martin, 2019). Our simulations (Materials and Methods), however, show that the capability of detecting fast neural activities underlying speech processing using fMRI is not necessarily constrained by this low temporal resolution as long as the fast speech events being studied are sufficiently variable at the temporal scaled captured by fMRI. For example, although the phoneme counts in speech are high, in a given TR there will be a small number of counts of any particular phoneme or diphone, etc. Thus, in order to have temporally sparse events in the BOLD signal, the sensitivity of a single voxel must be sufficiently high as to only be responsive to a small number of phonemic units. The threshold of sensitivity is determined by the signal to noise ratio (SNR) and the data size. It is also a function of the spatial resolution of the fMRI (Mumford & Nichols, 2008; Poldrack et al., 2017).

Another key finding from this study is that two principal regions where the phonemic/lexical representation to semantic meaning transition occurs were identified: a medial-lateral gradient in the LTC and an inferior-superior gradient in the IPFC around IFS. In the LTC, the phonemic/lexical representation dominates in the more medial parts around the STS, while the semantic representation dominates in the more lateral areas along ITS. In our study, the phonemic/lexical LTC area includes the posterior medial temporal gyrus (pMTG) and the posterior part of the inferior temporal sulcus (pITS). These regions have been assigned as a core phonological to lexical interface in a large body of research based on clinical, neurophysiological and imaging research (reviewed in (Poeppel et al., 2012)). While our results and conclusions are consistent with that published work, we also found that the posterior LTC extends into the ventral and posterior area of the IPL making a larger continuous region of phonemic/lexical representation to semantics transition. On the one hand, the junction of the temporal and the parietal cortex (in particular the regions that include the supramarginal gyrus and angular gyrus, that together are known as Wernicke's

area) has been extensively implied in speech processing. For example, it is known to be involved in categorical phoneme perception (Turkeltaub & Coslett, 2010) and more generally in auditory to motor integration (Hickok & Poeppel, 2016). On the other hand, the posterior IPL had not been identified previously as a region involved in the transition from phonemic/lexical representations to semantics. It has, however, been identified as a core region where visual representations are transformed into visual objects with semantic meaning (Huth et al., 2012). Similarly, it is a region of transition between a visual representation of the letters in written text to the meaning of the words in the text (Deniz et al., 2019). The ventral and posterior IPL therefore appears to be implicated in the extraction of meaning from sensory stimulus irrespective of the modality. For the speech sound to semantic meaning transformation, it remains to be seen whether the ventral and posterior IPL serves a different function as the neighboring region in the temporal cortex. For example, one could hypothesize that the subregion in the temporal cortex specializes in the phonemic/lexical to meaning transformation supported by linguistic priors, and that the subregion in the parietal cortex specializes in the phonemic/lexical to meaning transformation by engaging the visual imagery that can be elicited by listening to narrated speech stories.

In the inferior frontal cortex, we found an inferior-superior gradient with a dominant phonemic/lexical representation around IFS and a dominant representation for semantics along the MFG. This finding is also in accordance with prior studies. IPFC is known to play a significant role not only in speech related sensori-motor integration but also in phonological processing, in particular when it involves analyses requiring segmentation based on lexical features (Burton, 2001; Zhuang et al., 2011). In non-humates primates, high-level auditory neurons in IPFC have also been shown to respond categorically to species specific vocalizations according to their behavioral meaning (Gifford III et al., 2005).

In summary, by investigating phonemic segmentation in long fMRI recording sessions obtained while subjects listened to engaging narrative speech, two large regions involved in the representation of phonemic and lexical features were described: one in the superior and lateral temporal cortex and ventral/dorsal parietal cortex and a second one in the IPFC. In both regions phonemic features beyond those found in short words increase predictive power of encoding models. It indicates phonemic and lexical processing may be intertwined. In both LTC and IPFC, a systematic gradient in the relative dominance in phonemic/lexical versus semantic representation was also observed. The actual potential distinct roles of these two large cortical networks, or of sub-networks within these two large networks, in phonemic processing and in the phonemic/lexical to semantic transformation remain poorly understood. We also demonstrated here that fMRI analyses could be leveraged to study cortical speech processing, even those occurring at time scales that are faster than the temporal resolution of the BOLD signal. Thus, further exploring the predictive power of more complex VMs applied to fMRI data could allow for a finer functional parcellation of phonemic/lexical specific areas. For example, to gain further insights on networks dedicated to word segmentation, one could test the additional predictive power of phonemic based VM that also take into account word segmentation expectations based on linguistic context assessed at multiple time-scales (N. Ding et al., 2016; Klimovich-Gray et al., 2019; Leonard et al., 2015; Lerner

et al., 2011; Zhuang et al., 2011). Similarly, distinct roles for the subnetworks involved in phonemic/lexical to semantic transitions could also be tested. For example, linguistic priors that play a role in predicting meaning from the beginning of sentences could also be included as features in semantic based VMs (Piai et al., 2016). Lastly, VMs could also be used to explore the speech representations of bilingual or multilingual subjects listening to stories in multiple languages including some that subjects do not understand. Such experiments would allow researchers to further distinguish phonemic segmentation and identification that occurs independently of understanding, or that might be common across multiple languages. These analyses would therefore be useful to also more clearly delineate the purely phonemic regions from those implicated in lexical retrieval, if such regions exist.

2.5 Materials and Methods

Participants

Structural and functional brain data were collected from seven male subjects ($S1$: age 26 , $S2$: age 31 , $S5$: age 30, $S6$: age 25, $S7$: age 36, $S9$: age 24, $S10$: age 24), and four female subjects ($S3$: age 28, $S4$: age 25, $S8$: age 24, $S11$: 31). Each subject's handedness was evaluated by the Edinburgh handedness inventory (Oldfield, 1971). All subjects were healthy and had no reported hearing problems. The use of human subjects in this study was approved by the UC Berkeley Committee for the Protection of Human Subjects. A written statement of informed consent has been obtained from each subject.

Stimuli

The stimuli in these experiments were pre-recorded stories from the Public Radio Exchange (PRX) radio show "The Moth Radio Hour", which has been used in previous studies of our lab (de Heer et al., 2017; Deniz et al., 2019; Huth et al., 2016). These sounds have been annotated for their word and phonetic content. They are engaging stories that capture the attention of the subjects. The scanning sessions lasted approximately 2.5 hours and included time needed to play the audio/visual stimuli used as generic localizers and two one-hour sessions during which the subjects listened to Moth stories.

The stimuli were split into separate model estimation and model validation sets. The model estimation stimulus-set consisted of ten 10− to 15−minute stories played once each. The length of each scan was tailored to the story and also included 10 seconds of silence both before and after the story. Each subject heard the same 10 stories, 5 of which were told by male speakers and 5 by female speakers. The model validation stimulus-set consisted of a single 10−minute story told by a female speaker that was played twice for each subject in order to estimate voxel response reliability and noise ceiling. This resulted in 3737 time points (sampled at TR) for the training dataset and 291 time points for the validation dataset.

Auditory stimuli were played over Sensimetrics $S14$ in-ear piezoelectric headphones (Sensimetrics MA, USA). These headphones provide both high audio fidelity and some attenuation of scanner noise. A Berhinger Ultra-Curve Pro hardware parametric equalizer was used to flatten the frequency response of the headphones (Berhinger, Las Vegas, USA). The sampling rate of the stimuli in their digitized form was 44.1 kHz and the sounds were not filtered before presentation. Thus, the potential frequency bandwidth of the sound stimuli was limited by the frequency response of the headphones from 100 Hz to 10 kHz. The sounds were presented at comfortable listening levels. Stories were manually transcribed and converted into separate word and phoneme representations (see (de Heer et al., 2017; Huth et al., 2016)).

MRI data collection and preprocessing

Structural MRI data and blood oxygen level dependent (BOLD) fMRI responses from each subject were obtained while they listened to approximately 2 hours and 20 minutes of natural stories. For nine of the subjects, these data were collected during two separate scanning sessions that lasted no more than 2 hours each. For two of the subjects ($S1$ and $S5$) the validation data (two repetitions of a single story) were collected in a third, separate session. MRI data were collected on a $3T$ Siemens TIM Trio scanner at the UC Berkeley Brain Imaging Center, using a 32-channel Siemens volume coil. Functional scans were collected using a gradient echo-EPI sequence with repetition time (TR) = $2.0045s$, echo time (TE) = $31ms$, flip angle = $70degrees$, voxel size = $2.24x2.24x4.1$ mm, matrix size = $100x100$, and field of view = $224x224$ mm. 32 axial slices were prescribed to cover the entire cortex. A custom-modified bipolar water excitation radiofrequency (RF) pulse was used to avoid signals from fat tissue. Anatomical data were collected using a T1-weighted MP-RAGE (Brant-Zawadzki et al., 1992) sequence on the same $3T$ scanner.

Each functional run was motion-corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL 4.2 (Jenkinson & Smith, 2001). All volumes in the run were then averaged to obtain a high quality template volume. FLIRT was also used to automatically align the template volume for each run to the overall template, which was chosen to be the template for the first functional run for each subject. The overall template is obtained within each participant over all runs. These automatic alignments were manually checked and adjusted for accuracy. The cross-run transformation matrix was then concatenated to the motion-correction transformation matrices obtained using MCFLIRT, and the concatenated transformation was used to resample the original data directly into the overall template space. This overall template space is also obtained within each participant across all runs. Low-frequency voxel response drift was identified using a $2nd$ order Savitsky-Golay filter with a $120-$second window, and this was subtracted from the signal. After removing this time-varying mean, the response was scaled to have unit variance (i.e. z-scored). Structural and functional MRI were combined to generate functional anatomical maps that included localizers for known regions of interests (ROIs). The method used to determine these ROIs is explained in detail in (de Heer et al., 2017).

Feature space and model construction

In order to localize the phonemic brain regions, to investigate phonemic segmentation and the phoneme to word/meaning transitions, we constructed six distinct feature spaces: time-varying power spectrum, phoneme count, single phoneme, diphone, triphone, and semantic features.

The time-varying power spectrum was obtained by estimating the power for 2-second segments of the sound signal between $25Hz$ and $15kHz$, in $33.5Hz$ bands (number of frequency bands = 448), using the classic Welch method for spectral estimation density as we have done previously (de Heer et al., 2017).

The phonemes rate features consisted of phonemes' counts per TR. The time-varying power spectrum and the number of phonemes were combined to make the baseline feature space (number of features = $448 + 1 = 449$). The baseline feature space quantifies the presence or absence of sounds and speech sounds. As we were interested in brain regions that process phoneme identity and meaning, the fraction of the BOLD response explained by the baseline feature space was analyzed separately and then subtracted.

The single phoneme feature space is composed of a binary vector encoding the presence or absence of 15 consonants and 24 vowels of English that we generated based on the CMU pronouncing dictionary (Lenzo et al., 2007). Although the numbers of all possible diphones ($39^2 = 1521$) and triphones ($44^3 = 59319$) are large, English words utilize only a subset of all possible combinations. For the stories used in this study, we found 835 diphones and 2677 triphones. When we added combinations found in the larger IPhOD data set, we obtained altogether 858 possible diphones and 4841 possible tri-phones (Vaden et al., 2009). For diphone and triphone feature space, they are composed of one-hot code encoding the presence or absence of all possible diphones or triphones. In semantic feature space, each word is represented as a 985-dimensional vector based on word co-occurrence statistics with the 985 basis words from Wikipedia's "List of 1000 Basic Words" (de Heer et al., 2017; Huth et al., 2012). All the feature vectors were downsampled to the TR using Lanczos filtering (Figure 2.1).

Voxelwise model fitting and validation

Based on these feature spaces, we created different linear encoding models (linearized regression) in order to predict the time-varying BOLD response of each voxel of each subject from the time varying stimulus (Figure 2.1). We have called this approach voxelwise modeling (VM). First, we constructed a baseline VM that used power spectrum and phoneme count as regressors ($X_B \in R^{nt \times 449}$) to predict the BOLD response ($Y \in R^{nv \times nt}$), where nt is the number of time points ($nt = 3737$) and nv is the number of voxels (varies across subjects; range: $73,023 - 92,970$). We then subtracted this prediction from the measured BOLD response to obtain a response residual ($Y_{res} \in R^{nv \times nt}$) before fitting VMs based on phonemic and semantic identities. Subtracting the prediction from this baseline VM is needed in order to distinguish variance in the BOLD response that is simply due to the presence

versus absence of phonemes or words from the variance that is dependent on the identity of phonemic and semantic features. Second, one joint phonemic model, using single phonemes, diphones and triphones features ($X_{1+2+3} \in R^{nt \times 5738}$) as regressors, was fitted to predict the BOLD response residual, Y_{res}. Finally, the phoneme-semantic VM was obtained by using all three phonemic features with semantic features ($X_{1+2+3+4} \in R^{nt \times 6723}$). Before fitting the VM weights (also known as model coefficients in the context of linear regression), the regressors X and the BOLD response Y were z-scored. BOLD response was z-scored separately for each story to control for the random effect of the story.

In order to account for the temporal integration time constants of both the neural processing and the BOLD response, our linear models predict the BOLD response at time t from signal features evaluated in four time windows of 2 seconds each and starting at $t-2s$, $t-4s$, $t-6s$ and $t-8s$. This is accomplished by concatenating feature vectors that had been delayed by $2s$, $4s$, $6s$, and $8s$. This yields $1,796$ features (448 spectral power x 4 + 1 phoneme count x 4) for the baseline VM, $22,952$ features (39 phonemes x 4 + 858 diphones x 4 + 4841 triphones x 4) for the phonemic VM, and $26,892$ features (39 phonemes x 4 + 858 diphones x 4 + 4841 triphones x 4 + 985 words x 4) for the phoneme-semantic VM.

Due to the relatively large number of features relative to nt, we used regularized regression techniques to estimate the VM weights to prevent overfitting. Regularization was achieved using Tikhonov regression with different levels of regularization used for different feature spaces and delays for each voxel and each subject. Along the spatial dimension (features), a different regularization hyperparameter was used for each feature space. For example, all features belonging to the single phoneme feature space would shrink according to λ_1 (the regularization hyperparameter corresponding to the inverse variance of the Gaussian prior), while all features belonging to the diphones would shrink according to λ_2 (de Heer et al., 2017; la Tour et al., 2022; Nunez-Elizalde et al., 2019). We have called this type of regularization banded ridge regression because the diagonal covariance matrix that characterized the Gaussian prior has bands corresponding to the equal values of for all features belonging to the same feature space. In addition to the regularization along the spatial dimension, our regularization procedure also acted along the temporal dimension (the features in the four delay periods). As mentioned in the previous paragraph, the weights along the temporal dimension include the neural integration time and the time course of the hemodynamic response. As part of the regularization procedure, one can capture the expected effect due to the hemodynamic response by using a hemodynamic response function (HRF) temporal prior. This temporal prior was fixed and effectively enforces different levels of regularization for different time delays. An efficient analytical solution for banded ridge regression was implemented using the kernel method as described in (la Tour et al., 2022; Nunez-Elizalde et al., 2019).

For each feature space and each voxel, we tested 10 possible regularization hyperparameters, (log spaced between 10^2 hyperparameters for the baseline VM, 10^3 for the phonemic VM and 10^4 for the phoneme-semantic VM. This hyperparameter search was performed for each voxel of each subject. The optimal hyperparameter was found by maximizing the cross-validated coefficient of determination (R^2) using a 10 fold cross-validation procedure on the

training data set (see below). In order to prevent overfitting, these optimal hyperparameters were tested on a separate validation data set.

Finally to quantify the goodness of fit of the VMs, the fitted VMs were used to predict BOLD responses to a separate story that had not been used for hyperparameter optimization mentioned above and model fitting. Prediction performance was then estimated from the R^2 calculated between predicted and actual BOLD responses for each voxel over the 291 time points for this validation data.

Variance Partitioning

In order to quantify the variance and localize the cortical regions uniquely explained by individual feature spaces and any combination of feature spaces, we performed a variance partitioning procedure. We first fitted a joint phonemic model with all phonemic features (single phonemes, diphones and triphones)(de Heer et al., 2017). Then, we used the hyperparameters (to the optimal shrinkage) obtained from this joint phonemic model of each block of features in the banded regression so that the same sub space of features obtained by the ridge is used in all models. We computed the variance of seven possible partitions: the unique variance of each feature and all the possible combinations of these features (single phonemes + diphones, diphones + triphones, single phonemes + triphone, single phonemes + diphones + triphones). Afterwards, we performed a variance correction on R^2 that is used to eliminate biases in the R^2 estimation that could lead to nonsensical results.

In a separate analysis, we examined the contribution of distinct types of diphones features. First, we separated diphone into three categories: diphones that are short words (e.g. "M.AY" as word "my"), diphones that are at the beginning of words (e.g. "AE.N" in word "and"), and all the other diphones. We then estimated the contribution of each diphone category to the prediction by deleting the diphones that did not belong to the category being considered; more precisely we set the contribution to the diphone feature vector from the out-of-considered-category diphones to zero. With this procedure, we divided the predicted signal obtained from the diphones into three components, one obtained for each mutually exclusive diphone category.

Definition of ROIs

Regions of interests including auditory cortex broadly defined (AC), Broca's area, and ventral speech premotor area (sPMv) were defined based on standard functional localizer scans. Voxels that were responsive when the subject listened to 10 repetitions of a one-min auditory stimulus with 20-sec segments of music, speech and natural sound are considered to belong to AC. AC thus includes both primary and secondary auditory cortical areas. When the subject continuously subvocalizes self-generated sentences, the active voxels located at the triangular part of the inferior frontal gyrus are determined as Broca's area and voxels located at the premotor cortex as sPMv. The repeatability of the voxels' response was

calculated as an F statistic given by the ratio of the total variance responses over the residual variance.

Other ROIs were defined based on speech processing relevant anatomical landmarks (Binder et al., 2009; Bookheimer, 2002; Hickok & Poeppel, 2007; Huth et al., 2016; Price, 2010) obtained from the structural MRI scan. In particular, the inferior prefrontal cortex (IPFC) contains cortical regions ventral to inferior frontal sulcus (IFS), while superior prefrontal cortex (SPFC) contains regions dorsal to superior frontal sulcus (SFS). The peak of the superior parietal gyrus separates the lateral parietal cortex (LPC) from medial parietal cortex (MPC). LPC contains supramarginal gyrus (SMG) and angular gyrus (AG). The peak of the inferior temporal gyrus is used to separate the lateral temporal cortex (LTC) from ventral temporal cortex (VTC).

Statistical Analyses

The statistical significance of the cross-validated coefficient of determination R^2 was estimated using a permutation analysis. First, We obtained a set of regularization hyperparameters from fitting the phoneme-semantic VM (using single phoneme, diphone, triphone and semantic features) for each subject. Then, a permuted R^2 for each model of each subject was obtained by refitting the model using this set of hyperparameters and the shuffled regressors of each feature space within the model. This process was repeated $1,000$ times to generate the null distribution of cross-validated R^2 for each voxel of each subject. Based on the values of this null distribution, we used variable thresholds yielding a significance level of 1% corrected using the False Discovery Rate (FDR) procedure (Benjamini & Hochberg, 1995) to determine statistical significance.

The raw R^2 values, or the partial R^2 raw values obtained from variance partitioning, obtained for each voxel and each subject were then aggregated to make inferences about reliable patterns of correlations between features in the speech stimuli and BOLD activity in specific brain regions. Alternatively, we also used a winner-take-all approach assigning the feature space that yielded the highest prediction in a specific comparison (e.g. single phoneme, diphone, triphone) as a label to each voxel. The number of voxels for each label and each ROI was then taken as the measure of the effect (Table A.2 and Table A.5).

In such aggregations, it is incorrect to simply consider all voxels found in a given ROI for all subjects as independent observations. Because of pseudo-replication, in such erroneous approach, an effect (e.g. phonemic based R^2 > semantic based R^2) found in one single subject with many voxels but not in the other subjects could yield a false result of statistical significance with a grossly deflated p-value. More conservatively, one could first average the data for each subject (e.g. estimating the average "phonemic based R^2" - "semantic based R^2" for each subject) and perform statistical inference on those subject-averaged values. Although this approach is clearly more correct and conservative, it also gives the same weight to a subject with a small number of significant voxels in the ROI than to a subject that has a higher number of significant voxels in the same ROI. A more principled solution is therefore to explicitly model the statistical dependence of the raw voxelwise R^2 or counts

in each subject. For this purpose, we used statistical mixed effect models with the subject as the random factor. In mixed effect models, the penalty for the random effect in the overall likelihood that is being maximized for a subject with a smaller number of significant voxels is smaller than for a subject with more significant voxels. In this manner, a likelihood based weighting of the subject data is performed.

In this paper, we used this mixed-effect statistical modeling when reporting the effect sizes and statistical analyses in the text and figures in the main section of the paper. We also repeated all statistical analyses by first performing averages for each subject and then applying classical statistical inference tests on those data. The per-subject analyses and statistics are found in supplementary tables. The statistical results, effect sizes and conclusions obtained using these two approaches were practically identical.

For the statistical analyses presented in the main text, linear mixed-effects statistical models with subjects as random effect (R: *lmer*) were used to compare the raw R^2 for different phonemic and semantic based VMs across in PAC, STG, STS, LTC, Broca's area, and the entire cortex and to test for lateralization effects. Generalized mixed-effects statistical models (R:*glmer*) with binomial distributions were used to compare the fraction of best explained voxels by the semantic feature based VMs relative to the phonemic feature space VMs across ROIs. Multinomial mixed-effects statistical models (R: *mblogit*) were used to estimate the fraction of best explained voxels by the single phoneme vs diphone or triphone feature based VM while taking into account the variability across subjects (the random effect). Custom R code was written to calculate the expected likelihood of the null model corresponding to a fixed effect of equal probability ($\frac{1}{3}$) and a random effect yielding a saturated statistical model that predicts the actual empirical probability found for each subject. Likelihood ratio tests were then performed to estimate whether the likelihood of the mixed-effects multinomial statistical model with fitted fixed effects (i.e. non zero intercepts) was greater than the likelihood of the null model. The standard errors in the mixed-effects model were estimated by bootstrapping over subjects. The bootstrap was performed using the *emmeans* R package for *glmer* and with custom code for *mblogit*.

We used Cohen's d' to report the effect sizes of the difference of the mean prediction performance between phonemic and semantic based VM model, as well as the mean prediction performance among single phoneme, diphone and triphone features. In order to obtain the dprime of the phonemic and semantic based VM model, we first used a mixed-effect statistical model with subject as random factor and models as fixed factor. The estimated standard deviation of the random effect of this mixed-effect model captures the variance in mean prediction performance found across subjects, σ^2_{SUB}; the coefficients of the fixed effects of this model captures the difference in the mean prediction performance between VM models based on different features. For example $\mu_S - \mu_P$ where μ_S is the mean prediction performance for semantic based VM and μ_P is the mean prediction performance for the phonemic based VM, and the standard error of these same coefficients gives the standard error in this same mean prediction difference, $\sigma_{SE(S-P)}$, the effect size (d') was then estimated by dividing this difference in average performances by the standard deviation, both estimated in the mixed effect model across subjects:

$$d' = \frac{(\mu_S - \mu_P)}{\sigma_{SUB}} \tag{2.1}$$

The standard error of the d' was calculated from the standard error for the difference in mean predictions:

$$SE(d') = \frac{\sigma_{SE(S-P)}}{\sigma_{SUB}} \tag{2.2}$$

In order to obtain the dprime among the single phoneme, diphone and triphone features (d'_{phn}) with the same logic, we first calculated the dprime of the diphone versus triphone features (d'_{DiTri}) and the dprime of the diphone versus single phoneme features (d'_{SinDi}). Then, we simply averaged these two dprimes.

$$d'_{phn} = \frac{(d'_{DiTri} - d'_{SinDi})}{2} \tag{2.3}$$

For the standard deviation of the dprime among three phonemic features ($SE(d'_{phn})$), we first computed the sum of the square of the standard error of the dprime of the diphone versus triphone features ($SE(d'_{DiTri})$) and diphone versus single phoneme features ($SE(d'_{SinDi})$), and then we calculated the half of the square root of this sum.

$$SE(d')phn) = \frac{\sqrt{SE(d'_{DiTri})^2 + SE(d'_{SinDi})^2}}{2} \tag{2.4}$$

In addition, we also reported a d' for each subject in Table A.1 and Table A.4. It was calculated as the difference between the means of the average prediction performance between the phonemic and semantic based VM model for each subject divided by the pooled standard deviation. The mean and SE of those per-subject dprimes are summarized in the table as well.

To quantify the difference in the number of voxels that are best explained by phonemic or semantic based VMs, the log of the odds (or logit) that a voxel is labeled as "semantic" (vs "phonemic") was used to calculate the the effect size:

$$l = log(\frac{p_S}{1 - p_S}) \tag{2.5}$$

Similarly to compare the number of voxels that are best explained by single phonemes, diphones or triphones, the probability that a voxel is labeled as any of the phonemic label was estimated using a reference in the multinomial statistical model. Here we took the diphone as the reference and report negative logits, for example for the probability of single phonemes vs diphones:

$$l = -log(\frac{p_S}{p_D}) \tag{2.6}$$

The probability values are obtained from mixed effect binomial (semantic versus phonemic) or multinomial (single phonemes vs diphone vs triphone) statistical models in the main

text and, just as for d', this calculation is repeated with the "equal weight" per subject and shown in Table A.2 and Table A.5.

Simulation: Feasibility and Limitations

In this study we are proposing to use the VM approach to predict BOLD activity based on the segmentation of the speech stream occuring at the level of small phonemic units: single phonemes, diphones and triphones. The temporal resolution of the BOLD signal is known to be particularly slow due to the sluggish hemodynamic response function whose power is concentrated below 0.2Hz (Henson & Friston, 2007). Here, we investigate the processing of phonemic units which have rates of up to 20Hz in spoken speech. Stated in this manner, the time scales of phonemes and the BOLD response are mismatched by a factor of 100. Thus, can we prove that our approach could in fact measure correlations between specific phoneme identity (or their combination) and BOLD activity?

As an initial step, one can gain some insight on the feasibility of the approach by examining the signals generated at TR rates for single phonemes or combinations as shown in Figure 2.1. Indeed, although the phoneme counts in speech are high, in a given TR there will be a small number of any particular phoneme. One finds mostly counts of zero, one, two or three (and consequently, even smaller numbers of particular diphones and triphones). Moreover, this count changes from TR to TR yielding a clear signal. Thus, as long as voxels are sufficiently sensitive, in the sense of responding more strongly to a small number of phonemes, one should be able to detect this signal in the BOLD response given sufficiently long recording times. To quantify what is "sufficiently sensitive", we thus performed a series of simulations with realistic signal to noise ratio (SNRs) and data size (corresponding to our $2h$ recording time) to evaluate the recovered VM coefficients and compare them to their actual values used to generate the simulated data. The simulation results are based on the SNR values of average responsive voxels (SNR$=1$) for subjects with overall high SNR across all voxels (see Figure A.16). Thus, we are simulating the average case scenario in order to assess the limits of what we would be able to detect inn terms of phonemic feature sensitivity given our data size and for the most responsive voxels.

First, we examined whether our approach could distinguish between voxels sensitive to single phonemes versus those sensitive for diphones, or combinations of single phonemes and diphones. The single phoneme-only simulated voxel was by construction sensitive to three randomly chosen distinct phonemes with equal weights. Similarly, the diphone-only simulated voxel was sensitive to three randomly chosen diphones. The simulated voxel that was sensitive to the combination of single phonemes and diphones had a BOLD signal that was simply the sum of the BOLD signal for the single phoneme and the diphone model voxel. To generate a realistic BOLD signal, the signals obtained from the convolution were low-passed filtered below 0.1Hz and Gaussian white-noise was added to fit the SNR of one of a highly responsive voxel. The voxel SNR was estimated by calculating the coherence between two repeats of the same speech stream in the validation data set (Hsu et al., 2004). We repeated this procedure 30 times by resampling among single phonemes and diphones

to generate a distribution of simulated results. As shown in Figure A.3A, for the stimulus duration of 3737 TRs (2 hours), we were able to recover the VM coefficients and expected values for R^2. Thus, for average sensitive voxels, our procedure could distinguish BOLD responses reflecting single-phonemes-only sensitivity, diphone-only sensitivity and responses that resulted from sensitivity to the mixtures of particular single phoneme and diphone occurrences.

Next, we examined the effect of voxel sensitivity on our ability to determine whether representations based on phoneme identities would yield better results than those based on just the phoneme count. With the decrease of the sensitivity of putative voxels (in the sense that it responds to a growing number of single phonemes or diphones), we may fail to differentiate between VMs based on phonemic identity versus phonemic rate. As shown in Figure A.3B, as a voxel's sensitivity decreases (i.e. as it becomes tuned to a larger number of phonemic identities), the predictive power of the VM based on phoneme count increases and approaches that of the VMs based on phoneme identity. The roughly equal prediction performance between the phoneme count and identity based VMs was obtained around 10 single phonemes for the single phoneme based VM and 100 diphones for diphone based VM. Note that the observed decrease of predicted performance of the identity based VMs (green and cyan lines) for low sensitivity voxels is a result of overfitting. Also, the suboptimal performance of the diphone based VM for very sensitive putative voxels (i.e. sensitive for < 10 diphones) is a result of the sparsity of that signal. Particular diphone combinations can occur with low frequency in 2 hours yielding poor estimations of the VM coefficients.

Finally, we investigated the number of phonemic combinations that could be recovered from our dataset. For this purpose, we generated fake data for three model voxels: one voxel was sensitive to the phoneme count, one voxel sensitive to phonemic combinations of increasing order (i.e. single phoneme, diphone, triphone, tetraphone, pentaphone, and hexaphone) and one voxel to the mixture of both. As in our first stimulation, the BOLD signal for the phonemic combination was obtained by randomly assigning weights of three features (i.e. three particular single phonemes, or diphones, etc). Here, taking into account the increasing number of features in each VM, the sampling was repeated 30 times for single phoneme simulation, 50 times for diphones, and 150 times for triphone, tetraphone, pentaphone and hexaphone based VMs to obtain the standard deviation of the prediction performance. As shown in Figure A.4, we were able to recover the VM coefficients and expected values for R^2 of single, diphone and triphone based VMs but unable to recover that of tetraphone, pentaphone and hexaphone based VMs. We also more directly assessed the effect of the data size by comparing the results we obtained for the diphone vs triphone based VM models for a single subject for which we had 5 (5 hours) scanning sessions of Moth Radio Hour stimulus. Although the additional predictive power of the triphone based VM relative to the diphone based VM increased slightly, the absolute magnitude of the relative predictive power of the two models, including the dominance of the diphone representation, remained (Figure A.15).

Therefore based on these three simulations and on the analysis of one single subject for which we had additional data, with around five hours of narrated speech stimulus (Fig-

ure A.15), our approach should allow us to identify voxels that show sensitivity for sets of phonemic units and differentiate brain regions that are sensitive to the identity of single phonemes, diphones or triphones as long as they exhibit sufficient sensitivity (approximately for fewer than 10 particular single phonemes and 100 particular diphones). Increasing the duration of the data collection might allow one to assess sensitivity for higher order combinations of phonemes including entire words, or to detect voxels that are sensitive to phoneme identity with lower sensitivity, or to recover voxels sensitive to diphones or triphones with very high sensitivity.

Data Availability

Raw fMRI data has been made available on https://gin.g-node.org/gallantlab/story_listening. All data other than anatomical brain images has been be shared, as there is concern that anatomical images could violate subject privacy. However, we have also provided matrices that map from volumetric data to cortical flatmaps for visualization purposes. Data generated in this study for each figure are provided in the source data in this repository as well.

Code Availability

Custom code used to reproduce the results is available at https://github.com/theunissenlab/phoneme_segmentation

Acknowledgments

We thank S. Popham and A. Nunez-Elizalde for helpful suggestions on data analysis. Funding: This work was supported by a Weil Foundation grant to FET, a Dingwall Foundation grant in Neurolinguistics to XG and NSF grant 1912373 to JLG.

Author Contributions

XLG and FET conceived and designed the experiments, analyzed the data and wrote the paper, FD and AH collected the data, and JLG and KJ contributed the materials and analysis tools.

Competing Interests

There is no competing interest.

Chapter 3

Cerebral cortical representation of natural language processing in English

3.1 Abstract

Reading comprehension requires extracting meaning from visual patterns. Using functional MRI and voxelwise encoding models, we investigated how visual, orthographic and semantic aspects of reading comprehension are represented across the cerebral cortex. We collected the BOLD activity when nine participants read over two hours of engaging natural narrative stories. Contrary to traditional views that postulate an isolated feature processing, our finding indicates that reading involves a complex, simultaneous processing of multiple information levels across the cerebral cortex. Furthermore, we also examined specific tuning of visual and orthographic features in the cerebral cortex. We found that as the cerebral cortex extracts meaning from visual stimuli, it increases the sensitivities to the temporal variations and integrates information over progressively larger spatial areas. Additionally, our results show that the human cerebral cortex represents English letters orthographically according to their curvature and intersections.

3.2 Introduction

The primary objective of the human visual system is object recognition, enabling effortless semantic segmentation of visual scenes. Previous research has revealed the brain's step-by-step process for object recognition. The primary visual cortex first extracts basic features such as edges, motion, and colors (Hubel & Wiesel, 1962). These features are then processed along the ventral stream, known as the "what pathway", into increasingly complex features like shapes and textures in higher cortical areas such as V2, V3, and V4 (reviewed in Bracci and Op de Beeck, 2023). The inferotemporal cortex further integrates these features into coherent object representations, specializing in identifying complex shapes and objects like faces and body parts (reviewed in DiCarlo et al., 2012; Grill-Spector and Weiner, 2014).

Furthermore, top-down processes allow higher-level information to influence object perception based on context, expectations, and prior knowledge (Bar et al., 2006; Firestone and Scholl, 2016).

Reading, a specialized form of object recognition, involves extracting meaning from written texts. The neural recycling hypothesis suggests that reading repurposes the existing cortical systems of the ventral pathway due to the relatively recent development of writing (Dehaene & Cohen, 2011). However, the specifics of how the brain processes reading comprehension remain less understood. The primary visual cortex's role in recognizing written words, particularly how it contributes to identifying basic and complex orthographic features, is still being explored. Research indicates that the visual word form area (VWFA) in the left ventral occipitotemporal cortex is crucial for the invariant recognition of written words (Cohen et al., 2000). Further complexities arise in how orthographic information is transformed into semantic and syntactic understandings. Previous studies show that recognized words are mapped onto phonological representations in the Wernicke's area (Booth et al., 2002) to be matched with their lexical semantic meaning (Price, 2012) and processed syntactically in Broca's area (Friederici, 2011; Friederici et al., 2017). However, the specific phonological and orthographic representations during reading remain unclear.

Moreover, as a hierarchically organized processing system, cerebral cortical representation of reading could principally contain multiple, mixed representations of various features. However, most neuroimaging studies assume that each brain area is dedicated to one specific computational step or level of representation and are designed to isolate a single computational step from the rest of the reading-processing hierarchy. In a typical experiment, responses are measured to stimuli that differ along a single dimension of interest, for example, familiar versus unknown alphabets (Baker et al., 2007; Szwed et al., 2014), uppercase versus lowercase (Dehaene et al., 2004). Then subtraction is used to find brain regions that respond significantly more to one end of this dimension than to the other end. Although this approach provides substantial power for testing specific hypotheses about letter/word representation, investigating the entire reading-processing hierarchy this way is inefficient because it inevitably requires many separate experiments. Furthermore, examining each computational step individually provides little information about relationships between representations at different levels in the reading-processing hierarchy.

Another limitation of most studies of reading comprehension is that they do not use natural stimuli. Neuroimaging experiments often use isolated words that are very different from natural language. It is well known that context impacts letter and word perception (McClelland & Rumelhart, 1981; C. Perfetti, 2007). Factors such as sentence intelligibility (Price, 2012) and attention (Ungerleider & G, 2000) influence brain activity even in primary visual areas, and these factors are likely to differ between non-natural stimuli and natural narrative reading. Because the reading-processing hierarchy likely contains important nonlinearities, it is unclear how well standard neuroimaging studies actually explain brain responses during reading comprehension.

To address these limitations, this study investigates the cerebral cortical processing of reading at different representation levels. Using natural narrative stories, we simultaneously

examine the representations from low-level visual and language statistics features, through intermediate phonemes and orthographic features, to high-level semantic and syntactic features. Then, we used voxelwise modeling (VM; Nishimoto et al., 2011, Huth et al., 2016, de Heer et al., 2017, Deniz et al., 2019 to determine which specific reading-related features are represented in each voxel. The stimulus was first nonlinearly transformed into eleven feature spaces that span much of the image-to-meaning pathway. Then, banded ridge regression (la Tour et al., 2022; Nunez-Elizalde et al., 2019) was used to determine which specific features are represented in each voxel for each subject. To examine the relationships between these representations, we used a variance-partitioning analysis. For each voxel, this analysis quantified the variance in each voxel that could be distinctly attributed to individual feature spaces, compared to that explained by combinations of feature spaces (de Heer et al., 2017; Gong et al., 2023; Lescroart & Gallant, 2019). Our results reveal how different features of reading comprehension are represented across the cerebral cortex.

3.3 Results

The goal of this study was to characterize how the human cerebral cortex represents different levels of information in reading in English. To do this, nine English speaking participants read two hours of narrative stories presented one word at a time while BOLD activity was recorded by functional MRI (Figure 3.1, Materials and Methods).

Cerebral cortical regions involved in natural language processing

Before examining specific representations of reading in the cerebral cortex, we aimed to identify the cerebral cortical regions involved across all stages of reading processing and quantify the total variance in BOLD response of each voxel involved.

To achieve this goal, we fit a linear encoding model to every cortical voxel in each participant's brain (Deniz et al., 2019; Gong et al., 2023; Huth et al., 2012, 2016; la Tour et al., 2022; Nishimoto et al., 2011). The linear encoding model consisted of eleven feature spaces capturing different aspects of statistical structure in text and processing required for reading it. Basic descriptive speech stimuli statistics features were created by counting the number of letters, words and phonemes of each time of repetition (TR) to account for the BOLD response from the stimuli length, as well as mere presence versus absence of the text stimuli. Phonemic stimulus features, including single phoneme, diphones and triphones feature spaces, were constructed by creating binary vectors encoding the occurrence of individual phonemes, diphones and triphones (Chapter 2, Gong et al., 2023). Visual stimulus features were constructed by passing each stimulus frame through a fixed set of nonlinear spatiotemporal motion-energy filters (Nishimoto et al., 2011) to capture the graphic information of each stimulus frame. Orthographic feature quantifies the similarity between letters or words. This is measured by rendering texts into images and then calculating the maximum pixel overlap for each pair of letter or word images. Semantic stimulus features were

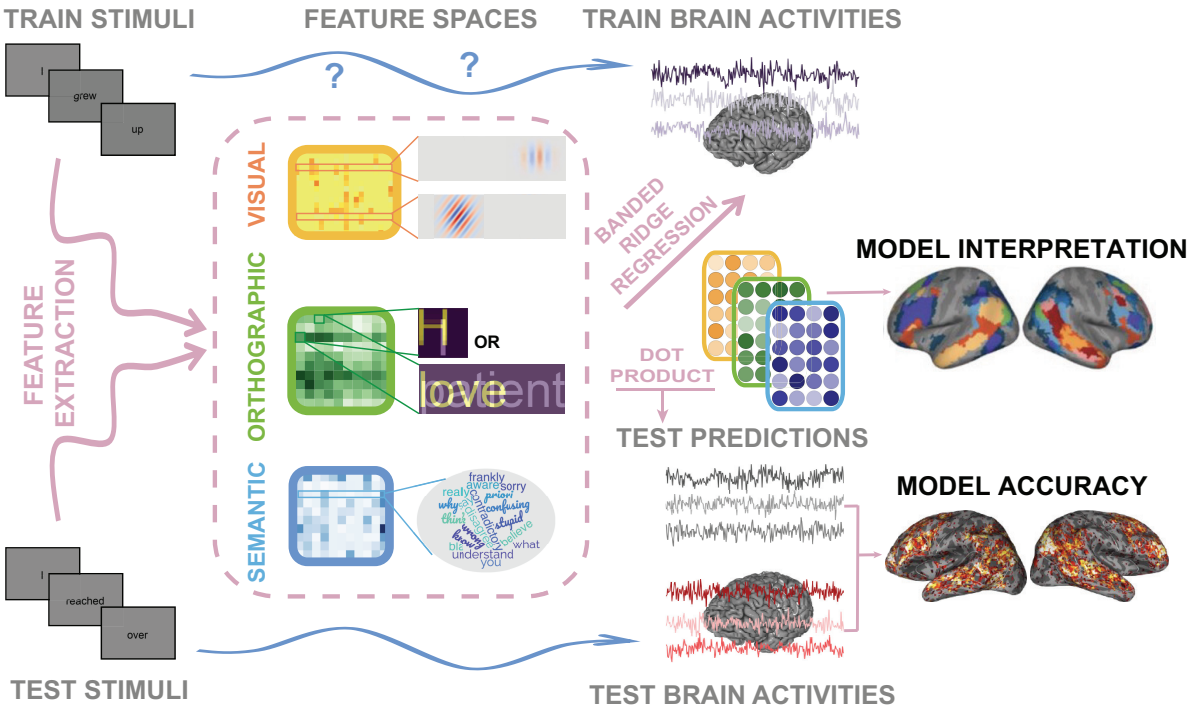


Figure 3.1: **Experimental procedure and voxelwise modeling (VM)** Eight participants read over two hours of natural narratives while BOLD responses were measured using fMRI. Visual stimulus features were constructed by passing each frame through a fixed set of nonlinear spatiotemporal motion-energy filters. Orthographic stimulus features were constructed by computing the maximum pixel overlap of all letter or word pairs. Semantic stimulus features were constructed by projecting each stimulus word into a 300-dimensional word embedding space (fastText). These stimulus features and BOLD responses were used to estimate a separate finite impulse response (FIR) ridge regression model for every voxel and participant. The estimated model weights describe the features that preferentially activates each voxel. These estimated model weights were used to predict BOLD responses for a separate held-out story that was not used for model estimation. Model accuracy was quantified as the coefficient of determination R^2 between predicted and recorded BOLD responses to the held-out narratives. Principal component analysis was performed on the model weights for interpretation.

constructed by projecting each stimulus word into a 300-dimensional word embedding space using fastText (Bojanowski et al., 2017) to capture the semantic content of the narrative stories. Syntactic features were created by constructing binary vectors encoding the identity of the part of speech tags of each word to capture their syntactic properties. The linearized encoding model was fit to the BOLD responses for each voxel and participant separately using a training set. Banded ridge regression (la Tour et al., 2022; Nunez-Elizalde et al., 2019) was used to fit all eleven feature spaces as a single joint model to the recorded BOLD responses in the training dataset. To validate that the model was not overfitted to the training dataset, the model was then used to predict the recorded BOLD responses in the hold-out test dataset. Model goodness of fit was assessed on the hold-out test dataset, and it was computed as the amount of variance in the recorded BOLD response explained by the model-predicted BOLD responses (the coefficient of determination R^2). In this study, this value is also referred to as the model prediction accuracy. To validate the robustness of our results and ensure generalization to new participants, the entire analysis was performed

for each participant separately, and data for two of the nine participants were not analyzed until the entire data analysis and interpretation pipeline was finalized.

Figure 3.2a shows the joint model prediction accuracy using all eleven feature spaces averaged across all nine participants, projected onto a standard brain (see Figure B.1 for individual participant results). Across all participants, model prediction accuracy is significantly high in many voxels located in bilateral visual cortex, superior temporal sulcus (STS), angular gyrus and supramarginal gyrus, inferior and medial parietal cortex, inferior and superior prefrontal cortex. The result indicates that extensive cortical regions are consistently involved in reading processing across subjects in the brain.

Major dimensions in cerebral cortical representations of natural language processing

To understand the important dimensions of reading processes across the human cerebral cortex, we estimated the contribution of each feature space to the joint model prediction accuracy, which is referred to as split prediction accuracy. This measure takes into account both the direct effect of the feature space, as well as its interaction with other feature spaces in the model.

Figure 3.2b shows the contribution of individual feature space averaged across all nine participants to the joint model prediction accuracy, projected to a standard fsAverage brain (Figure B.2 for individual participants). In general, the visual, orthographic and semantic feature spaces show significant prediction accuracy in many voxels in different cortical regions, while language statistics, phonemes and syntax feature spaces show significant prediction accuracy in very limited voxels. In specific, visual feature space best predicts voxels in the bilateral occipital cortex, superior temporal sulcus (STS), intraparietal sulcus (IPS) and sPMv. Orthographic feature space best predicts bilateral occipital, temporal and prefrontal cortices and IPS. Semantic feature space best predicts voxels in the bilateral temporal, parietal and prefrontal cortices. Statistics feature spaces predicted very limited voxels in the occipital cortex and STS. Phoneme feature spaces predicted very limited voxels along STS. Syntax feature space does not show significant prediction accuracy. This result indicates that visual, orthographic and semantic feature spaces are the three major dimensions in representing the natural language in the cerebral cortex.

To quantify the cortical regions involved in reading processes, we calculated the averaged prediction accuracy from the joint model across twenty-four regions of interests (ROIs) spanning the entire cerebral cortex. Figure 3.2c shows that all these twenty four regions are significantly involved in natural language processing. In particular, more than 5% of the BOLD responses of voxels in V1, V3, V5, LO1, pITG, STG can be significantly explained by the joint model. Additionally, we analyzed the specific contribution of the three major dimensions in each ROI. The visual features (Figure 3.2c) significantly predict more accuracy than orthographic or semantic features in V1, V3A, V3B, IPS0, IPS1, and IPS2. Visual and orthographic features predict similarly in V2 and V3. Orthographic features significantly

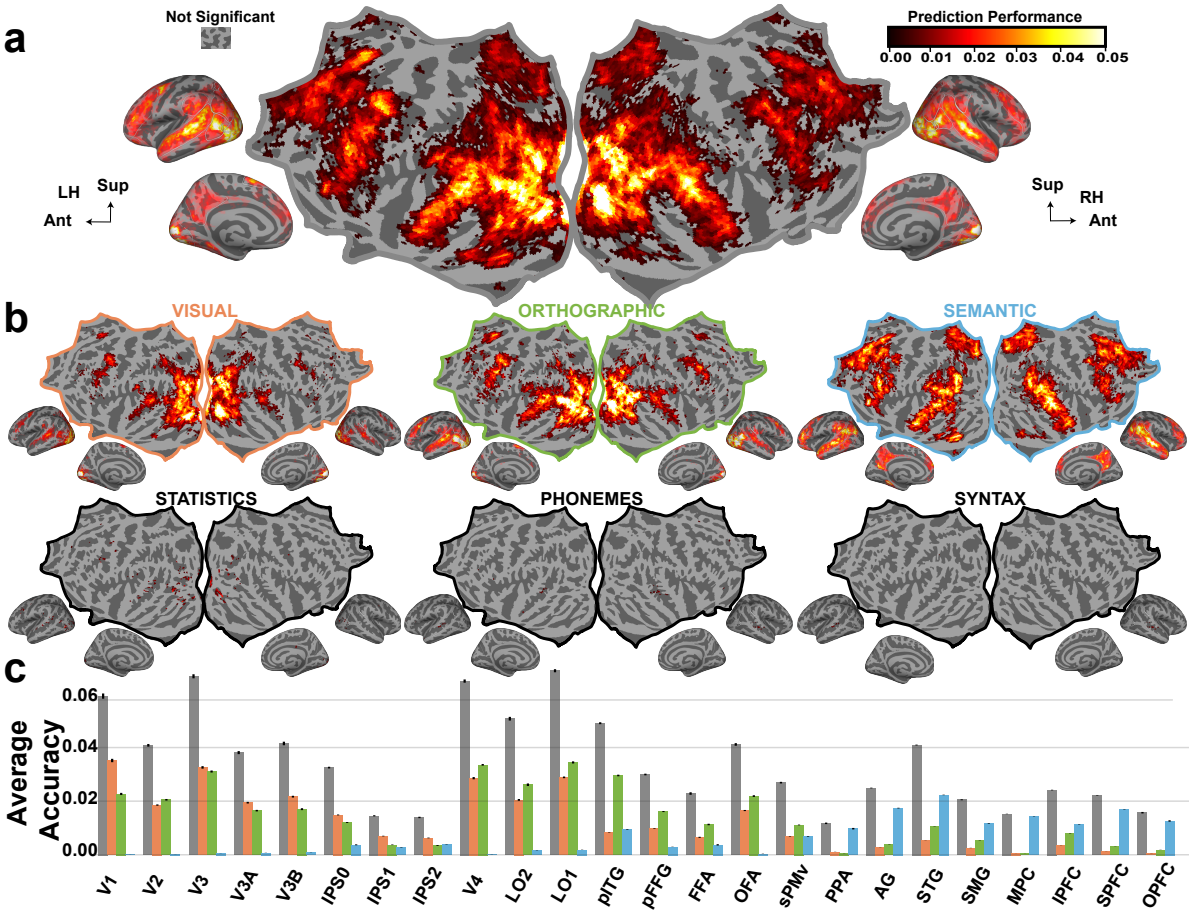


Figure 3.2: **Joint model prediction accuracy across the cerebral cortical surface.** To understand the cortical representation of reading, a joint voxelwise encoding model consisting of eleven feature spaces was constructed. **a** Group-averaged joint model prediction accuracy is shown on the flattened cerebral cortical surface of the fsAverage brain. Inflated cortices are shown for reference. Only vertices with significant accuracy were shown. Prediction accuracy is given by the color scale at the top right. Vertices for which the joint model prediction accuracy is not statistically significant are left transparent showing the gray curvature map. Averaged across participants, vertices in bilateral occipital, temporal, parietal, and prefrontal cortices are significantly predicted. **b** The contribution of individual feature space to the joint model prediction accuracy is shown on the flattened cerebral cortical surface of fsAverage brain. Visual feature space best predicts voxels in the bilateral occipital cortex, superior temporal sulcus (STS), intraparietal sulcus (IPS) and sPMv. Orthographic feature space best predicts bilateral occipital, temporal and prefrontal cortices and IPS. Semantic feature space best predicts voxels in the bilateral temporal, parietal and prefrontal cortices. Statistics feature spaces predicted very limited voxels in the occipital cortex and STS. Phoneme feature spaces predicted very limited voxels along STS. Syntax feature space does not show significant prediction accuracy. This result indicates that visual, orthographic and semantic feature spaces are the main features capturing the cerebral cortical representation of natural language processing. **c** The average prediction accuracy by ROIs for the joint model and the contribution of visual, orthographic and semantic feature space is shown in the bar plot. Visual features (orange) significantly predict more accuracy than orthographic or semantic features in V1, V3A, V3B, IPS0, IPS1, and IPS2. Visual and orthographic features predict similarly in V2 and V3. Orthographic features (green) significantly predict the most in V4, LO1, LO2, pITG, pFFG, FFA, OFA, and sPMv. Semantic features (blue) significantly predict the most in PPA, AG, STG, MPC, IPFC, SPFC, and OPFC. Gray bars show the joint model prediction performance.

predict the most in V4, LO1, LO2, pITG, pFFG, FFA, OFA, and sPMv. Semantic features significantly predict the most in PPA, AG, STG, MPC, IPFC, SPFC, and OPFC. This detailed map indicates how different cortical areas variably represent visual, orthographic and semantic aspects of reading processes.

Cerebral cortical representation of visual, orthographic and semantic information of natural language processing

To further compare the similarities and differences in the cortical representation of the three major dimensions, the split prediction accuracy of visual (red), orthographic (green), semantic (blue) feature spaces are summarized across participants, projected to one standard fsAverage brain (Figure 3.3a). The split prediction accuracy map in Figure 3.3a shows that many voxels of yellow, teal, magenta and white, suggesting that these voxels can be significantly predicted by more than one feature space. It indicates that the cerebral cortical representation of visual, orthographic and semantic features of natural language are not completely independent from each other. Instead, many cortical regions are representing combinations of these features together.

In order to test this hypothesis and quantify the feature representations in each voxel, we used variance partitioning to estimate the fraction of variance explained by visual, orthographic and semantic feature space individually and the fraction that might be equally well explained by any combination of these three feature spaces (de Heer et al., 2017; Gong et al., 2023; Lescroart & Gallant, 2019). For this purpose, we first estimated the BOLD residuals by removing the BOLD responses that could be explained by the joint model using language statistics, phoneme and syntax feature spaces. Afterwards, we fit a joint model using visual, orthographic and semantic feature spaces on the BOLD residual. We used banded ridge regression with the optimal hyperparameter selected for each voxel (la Tour et al., 2022). We fixed the hyperparameters for each voxel and refit the model with all the possible combinations of feature spaces as follows: three models based on a single feature space (visual, orthographic and semantic); three models based on paris of feature spaces (visual-orthographic, visual-semantic, orthographic-semantic); and a single model that used all three feature spaces together. Then, using set theory, we divided the variance explained by these feature spaces into the following seven partitions: the variance explained uniquely by each feature space; the variance explained jointly by each pair of feature spaces but excluding the third; and the variance explained jointly by all three feature spaces (Materials and Methods). With this approach, we were able to quantify the extent to which any particular voxel represented features from each of visual, orthographic and semantic feature spaces, after taking into account its selectivity to other feature spaces.

Figure 3.3b shows the partial variance explained for each of the seven partitions, projected to the cortical flatmap of a standard fsAverage brain. Three feature spaces jointly explained variance in bilateral occipital cortex, and superior temporal sulcus (STS) and sPMv. Unique contribution from visual feature space spans much of the bilateral visual

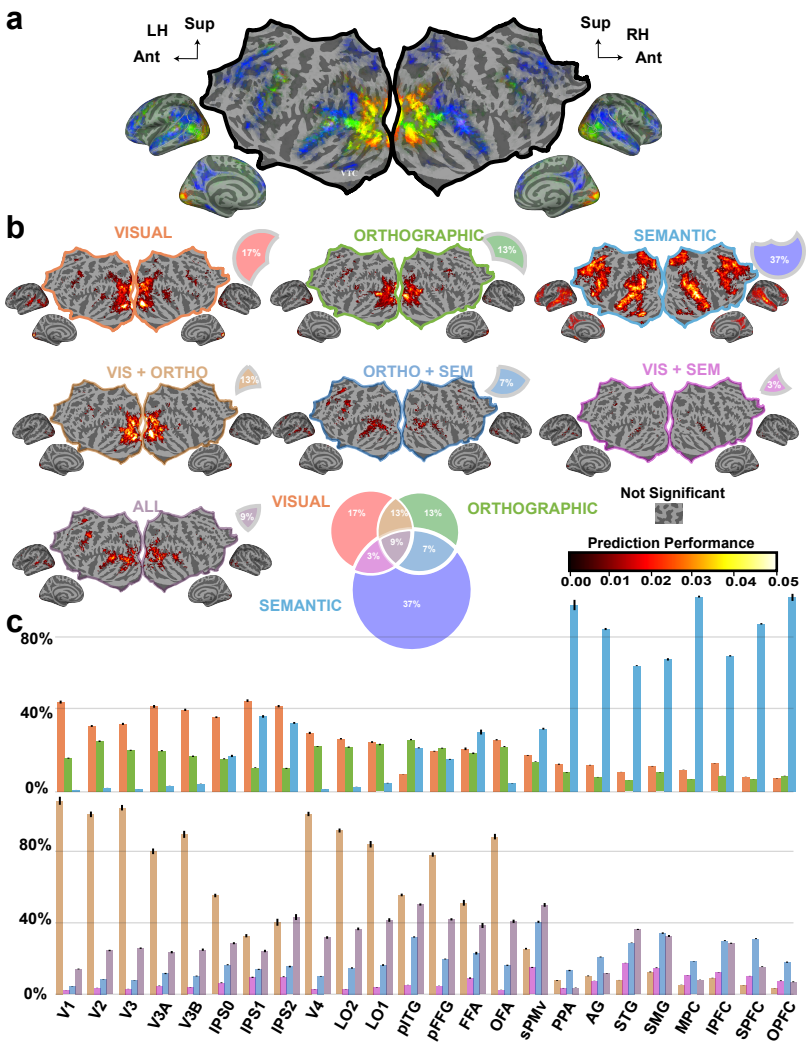


Figure 3.3: **Variance partitioning.** To compare the cortical representation of visual, orthographic and semantic information in details, variance partitioning was used to estimate the fraction of variance explained each feature space individually and the fraction that might be equally well explained by any combination of these three feature spaces. (**a**) Group-averaged prediction accuracy is shown for visual (orange), orthographic (green) and semantic (blue) feature spaces on the flattened cortical surface of the fsAverage brain. Only vertices with significant accuracy were shown. Across participants, visual features best explain response variance of vertices in the low and intermediate visual cortex, orthographic features best explain response variance of voxels in intermediate and higher visual cortex and limited parietal and prefrontal cortices, and the semantic features best explain variance of vertices in the bilateral temporal, parietal and prefrontal cortices. In addition, many vertices are in yellow, teal, magenta and white, suggesting that these vertices can be significantly predicted by more than one feature space. (**b**) A variance-partitioning analysis was used to separate the variance explained by the visual, orthographic, and semantic feature spaces into seven partitions. Venn diagram shows the percentage of the variance significantly explained by each partition across all subjects in the entire cortex. The flatmaps show variance explained for each of the seven partitions on the group averaged subject. (**c**) The average prediction accuracy by ROIs for each of the seven partitions is shown in the bar plot. The upper panel shows that visual features (orange) uniquely explain significantly more accuracy than orthographic or semantic features in V1, V2, V3, V3A, V3B, IPS0, IPS1, IPS2, LO2 and OFA. Visual and orthographic features predict similarly in LO1 and FFA. Orthographic features (green) significantly predict the most in pITG, and pFFG. Semantic features (blue) significantly predict the most in FFA, sPMv, PPA, AG, STG, MPC, IPFC, SPFC, and OPFC. The lower panel shows that three feature spaces jointly (purple) explained significantly more accuracy than any other combinations in sPMv and STG. The contributions from visual-orthographic (brown) intersection explained significantly more accuracy than any other combinations in V1, V2, V3, V3A, V3B, IPS0, IPS1, V4, LO1, LO2, pITG, pFFG, and FFA. The contribution from orthographic-semantic intersection (teal) explained significantly the most in PPA, AG, MPC, SPFC, and OPFC. The contribution from visual-semantic intersection (magenta) did not explain significantly more accuracy than any other feature combinations. In IPS2, there is no significant difference in the prediction accuracy of the contribution from the three joint features and the contribution from the visual-orthographic intersection. In SMG and IPFC, there is no significant difference in the prediction accuracy of the contribution from the three joint features and the contribution from the orthographic-semantic intersection.

cortex (including anterior visual areas, and intraparietal sulcus), superior temporal sulcus (STS), inferior parietal lobule, and left sPMv. The contributions from orthographic feature and visual-orthographic intersection were mostly in the bilateral visual cortex and limited voxels in the left inferior parietal lobule. The semantic feature space explained a large fraction of the unique variance in bilateral inferior temporal, inferior and medial parietal, inferior and superior prefrontal cortices. The contributions from orthographic-semantic intersection were mostly found in bilateral visual cortex, STS and left sPMv. The contributions from visual-semantic intersections are very limited voxels in intraparietal sulcus and superior STS. The Venn diagram in figure quantifies the magnitude of these observations by showing the percentage of BOLD response variance explained by each partition averaged across the entire cortex. Visual feature space uniquely explained 17% variance of the cerebral cortical BOLD responses, orthographic feature space uniquely explained 13% variance, and the semantic feature space uniquely explained 37% variance. Forty-three percent of the BOLD response variances are explained by any of the combinations of these three feature spaces. This result suggests that during natural language processing, the human cerebral cortex does not segregate visual, orthographic and semantic information clearly into distinct cortical areas. Instead, different cortical regions have a focus on representing one level of information, while simultaneously engaging in representing other levels of information.

To further quantify these observations, we calculated the averaged prediction accuracy from each of the seven partitions in twenty-four ROIs spanning the entire cerebral cortex. The upper panel in Figure 3.3c shows the unique contribution from the three major dimensions. Visual features (orange) uniquely explained significantly more accuracy than orthographic or semantic features in V1, V2, V3, V3A, V3B, IPS0, IPS1, IPS2, LO2 and OFA. Visual and orthographic features predict similarly in LO1 and FFA. Orthographic features (green) significantly predict the most in pITG, and pFFG. Semantic features (blue) significantly predict the most in FFA, sPMv, PPA, AG, STG, MPC, IPFC, SPFC, and OPFC. This result further strengthens the previous conclusion that visual, orthographic and semantic features are the three major dimensions in cerebral cortical representations of natural language processing, and they are represented differently in different cortical regions. In addition, the bottom panel in Figure 3.3c shows the contribution from the combinations of three major dimensions. Three feature spaces jointly (purple) explained significantly more accuracy than any other combinations in sPMv and STG. The contributions from visual-orthographic intersection (brown) explained significantly more accuracy than any other combinations in V1, V2, V3, V3A, V3B, IPS0, IPS1, V4, LO1, LO2, pITG, pFFG, and FFA. The contributions from orthographic-semantic intersection (teal) explained significantly the most in PPA, AG, MPC, SPFC, and OPFC. The contribution from visual-semantic intersection (magenta) did not explain significantly more accuracy than any other feature combinations. In IPS2, there is no significant difference in the prediction accuracy of the contribution from the three joint features and the contribution from the visual-orthographic intersection. In SMG and IPFC, there is no significant difference in the prediction accuracy of the contribution from the three joint features and the contribution from the orthographic-semantic intersection. This result suggests that cerebral cortical representation of natural language does not rely on straight-

forward, independent or sequential processing. Instead, it is highly parallel and complex. It is necessary to simultaneously represent diverse combinations of visual, orthographic, and semantic information across distinct cortical areas. Moreover, this process also requires the concurrent integration of information at all levels.

Visual feature tuning of natural language processing

To determine how specific visual properties are represented in the human cerebral cortex during natural language processing, we looked at the direction, position, spatial frequency and temporal frequency tuning of the voxels that are significantly predicted by visual features. These voxels are referred to as visually selective voxels. The visual feature tuning of each voxel is given by its 2772-dimensional vector of estimated visual feature model weights, one weight value for each of 2772 spatio-temporal Gabor filters. The filters are created at multiple spatial and temporal frequencies, directions of motion, x-y positions and sizes.

Figure 3.4a-d shows the projection of the optimal direction, position, spatial frequency, and temporal frequency tuning of each voxel averaged across all twelve participants onto a flattened map of the fsAvg cortical surface. Figure 3.4a shows that the voxels in the visual cortex (V1, V2, V3, V3A, V3B, V4, IPS0, IPS1, IPS2, LO1, LO2, FFA, and OFA) are mostly tuned to directions between 90 and 140 degrees, while voxels in the language related areas (pITG, STG, SMG, AG, sPMv, IPFC, and SPFC) are mostly tuned to directions between 260 and 310 degrees. It indicates that the beginning and end of visual stimuli (words) are critical for cerebral critical processing of reading. Figure 3.4b shows that visually selective voxels are tuned to the center of the screen. This result confirms that subjects followed the instructions well to fix at the center of the screen during the reading task. Figure 3.4c shows the optimal spatial frequency eliciting the BOLD responses in the visually selective voxels. Voxels in the primary visual cortex (V1) are tuned to high spatial frequency between 7 and 8 cycles per image. Voxels in the intermediate visual cortex, including V2, V3, V3A, V3B, V4, IPS0, LO1, LO2 are turned to lower spatial frequency between 5 and 6 cycles per image. Voxels in higher visual cortex, including FFA and OFA, and language related areas pITG, pFFA, AG, STG, and SMG are tuned to even lower spatial frequency between 3 and 4 cycles per image. Voxels in sPMv are tuned to very low spatial frequency between 1 and 3 cycles per image. This results reveals a gradient of decreasing spatial frequency tuning gradient spanning from primary visual cortex to the prefrontal cortex. This gradient indicates that when the cerebral cortex extracts meaning from visual stimuli, it integrates information over progressively larger spatial areas. Figure 3.4d shows the optimal temporal frequency eliciting the BOLD responses in the visually selective voxels. Voxels in the early visual cortex, including V1, V2, V3, are tuned to temporal frequency lower than 1.5Hz. Voxels in the intermediate visual cortex, including V3A, V3B, V4, IPS0, LO1, LO2 are turned to higher temporal frequency between 1.5 and 2.5Hz. Voxels in higher visual cortex, including FFA and OFA, and language related areas pITG, pFFA, AG, STG, SMG, sPMv are tuned to even higher spatial frequency between 2.5 to 4Hz. This results reveals a gradient of increasing temporal frequency tuning gradient spanning from primary visual cortex to the

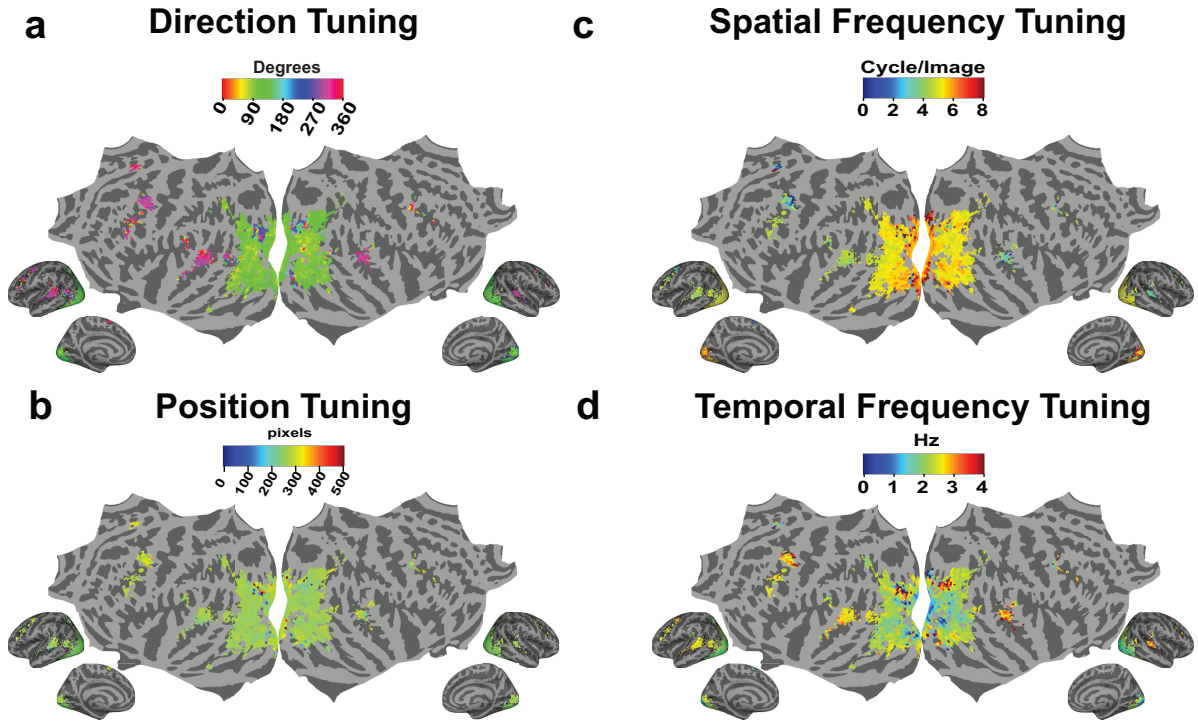


Figure 3.4: **Visual feature tuning.** To understand the selectivity of visual properties for each visually selective vertex, the visual tuning for each vertex is examined. Group-averaged optimal direction, position, spatial frequency and temporal frequency tuning of each vertex are shown on the flattened cortical surface of the fsAverage brain. (**a**) Voxels in the visual cortex are mostly tuned to directions between 90 and 140 degrees, while voxels in the language related areas are mostly tuned to directions between 260 and 310 degrees. (**b**) Voxels are tuned to the center of the screen. (**c**) Voxels in V1 are tuned to spatial frequency between 7 and 8 cycles per image. Voxels in V2, V3, V3A, V3B, V4, IPS0, LO1, LO2 are turned to spatial frequency between 5 and 6 cycles per image. Voxels in FFA and OFA, pITG, pFFA, AG, STG, and SMG are tuned to spatial frequency between 3 and 4 cycles per image. Voxels in sPMv are tuned to spatial frequency between 1 and 3 cycles per image. (**d**) Voxels in V1, V2, V3, are tuned to temporal frequency lower than 1.5 Hz. Voxels in V3A, V3B, V4, IPS0, LO1, LO2 are turned to temporal frequency between 1.5 and 2.5 Hz. Voxels in FFA and OFA, pITG, pFFA, AG, STG, SMG, sPMv are tuned to spatial frequency between 2.5 to 4 Hz.

prefrontal cortex. This gradient indicates that when the cerebral cortex extracts meaning from visual stimuli, it increases the sensitivities to the temporal variations.

Orthographic feature tuning of natural language processing

To determine how the human cerebral cortex represents the orthographic information of natural language processing, we looked at the letter and word orthographic tuning of the voxels that are significantly predicted by orthographic features. These voxels are referred to as orthographically selective voxels. The orthographic feature captures the orthographic similarities of all English letters and words in the stimuli by calculating the maximum pixel overlay of all letter (52) and word (2978) pairs. For letter orthographic features, the tuning of each voxel is given by its 52-dimensional vector of estimated model weights, one weight value for each of 52 letter pairs. For the word orthographic features, the tuning of each voxel

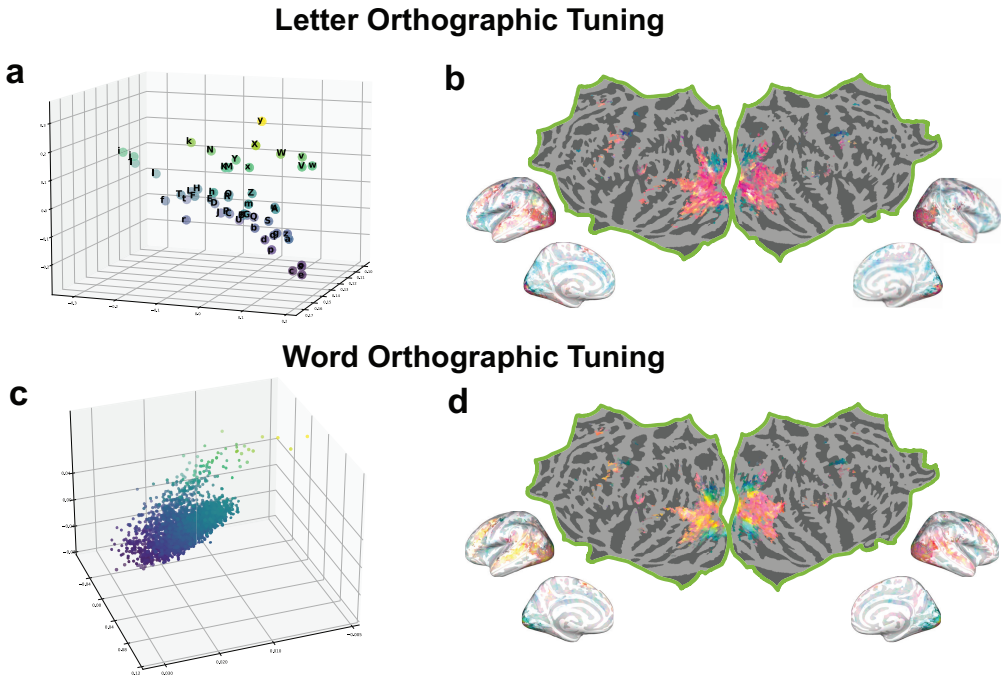


Figure 3.5: **Orthographic feature tuning.** The orthographic feature captures the orthographic similarities of all English letters and words in the stimuli by calculating the maximum pixel overlap of all letter (52) and word (2978) pairs. Principal component analysis (PCA) was used to determine the main dimensions of the voxelwise letter/word orthographic feature tuning. (**a**) Fifty-two English letters are projected onto the first three principal dimensions of the group-averaged voxelwise model weights of letter orthographic feature space. Letters are colored according to the projected values onto the first and third principal components. The largest axis of variation lies roughly along the diagonal of first and second principal dimensions, and separates linear letters (i, j, l) from curved letters (a, e, o). The second largest axis of variation lies roughly along the third principal dimension, and separates letters with intersections (x, y, w) from the rest. (**b**) Group-averaged voxelwise model weights for letter orthographic feature space were projected onto its top three principal components. This projection is shown on the flattened cortical surface of the fsAverage brain. Each vertex is colored according to its projection onto the first (red), second (blue) or third (green) letter orthographic principal components (PC). Voxels in the central visual cortex and superior sPMv are tuned more to the linear and curved letters, while voxels in the peripheral visual cortex and inferior sPMv are tuned more to letters with intersections. (**c**) All English words from the stimuli (2978) are projected onto the first three dimensions of the word orthographic feature space. Words are spanning diagonally across this three dimensional space. (**d**) Group-averaged voxelwise model weights for word orthographic feature space were projected onto its top three principal components. This projection is shown on the flattened cortical surface of the fsAverage brain. Each vertex is colored according to its projection onto the first (red), second (blue) or third (green) word orthographic principal components (PC).

is given by its 2978-dimensional vector of estimated model weights, one weight value for each of 2978 word pairs. Principal component analysis (PCA) was used to determine the main dimensions of the voxelwise letter and word orthographic feature tuning.

Figure 3.5a shows that 52 English letters (26 upper case and 26 lower case) are projected onto the first three principal dimensions of the group-averaged voxelwise model weights of the letter orthographic feature space. The largest axis of variation lies roughly along the diagonal of first and second principal dimensions, and separates linear letters (i, j, l)

from curved letters (a, c, s). The second largest axis of variation lies roughly along the third principal dimension, and separates letters with intersections (x, y, w) from the letters without intersections (o, c, s). This result indicates that the human cerebral cortex represents English letters orthographically according to their curvature and intersections. To visualize the pattern of orthographical selectivity across the cerebral cortex, we project the group-averaged voxelwise model weights for letter orthographic feature space onto the corresponding top three principal components (figure 5b). A red voxel produces greater BOLD responses to the curved letters, yellow voxels to linear letters, and blue voxels to letters with curvatures. This visualization suggests that in the visual cortex, orthographically selected voxels in the central areas are tuned more to the linear and curved letters, while voxels in the peripheral cortical regions are tuned more to letters with intersections. In the sPMv region, orthographically selected voxels in the inferior areas are tuned more to the linear and curved letters, while voxels in the superior areas are tuned more to letters with intersections.

This analysis was extended to explore the cortical representation of word orthographic information. Figure 3.5c shows the projection of all English words used in the stimuli (2978) onto the first three dimensions of the word orthographic feature space. It suggests that the human cerebral cortex represents English words in a low dimensional space during natural language processing. Figure 3.5d presents the projection of the group-averaged voxelwise model weights for word orthographic feature space onto its top three principal components, providing a visualization of how the orthographic information of English words are represented across different cortical regions. This visualization indicates that the orthographic information presented in the central visual cortical regions is different from the peripheral visual cortical regions. However, the interpretation of these projections and their implications for natural language processing requires further analysis and elucidation.

3.4 Discussion

This study examined how different levels of information of natural language processing are represented across the human cerebral cortex. To explore this issue, we recorded BOLD activity elicited by natural narrative stories. Firstly, we build a voxelwise encoding model based on eleven distinct feature spaces based on language statistical, visual, phonemic, orthographic, semantic and syntactic information of reading. We found that visual, orthographic and semantic feature spaces are the three major dimensions in representing the natural language in the cerebral cortex. Also, visual, orthographic and semantic features of language processing are represented differently in different cerebral cortical areas. Secondly, because visual, orthographic and semantic features can be highly correlated in natural language, we used a variance-partitioning analysis to determine how much of the variance was uniquely predicted by each feature space (de Heer et al., 2017; Gong et al., 2023; Lescroart & Gallant, 2019). We found that during natural language processing, the human cerebral cortex does not segregate visual, orthographic and semantic information clearly into distinct areas. Different cortical regions have a focus on representing one level of information, while

simultaneously engaging in representing other levels of information. Lastly, we examined specific tuning of visual and orthographic features in the cerebral cortex. We found that when the cerebral cortex extracts meaning from visual stimuli, it increases the sensitivities to the temporal variations and integrates information over progressively larger spatial areas. Furthermore, the human cerebral cortex represents English letters orthographically according to their curvature and intersections.

One key finding from this study is that cortical areas, including V4, LO1, LO2, pITG, pFFG, FFA, OFA, and sPMv are involved in orthographic representation and processing of written texts during reading. After carefully partitioning out the contributions from the correlations between orthographic, visual and semantic information, we identified the posterior inferior temporal gyrus (pITG) and posterior fusiform gyrus (pFFG) as being specifically dedicated to orthographic representation and processing of written texts in reading natural narratives. This result makes sense and is in congruent with previous studies of the human ventral visual pathway. Previous studies show that V4 is a midtier visual cortical area in the ventral visual pathway that exhibits diverse receptive field preferences, particularly to shape of orientation and curvature (reviewed in Roe et al., 2012). Lateral occipital complex, including LO1 and LO2, exhibited selectivity toward convex and concave shape (Decramer et al., 2019; Larsson & Heeger, 2006; Silson et al., 2013). These curve form tunings are critical for recognizing letters and words. In addition, prior neuroimaging studies show that the visual word form area (VWFA), located between left fusiform and inferior temporal gyri, has been considered to play a pivotal role in processing orthographic information of recognizing written words and letters (Dehaene et al., 2002; McCandliss et al., 2003; Price, 2012). Clinical studies also show that lesions in the posterior part of the inferior longitudinal fasciculus (ILF), which is connected to the pITG, have been associated with visual agnosia and alexia (Herbet et al., 2018). Furthermore, the fusiform face area (FFA) and occipital face area (OFA) are crucial for face perception and has also been reported to be involved in visual form perception (Zhan et al., 2023). Previous research has emphasized the similarities between word and face recognition, both of which require foveal processing and elicit category-specific activation (Hasson et al., 2002; Puce et al., 1996). Lastly, sPMv is shown to be involved in visual form perception and reading by its key role in multisensory integration, and in creation and storage of motor vocabularies for speech and manual features (Callan et al., 2014; Hickok, 2012).

Another important finding of this study is that the cerebral cortical representation of natural language does not simply rely on independent or sequential processing. Instead, it is highly parallel and complex. Visual, orthographic, and semantic information and the diverse combinations of these information are represented simultaneously across distinct cortical areas. In contrast, most prior studies trying to understand orthographic representation of reading comprehension in the human cerebral cortex focused on limited regions, mainly VWFA, and failed to identify other cortical regions involved in this process. In addition, the complex and intertwined nature of visual, orthographic and semantic representation in reading comprehension was not emphasized. These neglects could be caused by three methodological differences. First, we did not smooth our data in this study. We analyzed

data for each participant in their native brain space, and we did not spatially smooth the data across voxels. In contrast, almost all prior studies transformed their data into a standard brain space and spatially smoothed the data across voxels before analysis. These transformation and smoothing procedures can incorrectly assign signal to voxels and average away meaningful signal and individual variability. Thus, past studies may have failed to identify the additional brain regions we identified in this study because the signal in these regions was averaged out. Second, we collected two hours of fMRI data per participant for six participants. In contrast, prior studies collected a small amount of fMRI data (less than an hour) per participant from many participants. Because fMRI data is noisy, these studies transformed their data into a common brain space (to average across participants) and/or spatially smoothed their data to observe the effects of interest. However, as discussed in the previous point, smoothing fMRI data can average out the effects of interest. Finally, our experiment uses engaging natural narratives with rich contextual information. In contrast, participants in most prior studies were essentially performing letter or word identification, instead of reading comprehension.

Moreover, previous neuroimaging studies has uncovered a host of functional features of the VWFA, which is tuned to familiar vs. unknown alphabets (Baker et al., 2007; Dehaene et al., 2010), partial invariance for retinal location (Dehaene-Lambertz et al., 2018; Szwed et al., 2014), invariance for uppercase/ lowercase (Hasson et al., 2002; Long et al., 2018), or sensitivity to the frequency of word occurrence (Bouhali et al., 2014; Saygin et al., 2016). However, they were not able to uncover the main dimensions that the letters are organized and encoded. Therefore, the third important finding of this study is that the human cerebral cortex represents English letters orthographically according to their curvature and intersections. This makes sense because curvature is the key tuning feature in V4, LO1 and LO2 of the human ventral visual pathway (Lawrence et al., 2023; reviewed in Kourtzi and Connor, 2011). Also, intersection is a critical feature in edge detection and figure/ground segmentation for object recognition in the brain (Hesse & Tsao, 2016; Ko & von der Heydt, 2018; Zhou et al., 2000). This evidence supports the neuronal recycling hypothesis that the brain reuses pre-existing neural circuits involved in face and object recognition to process written letters and words for reading comprehension (Dehaene, 2005).

One surprising finding is that visual, orthographic and semantic feature spaces predicted responses about equally well in both cortical hemispheres. This contradicts with many previous studies suggesting that reading-related signals are lateralized at different stages of reading comprehension (Toga & Thompson, 2003; Wu et al., 2022). One potential explanation for this discrepancy is that we used narrative stories rather than isolated letters, words, phrases, sentences or other reading stimuli without contextual information. At a high level of linguistic processing, narrative comprehension is both engaging and demanding, and it likely recruits brain circuits that are simply not required to perform simpler tasks. And indeed, earlier studies that use narrative speech and stories also reported that evoked brain activities were bilateral (Deniz et al., 2019; Huth et al., 2016). In addition, previous studies also show that the processing of low-level text-like but nonlinguistic features is bilateral (Behrmann & Plaut, 2014; L. Chen et al., 2019). Second, the fact that all three feature spaces predicted

activities in both hemispheres does not imply that the same specific features are represented in the two hemispheres. A more detailed analysis of the voxel selectivity/tuning in each feature space might provide more insights concerning lateralizations.

Another very surprising result that should be emphasized is the absence of phonemic representation in this reading comprehension experiment. Traditionally, phonemic representation is considered a pivotal component in auditory processing (Chapter 2, Gong et al., 2023). This discrepancy suggests that the reading process may bypass the phonemic segmentation stage, which is essential in auditory perception. This observation aligns with the Dual Route Cascaded model (DCR, Coltheart et al., 2001), where reading transitions directly from visual stimuli to word recognition and semantic interpretation, omitting the intermediate phonemic parsing typically observed in speech processing. Consequently, these findings support a view of reading that is distinct and more streamlined than auditory processing, operating without the need for segmenting phonemic units within words (Grainger & Ziegler, 2011). Such a model challenges traditional theories, such as the activation-verification model, which claims that "a word's phonology is the initial and primary code by which a word accesses its representation in the internal lexicon" and serves as an early source of constraint on the visual recognition of words (Lukatela & Turvey, 1994; Van Orden, 1987).

In sum, our naturalistic experiment and VM framework provides a powerful and efficient method for identifying cortical regions involved in visual, orthographic and semantic representations in individual human brains. However, there are still several questions open to be explored. First, to explore how English words are organized orthographically in the cerebral cortex, we used principal component analysis (PCA) to project the high dimensional model weights from word orthographic features to a low dimensional 3D space. However, we were unable to obtain a meaningful interpretation of these dimensions. In addition, to explore the phonological representations of reading words in the cerebral cortex, we fitted a phonemic model using features including the identities of single phoneme, diphones, triphones and the phoneme rate. However, this model only produced significant accuracy in very limited voxels in STS after controlling the correlations between phonemic features with all the other features extracted from visual stimuli. Therefore, we were unable to explore the relationship between phonological representation and orthographic representation in reading.

3.5 Materials and Methods

Participants

Structural and functional brain data were collected from twelve English speakers: S1 (male, age 31), S2 (male, age 31), S3 (female, age 28), S4 (male, age 30), S5 (male, age 25), and S6 (male, age 36), S7 (female, age 24), S8 (male, age 24), S9 (female, age 29), S10 (male, age 25), S11 (female, age 25), S12 (male, age 25). All participants were healthy and had normal or corrected-to-normal vision. Each subject's handedness was evaluated by the Edinburgh handedness inventory (Oldfield, 1971). The use of human subjects in this study

was approved by the UC Berkeley Committee for the Protection of Human Subjects. A written statement of informed consent has been obtained from each subject.

Stimuli

The stimuli in these experiments consisted of eleven 10− to 15− min stories that were pre-recorded from the Public Radio Exchange (PRX) radio show "The Moth Radio Hour", which has been used in previous studies of our lab (C. Chen et al., 2024; de Heer et al., 2017; Deniz et al., 2019; Gong et al., 2023; Huth et al., 2016). They are engaging stories performed verbally in English that capture the attention of the subjects. The audio recording of each story was manually transcribed, and the written transcription was aligned to the audio recording (see (Deniz et al., 2019) for details). The stimuli were split into separate model estimation and model validation sets. The model estimation stimulus-set consisted of ten 10- to 15-min stories played once each. The length of each scan was tailored to the story and also included 10 seconds of silence both before the story and 10 seconds after the story. Each subject read the same 10 stories. The model validation stimulus-set consisted of a single 10-min story told that was read four times for each subject in each language in order to estimate voxel response reliability and noise ceiling. This resulted in 3737 time points (sampled at TR) for the training dataset and 291 time points for the validation dataset.

Stimulus presentation format

The words of each story were presented one-by-one at the center of the screen using a Rapid Serial Visual Presentation (RSVP) procedure (Buchweitz et al., 2009; Forster, 1970). Each word was presented for a duration equal to the duration of that word in the spoken version of the story.

The Python pygame library was used to display text on a gray background at 34 horizontal, and 27 vertical degrees of visual angle. Black letters were used to present the words. For English stimuli, words were presented at average 6 (min = 1, max = 16) horizontal and 3 vertical degrees of visual angle. A white fixation cross was present at the center of the display throughout the experiment. Participants were instructed to fixate on a center cross while reading the story. Participants' eye movements were monitored at 60 Hz throughout the scanning sessions using a custom-built camera system equipped with an infrared source (Avotec) and the View-Point EyeTracker software suite (Arrington Research). The eye tracker was calibrated at the end of each run of data acquisition.

MRI data collection and preprocessing

The experiment contains two two-hour and two three-hour scanning sessions. The first session collects structural MRI data and functional MRI data using generic localizers to delineate regions-of-interest (ROI) for each subject (V1, V2, V3, V4, FFA, PPA, OPA, RSC, AC, STG, Broca's area, and sPMv) (see Deniz et al., 2019 for details). For the remaining

two three-hour sessions, Functional MRI data was collected on different days. Each scanning session consisted of seven functional runs. Two of these runs presented the test story. The remaining five runs presented five different training stories.

MRI data were collected on a 3T Siemens TIM Trio scanner at the UC Berkeley Brain Imaging Center, using a 32- channel Siemens volume coil. Functional scans were collected using a gradient echo-EPI sequence with repetition time (TR) = 2.0045 s, echo time (TE) = 31 ms, flip angle = 70 degrees, voxel size = 2.24 × 2.24 × 4.1 mm, matrix size = 100 ×100, and field of view = 224 × 224 mm. 32 axial slices were prescribed to cover the entire cortex. A custom-modified bipolar water excitation radiofrequency(RF) pulse was used to avoid signals from fat tissue. Anatomical data were collected using a T1- weighted MP-RAGE45 sequence on the same 3T scanner. Each functional run was motion-corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL 5.0 (Jenkinson et al., 2012). All volumes in the run were then averaged to obtain a high quality template volume. FLIRT was also used to automatically align the template volume for each run to the overall template, which was chosen to be the template for the first functional run for each subject. The overall template is obtained within each participant over all runs. These automatic alignments were manually checked and adjusted for accuracy. The cross-run transformation matrix was then concatenated to the motion-correction transformation matrices obtained using MCFLIRT, and the concatenated transformation was used to resample the original data directly into the overall template space. This overall template space is also obtained within each participant across all runs. The noise of the data was estimated using six significant principal components derived from the template volume and this was subtracted from the signal (Behzadi et al., 2007). Afterwards, the response was scaled to have unit variance (i.e. z-scored). Ten TRs from the beginning and the end of each narrative were discarded in order to account for the 10 seconds of silence at the beginning and end of each scan and to account for non-stationarity in brain responses at the beginning and end of each scan.

Cortical surface reconstruction and visualization.

Cortical surface meshes were generated from the T1-weighted anatomical scans using Freesurfer software (Dale et al., 1999). Before surface reconstruction, anatomical surface segmentations were carefully checked by hand and corrected using Blender software and pycortex (Gao et al., 2015) (http://pycortex.org). Relaxation cuts were made into the surface of each hemisphere. Blender and pycortex were used to remove the surface crossing the corpus callosum (medial wall). The calcarine sulcus cut was made at the horizontal meridian in V1 using retinotopic mapping data as a guide. Functional images were aligned to the cortical surface using pycortex. Functional data were projected onto the surface for visualization and analysis using the line-nearest scheme in pycortex. This projection scheme samples the functional data at 64 evenly spaced intervals between the inner (white matter) and outer (pial) surfaces of the cortex and then averages together the samples. Samples are taken using nearest-neighbor interpolation, wherein each sample is given the value of its enclosing voxel.

Localizers for known ROIs.

Known ROIs were localized separately in each participant using standard techniques (Hansen et al, 2007) For all participants, ROIs were defined using a visual category localizer, and a retinotopic localizer. For details of localizer experiments please refer to (Deniz et al., 2019; Huth et al., 2016).

Voxelwise Modeling

Voxelwise modeling (VM) was used to model BOLD response (Deniz et al., 2019; Gong et al., 2023; Huth et al., 2012, 2016; la Tour et al., 2022; Nishimoto et al., 2011). A single joint VM model that included all feature spaces was estimated for each voxel in each subject separately using banded ridge regression (la Tour et al., 2022; Nunez-Elizalde et al., 2019). Banded ridge regression assigns a different regularization parameter for each feature space to reduce bias caused by correlations between feature spaces. Mathematically, the m delayed feature spaces $F_i(X), i \in 1, 2, ..., m$ (each of dimension p) were concatenated to form a feature matrix $F'(X)$ (dimension $(m \times p) \times n$).Then banded ridge regression was used to estimate a mapping B (dimension $v \times (\sum_{i=1}^{f} p)$) from $F'(X)$ to the matrix of voxel responses Y (dimension $v \times n$). B is estimated according to $\hat{B} =_B ||Y - BF'(X)||_2^2 + \lambda||CB||_2^2$. The diagonal matrix C of regularization hyperparameters for each feature space and each voxel is optimized over 10-fold cross-validation. The features spaces were language statistics features, including letter rate (1 parameter), word rate (1 parameter) and phoneme rate (1 parameter); phoneme features, including single phonemes (39 parameters), diphones (858 dimension) and triphones (4841 dimension); motion-energy features (2772 parameters); orthographic similarity features including letter orthographic (52 parameters) and word orthographic (2978 parameters); semantics (300 parameters) and syntactic features (33 parameters).

Feature Spaces

Language statistics feature construction.

To account for the variable speech rate, variable word length during visual presentation and the BOLD response simply elicited from presence versus absence of the reading stimuli, letter rate, word rate and phoneme rate feature spaces were constructed by counting the number of letters, words and phonemes during the acquisition of each fMRI volume (2.0045s).

Phoneme feature construction.

To account for response variance caused by phonemic content of the stories and internal speech when participants were reading the stories, single phoneme, diphones and triphones feature spaces were constructed. Each feature space was constructed by creating binary vectors encoding the presence or absence of individual phonemes (39 dimension), diphones

(858 dimension) and triphones (4841 dimension) to capture how often each phonemic unit was presented over time (Gong et al., 2023).

Motion-energy feature construction

A spatiotemporal Gabor pyramid was used to extract low-level visual features from the sequence of word frames (Adelson & Bergen, 1985; Watson & Ahumada, 1985). The word frames were first cropped to $95x500$ pixels (20 horizontal, and 5 vertical degrees of visual angle) to include all the words. The word frames were then converted to the CIE L*A*B* color space (McLaren, 1976) and the color information was discarded. The spatiotemporal Gabor pyramid consisted of a total of 2772 three-dimensional Gabor filter pairs of orthogonal quadrature spanning a square grid that covered the screen. The filters consisted of two spatial and one temporal dimension and were created using five spatial frequencies (0, 2, 4, 6, and 8 cycles/image), three temporal frequencies (0, 2 and 4 Hz), and twenty-four directions of motion (at a 15 degree interval from 0 to 360 degree). Each of the filters was convolved with the sequence of word frames. The resulting filter activations were squared and summed for each quadrature pair, resulting in a 2772-dimensional feature vector for each word frame. The output was down-sampled to the functional image acquisition rate (2.0045 s) using sinc interpolation (Oliphant, 2007). For more details, see Nishimoto et al., 2011.

Orthographic similarity feature construction.

A letter pixel-overlay feature was constructed to capture the distinctiveness of each English letter to represent its orthographic information. Distinctiveness of a letter is a function of the average overlap of that letter with every other letter in the alphabet (Wiley & Rapp, 2019). In order to calculate the distinctiveness of each letter, we first generated images of each letter with the font "Arial" and image size of $100x100$ pixels. Each image was then converted to binary arrays with background (0) and stroke (1). The overlap of two letters is then simply the percentage of pixels that contain strokes for both letters. In order to find the maximum overlap between two letters, we kept one letter fixed and shifted the other letter over to find the position of the maximum overlap. In sum, the letter pixel-overlay feature for each letter is a 52-dimension vector that measures the maximum overlap between this letter with all 52 English letters (26 upper case and 26 lower case).

$$distinctiveness_i = 1 - \frac{1}{N-1}\sum_{j!=1}^{N} overlap(i,j) \tag{3.1}$$

To represent the orthographic information of each word by extracting semantic meaning from visual representations, we also created a word orthographic feature space with the same line of thought. We first generated images of all the unique English words (2978) in the stimuli. Then, we calculated the maximum overlap between all the English words. In sum, the word orthographic feature for each word/character is a 2978-dimension vector that measures the maximum overlap between this word with all the other words.

Semantic feature construction.

To capture semantic content of stimulus narratives, each word of stimulus narrative was projected to a 300-dimensional word embedding space (Bojanowski et al., 2017). The word embeddings are from the FastText model that considers both the whole word and sub-word components to generate meaningful embeddings for out-of-vocabulary words by aggregating the vectors of its sub-word units.

Syntax feature construction.

The syntactic properties of each spoken word were labeled. A pre-trained neural network was used to create a parse tree for each sentence of the stories (Andor et al., 2016). Part-of-speech tags (e.g., noun, verb) were extracted from the parse trees to label each work. The syntax feature space (pos) was constructed by creating binary vectors encoding the presence or absence of each pos tag (33 dimension). The syntactic features were then down-sampled to the rate of acquisition of the functional images (2.0045s) using a Lanczos filter.

Feature space preprocessing

Before down-sampling to the fMRI acquisition rate, the phoneme, orthographic and semantic feature spaces were represented as unevenly sampled impulse trains. A three-lobe Lanczos filter with cutoff frequency set to the fMRI Nyquist rate (0.249Hz) was used to resample these impulse trains at evenly spaced time points corresponding to the middle of each fMRI volume. Afterwards, each feature channel was z-scored within each story (training and testing features were z-scored independently) by subtracting the mean and dividing by the standard deviation. This was done to match the features to the fMRI responses, which were also z-scored within each story. In addition, 10TRs from the beginning and 10TRs at the end of each story were discarded before VM.

Regularization hyperparameter selection

Ten-fold cross-validation was used to find the optimal regularization hyperparameters for each feature space and each voxel. Hyperparameter candidates were chosen with a random search procedure (Bergstra & Bengio, 2012): 1000 normalized hyperparameter candidates were randomly sampled from a dirichlet distribution and were then scaled by 20 log-spaced values ranging from 10^{-10} to 10^{10}. The regularization hyperparameters for each feature space and voxel were selected as the hyperparameters that produced the minimum squared error (L2) loss between the predicted voxel responses and the recorded voxel responses ($_{hyperparameters}||\hat{y} - y||_2^2$). Regularization hyperparameters were chosen separately for each participant and language. This hyperparameter search was performed using the Himalaya Python package (la Tour et al., 2022).

Model estimation and evaluation

The selected regularization hyperparameters were used to estimate model weights that map from the corresponding feature space to voxel BOLD response. Model weights were estimated separately for each voxel in each participant.

The test dataset was not used to select hyperparameters or to estimate regression weights. The prediction accuracy R^2 of the feature spaces was computed per voxel as the coefficient of determination between the predicted voxel responses and the recorded voxel responses.

Group-level prediction accuracy

Group-level prediction accuracy was computed by first computing prediction accuracy for each participant in the participant's native brain space, and then projecting individual participant results into a standard fsAverage space. Average prediction accuracy across all the participants was computed for each fsAverage vertex.

Generalization to new participants

To ensure generalization to new participants, two important steps were performed. First, the entire analysis was performed at the individual participant level – group-averaged results are shown only as summary statistics. Second, before any analyses were performed, two out of all the participants were set aside as held-out participants. The data for these participants were not analyzed until the entire data analysis and interpretation pipeline was finalized (Popham et al., 2021). Analysis of the two held-out participants was performed immediately prior to manuscript submission.

Stepwise regression procedure

To remove confounds from stimulus correlations between semantics and low-level sensory stimulus features, a stepwise regression procedure was used. First banded ridge regression was used to jointly estimate encoding models that predict BOLD responses from low-level stimulus features: visual spatial and motion features, word count, phonemes, diphones, triphones, pixel overlap, letter count (for English only), and character count (for Chinese only). Only data from the train narratives was used to estimate models. Then, the low-level models were used to predict BOLD responses to the train and test narratives. The predicted BOLD responses $\hat{Y}_{lowlevel,train}, \hat{Y}_{lowlevel,test}$ were subtracted from the true BOLD responses Y_{train}, Y_{test}. The residual BOLD responses $Y_{train} - \hat{Y}_{lowlevel,train}, Y_{test} - \hat{Y}_{lowlevel,test}$ were zscored. During z-scoring, for each voxel separately the mean response across time was subtracted and the remaining response was scaled to have unit variance. The zscored residual BOLD responses were used to estimate encoding models that predict BOLD responses from lexical semantic stimulus features.

Principal Component Analysis

We used principal components analysis (PCA) to recover a low-dimensional letter/ word orthographic space from the estimated orthographic model weights. We first selected only the significantly predicted voxels in each subject according to the average bootstrap correlation (for the selected regularization parameter value) obtained during model estimation. This was done to avoid including noise from poorly modelled voxels. Then we removed temporal information from the voxel-wise model weights by averaging across the four delays for each feature. Finally, we applied PCA to the orthographic weights, yielding 52/2978 (letter/word) principal components (PCs).

Chapter 4

Cerebral cortical representation of natural language processing in Mandarin-English bilinguals

4.1 Abstract

Bilingual brains are capable of effortlessly transforming drastically different written texts into the same meaning. To start exploring this process, we collected BOLD activity using fMRI while six Mandarin-English bilinguals read engaging narrative stories in each language. We found that similar cerebral cortical regions are involved in reading comprehension across languages, indicating a shared language processing network. However, this study also revealed several distinctions: while the visual and semantic information are represented in similar cortical regions, orthographic information is represented in different cortical areas with different features encoded across languages. Additionally, critical visual properties essential for reading also vary across languages. These differences highlight adaptive capabilities of the bilingual brains to process the specific orthographic demands of multiple languages.

4.2 Introduction

More than four billion people worldwide are multilingual (Ansaldo et al., 2008; Grosjean, 2021). These individuals can extract the same meaning from drastically different visual, orthographic, phonological, morphological, and syntactical information of multiple languages. How does the human brain achieve these tasks? Are the same cortical regions involved in processing different languages? If so, are the brain responses similar across languages? What are the visual, orthographic, and semantic features represented in each language?

Because of the diversity of experimental tasks (e.g., word reading, picture naming, semantic judgment, etc) and vast differences in participant backgrounds, including the language spoken, age of acquisition, language proficiency, the bilingual neuroimaging studies yielded

many conflicting results for the localization of each language (reviewed in Comstock, 2024). Early evidence suggests that reading in first (L1) and second (L2) languages activate similar brain regions (reviewed in Perani and Abutalebi, 2005), while more recent studies indicates that L2 reading may require additional neural resources (Xu et al., 2017; reviewed in H. Li et al., 2021). In particular, the left inferior parietal lobule showed greater activation during L2 processing (Cargnelutti et al., 2019; Zhang et al., 2023). Other meta-analysis studies reveal a much more complex picture, suggesting a shared neural network that varies depending on the linguistic level of grammar and phonology (Sulpizio et al., 2020).

At the visual perceptual level, studies have mainly focused on examining the impact of visual attention span (VAS), which is the number of distinct visual elements that can be simultaneously processed at a glance (Bosse et al., 2007). Fluent reading requires strong VAS capacity to simultaneously process several orthographic units (Huang et al., 2019; Lobier et al., 2013). Deficits in VAS could lead to dyslexia, characterized by difficulties in multielement parallel processing (Lobier et al., 2012; Valdois, 2022). At the orthographic representational level, differences in writing systems: alphabetic (e.g. English) versus logographic (e.g. Mandarin) can result in distinct brain activation patterns (Shen & Del Tufo, 2022). For bilinguals sharing a writing system (e.g. English and French), there is nearly complete overlap in the visual word form area (VWFA), found in the left ventral occipitotemporal cortex (VOTC) and considered to be involved in orthographic analysis of written texts. However, in fMRI, distinct VWFA patches appear for languages with different writing systems (e.g English versus Chinese) (Zhang et al., 2023). At the semantic representational level, evidence suggests that the same brain regions and meaning specific activations are shared across languages. For example, in bilateral occipitotemporal cortex and the inferior and middle temporal gyrus, the semantic information of one language could be decoded from the other (Van de Putte et al., 2017). However, language-specific semantic representations are also evident. Brain lesions in multilinguals sometimes selectively impair one language while sparing another (Fabbro, 2001; Mooijman et al., 2022), and multilinguals perceive semantic information differently between their primary and secondary languages (Malik-Moraleda et al., 2023).

Despite these insights, the systematic investigations of visual, orthographic, and semantic representation in bilingual brains remains underexplored. The mixed results from bilingual studies is partly due to reliance on null hypothesis testing, which often forces a binary choice between shared or separate representations (Comstock, 2024; H. Li et al., 2021; Sulpizio et al., 2020). Furthermore, most studies have investigated languages in isolation, ignoring the dynamic interplay and integration of both languages in bilingual individuals (Chee et al., 2004). Thirdly, most prior studies use unnatural and strictly controlled stimuli, such as isolated letters, words, phrases, or pseudowords, which may not accurately represent real-world reading comprehension (Comstock, 2024; Sulpizio et al., 2020). To overcome these limitations, our study employs large-scale, data-driven methods to analyze brain representations using naturalistic stimuli, aiming to go beyond a dichotomy of shared or separate representations and enrich our understanding of cerebral cortical representation of different languages in bilinguals.

Here we designed a study to perform comprehensive, fine-grained, within-participants

comparisons of visual, orthographic, and semantic representations between languages. We used functional magnetic resonance imaging (fMRI) to record brain responses while six Mandarin-English bilingual participants engaged in a naturalistic reading task for over four hours in two languages, English and Mandarin (Figure 4.1b). We used voxelwise encoding models to map visual, orthographic, and semantic representations throughout the cerebral cortex, and used these maps to assess the similarities and differences of these representations between the two languages.

4.3 Results

The goal of this study was to investigate the similarities and differences in the cerebral cortical representations of reading comprehension across Mandarin-English bilinguals. To do this, we used fMRI to record the BOLD activity of six Mandarin-English bilinguals while they read narrative stories for two hours in each language (English and Mandarin). Words were presented on the screen one at a time at normal reading speeds (Figure 4.1, Materials and Methods).

Cerebral cortical regions involved in natural language processing are similar

Before examining specific features, we first aimed to compare the similarities and differences in which cerebral cortical regions might be involved in any stage of reading across languages. In order to achieve this goal, we quantified the signal-to-noise ratio (SNR) of each cortical vertex involved throughout all stages of reading across languages. In particular, we showed one 10 min narrative story four times in Mandarin and English for each subject. For each vertex in the cerebral cortex of each subject in each language, we then computed the pairwise Pearson correlation coefficients, which is referred to as explainable variance (EV, Materials and Methods). Areas of the cerebral cortex that are involved in any aspect of reading should respond reliably across the four presentations, while areas not involved in this processing should not respond reliably. Figure C.1a shows the EV for each language and each participant separately. In all participants, EV is high in both languages in voxels located in early visual, temporal, lateral inferior parietal, medial parietal, and prefrontal cortices, including superior temporal sulcus (STS), fusiform gyrus, angular gyrus (AG) and supramarginal gyrus, precuneus, superior frontal gyrus, and inferior frontal gyrus (IFG). The activation pattern is also found bilaterally. The 2D density plot shows that EV of all the significantly explained voxels for each participant and both languages cluster along the diagonal line, indicating that the SNR of the recorded BOLD responses was comparable for each participant across languages (Figure C.1b).

Group-level EV was computed to summarize results across participants. EV for each participant was projected to a standard fsAverage brain. Group-averaged EV for each fsAverage vertex was computed as the average EV across participants, for each language separately.

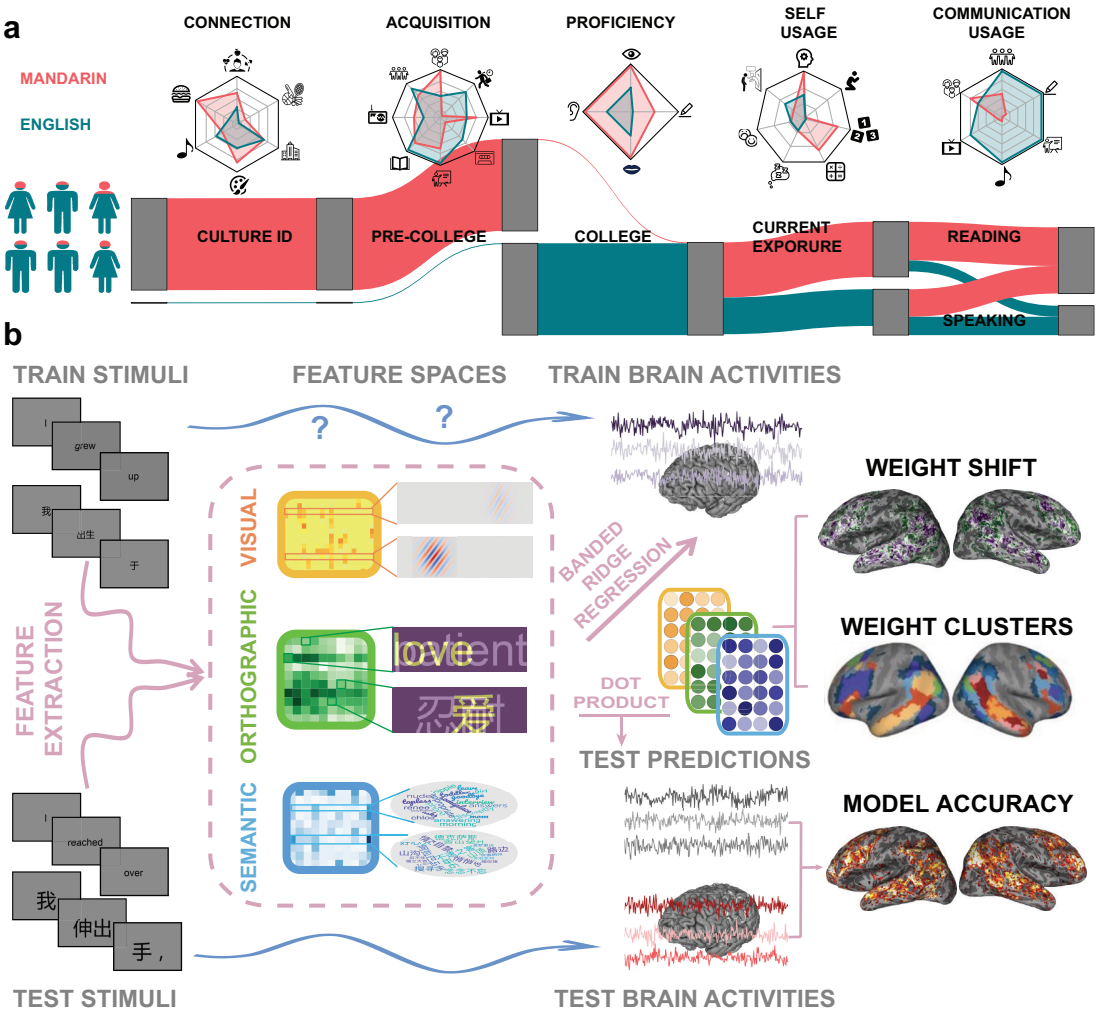


Figure 4.1: **Participant language demographic information and experimental procedure.** (**a**) Participant's language demographic information. The language proficiency and experience measured by the Language Experience and Proficiency Questionnaire (LEAP-Q) and Language History Questionnaire (LHQ3) of three male and three female Mandarin-English bilinguals (human figure) participated in the experiment are summarized and shown. Participants' primary language is Mandarin (coral background) and are fluent in English with standard scores (TOEFL) over 80% (teal filling). The Sankey plot tracks the major shifts in participant's language history. All participants were born and raised in China, and migrated to live in an English speaking country since college. All participants identified with Chinese culture. The connection plot shows participants' average ratings of connection in way of life, food, music, art, cities and sports of each culture related to the language. Participants started English acquisition on average at 7 years old, and the acquisition plot shows how much friends, families, reading, tape, tv, radio, school and work contributed to their language acquisition. The proficiency plot shows the averaged self reported language proficiency in reading, speaking, listening and writing. The self-usage plot shows the frequency of their language usage to think, self talk, express emotions, dream, perform arithmetics, remember numbers, and pray. The communication plot shows their language exposure through interaction with friends, families, watching television, listening to music, in school or writing. (**b**) Experimental procedure and voxelwise modeling (VM). Six fluent Mandarin-English bilingual participants read over two hours of natural narratives in each language while BOLD responses were measured using fMRI. Visual stimulus features were constructed by passing each frame through a fixed set of nonlinear spatiotemporal motion-energy filters. Orthographic stimulus features were constructed by computing the maximum pixel overlap of all word or character pairs. Semantic stimulus features were constructed by projecting each stimulus word into a 300-dimensional word embedding space (fastText). These stimulus features and BOLD responses were used to estimate a separate FIR ridge regression model for every voxel, participant and language. These estimated model weights were used to predict BOLD responses for a separate held-out story that was not used for model estimation. Model accuracy was quantified as the coefficient of determination R^2 between predicted and recorded BOLD responses to the held-out narratives. Principal component analysis was performed on the model weights for interpretation. The estimated model weights describe the features that preferentially activate each voxel in each language ("feature tuning" in each language). For each voxel the feature tuning shift describes which feature elicit higher BOLD activity in one language relative to the other.

Figure C.1c shows vertex-wise group-averaged EV on the flattened cortical surface of the fsAverage brain for each language, revealing that similar to results for individual participants, vertices within the bilateral occipital, temporal, parietal, and prefrontal cortices are well-explained across participants for both English and Mandarin. In Figure 4.2a, group-level averaged EV is directly compared between English and Mandarin by using blue vs red color code. Most vertices throughout bilateral occipital, temporal, parietal, and prefrontal cortices are white showing that most cerebral cortical regions involved in natural language processing are shared across English and Mandarin. Quantitatively, the 2D density plot as the inset in Figure 4.2c shows that group-level averaged EV of both languages of all vertices cluster along the diagonal line. It also directly quantifies the group-level EV averaged across all the vertices, showing that 4.97% variance of English BOLD responses are reliably explained by the story, and 4.91% of Mandarin. This result indicates that the SNR of the recorded BOLD responses toward stories averaged across the entire cerebral cortex are comparable across languages at the group level.

To further quantitatively compare the SNR in different cortical regions involved in reading in English and Mandarin, we calculated the averaged EV in twenty-four regions of interests (ROIs) spanning the entire cerebral cortex in both languages. Figure 4.2c shows that all these twenty four regions are significantly involved in reading in both languages. More than 5% of the BOLD responses of voxels in these areas could be explained by the joint model. Significantly story-responsive vertices span the entire putative language-processing pathway from early visual cortex through semantic areas and high-level association cortex. This suggests that we should be able to use this data to build and test voxelwise encoding models that capture the representations and computations at many different stages of natural language processing. Furthermore, Figure 4.2c also shows that English EV is significantly higher than Mandarin in V1, while Mandarin EV is significantly higher than English in V4, LO1, LO2, pITG, pFFG, FFA, OFA, STG. There is no significant difference in EV between languages in V2, V3, V3A, V3B, IPS0, IPS1, IPS2, sPMv, PPA, AG, SMG, MPC, IPFC, SPFC, OPFC. This result indicates that the SNR is higher in English in low level visual areas, higher in Mandarin in high visual areas, and it is comparable in intermediate visual areas and semantic and language areas between English and Mandarin. This result indicates that in the group level of Mandarin-English bilinguals, the SNR shows differences in some cortical area between English and Mandarin that needs further examinations.

Cerebral cortex is more involved in representing primary language

The repeatability analysis showed that a relatively large fraction of cortical voxels responds reliably to natural language processing and thus may represent different information in the natural language processing. To identify what information is represented in the BOLD response of each voxel in each language, we fit a linear encoding model to every cortical voxel in each participant's brain for each language (Deniz et al., 2019; Gong et al., 2023; Huth et al., 2012, 2016; la Tour et al., 2022; Nishimoto et al., 2011). Before fitting the linear encoding model, BOLD responses from fourteen voxels around the eye balls were regressed out to

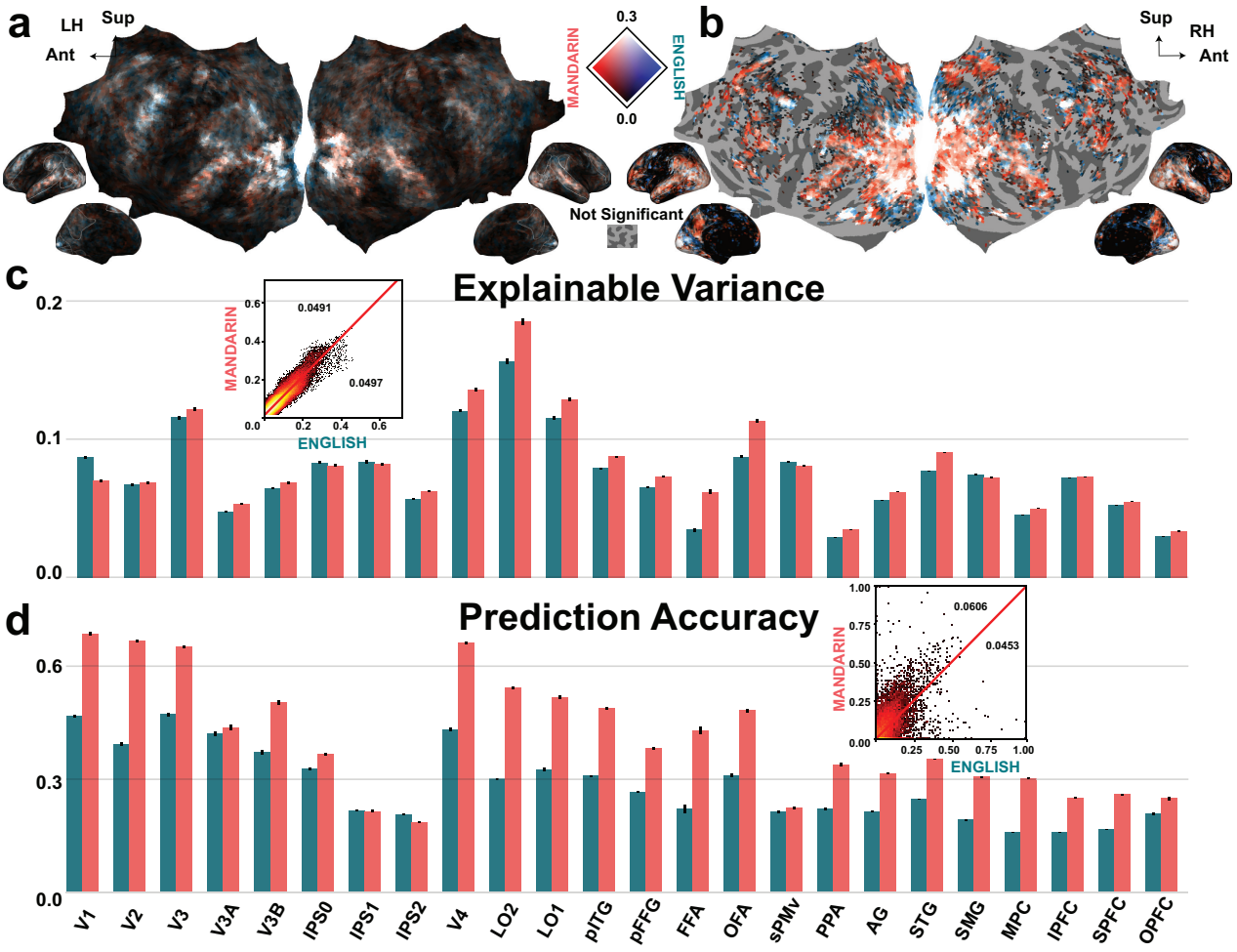


Figure 4.2: **Explainable variance and joint model prediction accuracy of English versus Mandarin.** Explainable variance and joint model prediction accuracy of English versus Mandarin. Subjects read one validation story four times in both English and Mandarin, in order to estimate the signal-to-noise ratio (SNR) of each cortical voxel involved throughout all stages of reading comprehension. Voxelwise encoding modeling (VM) was used to estimate joint model weights using all eleven feature spaces for each of the six participants and two languages separately. (**a**) Group-averaged explainable variance (EV) for English and Mandarin are shown on the flattened cortical surface of the standard fsAverage brain with a 2D colormap. EV for each subject was projected into a standard fsAverage space, and then the average EV across subjects was computed for each fsAverage vertex per language separately. Vertices with higher English EV are shown in blue. Vertices with higher Mandarin EV are shown in red. Vertices with similar EV in English and Mandarin are shown in white. Many vertices throughout bilateral occipital, temporal, parietal, and prefrontal cortices are white, indicating that the SNR is comparable between English and Mandarin. (**b**) Grouped-averaged joint model prediction accuracy for English and Mandarin are shown on the flattened cortical surface of the standard fsAverage brain with a 2D colormap. Only vertices with significant accuracy in both languages were shown. Prediction accuracy is given by the same color scale as the EV. Vertices for which the joint model prediction accuracy is not statistically significant are shown with the curvature (gray). The joint model consists of word rate, letter rate, visual, orthographic, and semantic feature spaces. Averaged across participants, vertices in bilateral occipital, temporal, parietal, and prefrontal cortices are significantly predicted, indicating that similar cortical regions are involved in natural language processing across languages. Many vertices are red, showing that Mandarin prediction accuracy is systematically higher than English prediction accuracy. (**c**) The average EV by ROIs for each language is shown in the bar plot. In V2, V3, V3A, V3B, IPS0, IPS1, IPS2, sPMv, PPA, AG, SMG, MPC, IPFC, SPFC, OPFC, there is no significant difference in EV between languages. English EV is significantly higher than Mandarin in V1, while Mandarin EV is significantly higher than English in V4, LO1, LO2, pITG, pFFG, FFA, OFA, STG. The inset shows the 2D density plot of group-averaged EV for English (x-axis) vs Mandarin (y-axis). Points indicate all vertices. Brighter colors indicate a larger number of vertices in the corresponding bin. EV of vertices cluster along the diagonal line. (**d**) The average prediction accuracy of the joint model by ROIs for each language is shown in the bar plot. In IPS1 and IPS2, there is no significant difference in the prediction accuracy between languages. The joint model systematically and significantly predicts better in Mandarin than English in all the other ROIs. The inset shows the 2D density plot of group-averaged prediction accuracy from the joint model for English vs Mandarin. Most vertices are better predicted by Mandarin than English.

remove the whole brain BOLD responses that could be explained by the eye movement and mere presence versus absence of the reading stimuli. The linear encoding model consisted of five feature spaces, each of which captured a different kind of information in the experiment. Word count and letter (English)/character(Mandarin) count feature spaces capture the number of words and letters/characters of each TR to account for the BOLD response from the length of reading stimuli. Visual stimulus features were constructed by passing each frame through a fixed set of nonlinear spatiotemporal motion-energy filters (Nishimoto et al., 2011) to capture the low-level graphic information of each word frame. Orthographic stimulus features were constructed by rendering texts (words in English and characters in Mandarin) as images and computing the maximum pixel overlap of images of all word and character pairs to capture the intermediate level representations of texts based on orthographic similarity. Semantic stimulus features were constructed by projecting each stimulus word into a 300-dimensional word embedding space (fastText, Bojanowski et al., 2017). The word embedding space is aligned between languages such that words with similar meanings project to nearby vectors. The linearized encoding model was fit to the BOLD responses for each voxel, participant and each language separately using a training set. Banded ridge regression (la Tour et al., 2022; Nunez-Elizalde et al., 2019) was used to fit all five feature spaces as a single joint model to the recorded BOLD responses in the training dataset. To validate that the model was not overfitted to the training dataset and evaluate the model accuracy both within and across languages, the model was then used to predict the recorded BOLD responses in the hold-out test dataset of each language. Model goodness of fit was assessed on the hold-out test dataset, and it was computed as the amount of variance in the recorded BOLD response explained by the model-predicted BOLD responses (the coefficient of determination R^2). In this study, this value is also referred to as the model prediction accuracy. To compare the similarities and differences between the visual, orthographic and semantic representations between languages, we compared the estimated model weights of each feature space between English and Mandarin. To validate the robustness of our results and ensure generalization to new participants the entire analysis was performed for each participant separately, and data for two of the six participants were not analyzed until the entire data analysis and interpretation pipeline was finalized.

Figure 4.2b shows the joint model prediction accuracy using all five feature spaces averaged across all six participants, projected to a standard fsAverage brain (Figure C.2 for individual participants). Across all participants, model prediction accuracy is significantly high in many voxels located in bilateral visual cortex, superior temporal sulcus (STS), angular gyrus and supramarginal gyrus, inferior and medial parietal cortex, inferior and superior prefrontal cortex for both languages. The result shows that large and similar cortical regions are involved in natural language processing in both languages, and these cortical regions are consistently engaged across subjects. In addition, many vertices are red, showing that the joint model systematically produces higher prediction accuracy in Mandarin than in English for Mandarin-English bilinguals. This observation was further quantified by the 2D density plot as the inset in Figure 4.2d, showing that group-level averaged prediction accuracy of both languages of all vertices cluster above the diagonal line: 4.53% variance of English

BOLD responses are reliably predicted by the joint model, and 6.06% of Mandarin. To further quantitatively compare the prediction accuracy in different cortical regions involved in reading in English and Mandarin, we calculated average prediction accuracy of the joint model in different ROIs. Figure 4.2d shows that the joint model systematically and significantly predicts better in Mandarin than English in all the ROIs apart from IPS1 and IPS2. These results suggest that primary language is better represented in the cerebral cortex of bilinguals.

Within language comparison of cerebral cortical representation of visual, orthographic and semantic information in reading

To further compare the similarities and differences in the cortical representation of the visual, orthographic and semantic information of reading between English and Mandarin, the split prediction accuracy of visual (red), orthographic (green), and semantic (blue) feature spaces are summarized across participants in both languages, projected to one standard fsAverage brain (Figure 4.3a, Figure C.6 for individual participants). For English, visual features best explain response of vertices in low and intermediate level visual cortex, and orthographic features best explain vertices in high level visual cortex, angular gyrus and supramarginal gyrus. For Mandarin, visual and orthographic features both explain vertices in visual cortex well, and orthographic features explain more variance of vertices in angular gyrus, inferior temporal gyrus. Semantic features best explain vertices in superior temporal sulcus, medial parietal, superior prefrontal and inferior prefrontal cortices in both English and Mandarin. This result indicates that the cerebral cortical regions involved in visual and semantic representations are shared across languages, while the cortical regions involved in orthographic representations show differences across Mandarin and English.

These RGB joint model prediction accuracy maps show that many voxels are colored in yellow, teal and magenta, indicating that response of many voxels can be significantly predicted by more than one feature space. To facilitate feature comparison, we used variance partitioning method to quantify the fraction of variance explained for each language by each feature space individually and the fraction that might be equally well explained by any combination of feature spaces (de Heer et al., 2017; Gong et al., 2023; Lescroart & Gallant, 2019). For this purpose, we first fit the joint models with all five features (word count, letter/character count, visual, visual, orthographic and semantic) using banded ridge regression with the optimal hyperparameter for each voxel (la Tour et al., 2022). Afterwards, we fixed the hyperparameters for each voxel and refit the model with all the possible combinations of feature spaces: as follows: three models based on a single feature space (visual, orthographic and semantic); three models based on paris of feature spaces (visual-orthographic, visual-semantic, orthographic-semantic); and a single model that used all three feature spaces together. Then, using set theory, we divided the variance explained by these feature spaces into the following seven partitions: the variance explained uniquely by each feature space; the variance explained jointly by each pair of feature spaces but excluding the third; and

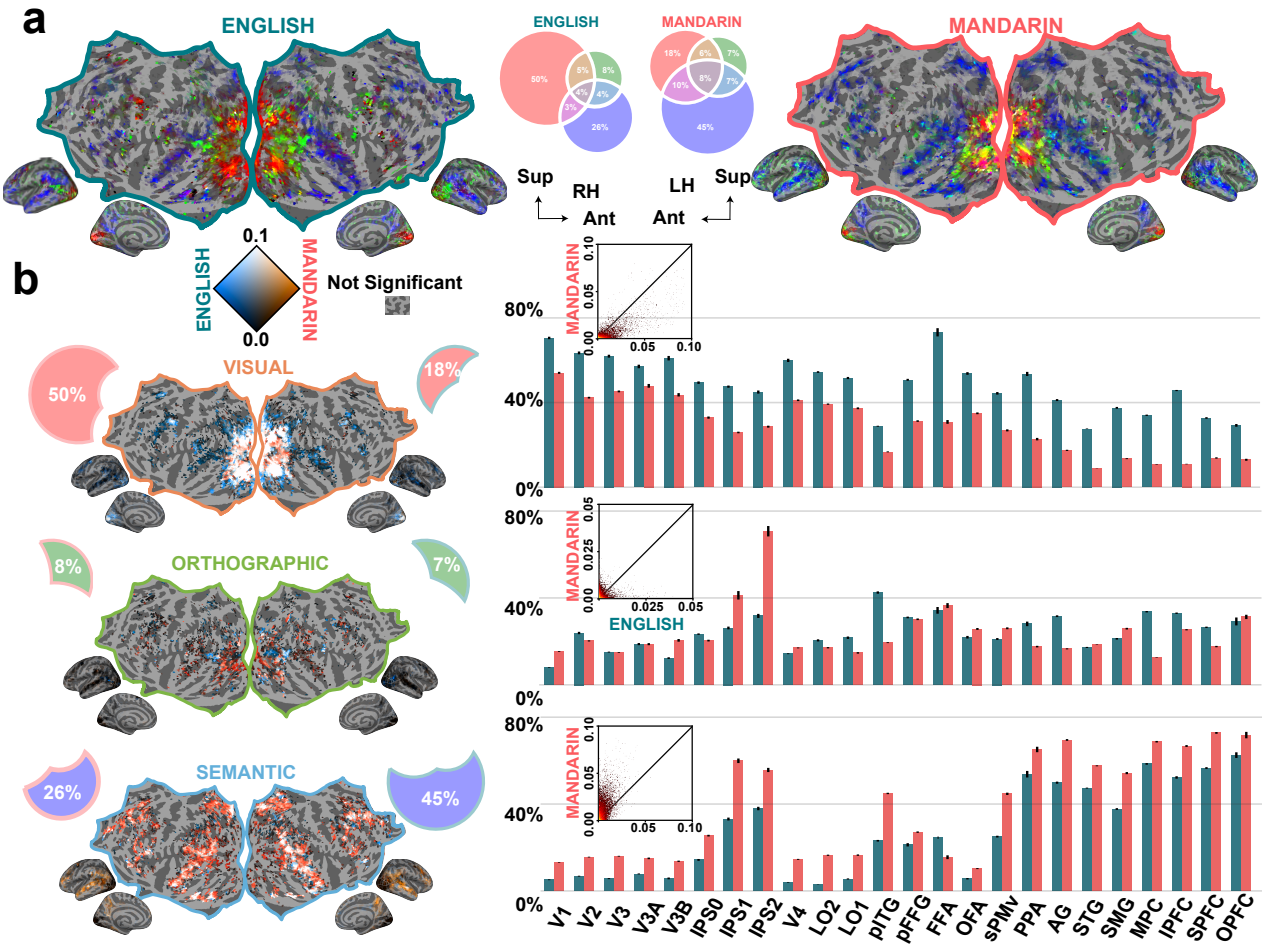


Figure 4.3: **Variance partitioning.** To understand the cortical representation of natural language processing in English and Mandarin, a voxelwise encoding model consisting of visual, orthographic and semantic feature spaces was constructed for each of the six participants and two languages separately. (**a**) Group-averaged prediction accuracy is shown for visual (red), orthographic (green) and semantic (blue) feature spaces on the flattened cortical surface of the fsAverage brain for English (teal, left) and Mandarin (coral, right). Only vertices with significant prediction accuracy are shown. Across participants in English, visual features best explain response of vertices in low and intermediate level visual cortex, orthographic feature best explain vertices in high level visual cortex, angular gyrus and supramarginal gyrus, and semantic features best explain vertices in superior temporal sulcus, medial parietal, superior prefrontal and inferior prefrontal cortices. Across participants in Mandarin, visual and orthographic features both explain vertices in visual cortex well, and orthographic features explain more variance of vertices in angular gyrus and inferior temporal gyrus. Semantic features best explain Mandarin BOLD response of vertices in superior temporal sulcus, medial parietal, superior prefrontal and inferior prefrontal cortices. To quantify the cortical representation by each feature space, a variance-partitioning analysis was used to separate the variance explained for each language by the visual, orthographic, and semantic feature spaces into seven partitions. Venn diagram shows the percentage of the variance explained by each partition across all subjects (calculated using only significantly predicted vertices) in the entire cortex. (**b**) Grouped-averaged prediction accuracy for the visual, orthographic and semantic feature spaces in English and Mandarin are shown on the flattened cortical surface of the standard fsAverage brain with a 2D colormap. Vertices better explained by English are shown in blue. Vertices better explained by Mandarin are shown in red. Vertices equally well explained by both English and Mandarin are shown in white. The average prediction accuracy of each feature space by ROIs for each language is shown in the bar plot. For visual features, prediction accuracy is significantly higher in English than in Mandarin throughout all cortical regions. For orthographic features, prediction accuracy is significantly higher in English in V2, IPS0, LO1, LO2, pITG, PPA, AG, MPC, IPFC, SPFC, while prediction accuracy is significantly higher in Mandarin in V1, V3B, IPS1, IPS2, V4, OFA, sPMv, SMG. There is no significant difference in the accuracy between English and Mandarin in V3, V3A, pFFG, FFA, STG. For semantic features, prediction accuracy is significantly higher in English in FFA, while prediction accuracy is significantly higher in Mandarin in all the other ROIs. The inset shows the 2D density plot of group-averaged prediction accuracy across the entire cerebral cortex of each feature space for English (x-axis) vs Mandarin (y-axis), showing that more vertices are better explained by visual features in English than Mandarin, while more vertices are better explained by semantic features in Mandarin than English. There are no significant nor systematic differences in the amount of vertices explained by orthographic features between English and Mandarin.

the variance explained jointly by all three feature spaces (Materials and Methods). With this approach, we were able to quantify the extent to which any particular voxel represented features from each feature space, after taking into account its selectivity to other feature spaces. The Venn diagram in Figure 4.3a shows the percentage of the variance of BOLD responses explained by each partition across all participants (calculated using only significantly predicted vertices). It reveals that averaged across the entire cortex, more BOLD responses are involved in visual representation in English than Mandarin, while more BOLD responses are involved in semantic representation in Mandarin than English.

To further quantify this observation, grouped-averaged unique variance explained for the visual, orthographic and semantic feature spaces in English and Mandarin are shown on the flattened cortical surface of the standard fsAverage brain with a 2D colormap (Figure 4.3b; Figure C.7 for other partitions). The unique contribution from visual features space is significantly higher in English than in Mandarin throughout all cortical regions. The unique contribution from orthographic feature space is significantly higher in English in V2, IPS0, LO1, LO2, pITG, PPA, AG, MPC, IPFC, SPFC, while it is significantly higher in Mandarin in V1, V3B, IPS1, IPS2, V4, OFA, sPMv, SMG. There is no significant difference in the unique contribution from the orthographic features between English and Mandarin in V3, V3A, pFFG, FFA, STG. The unique contribution of semantic features is significantly higher in English in FFA, while it is significantly higher in Mandarin in all the other ROIs. Furthermore, we summarized the group-averaged unique contribution across the entire cerebral cortex of each feature space for English and Mandarin. Figure 4.3b insets show that more vertices are better explained by visual features in English than Mandarin, while more vertices are better explained by semantic features in Mandarin than English. There are no significant nor systematic differences in the amount of vertices explained by orthographic features between English and Mandarin. This result indicates that for Mandarin-English bilinguals in the group level, different cortical areas are systematically more involved in visual representation for English, while more involved in semantic representation for Mandarin. Cortical representation for orthographic features show differences in different cortical regions across English and Mandarin that need further examination.

Across language comparison of cerebral cortical representation of visual, orthographic and semantic information in reading

To directly compare the cerebral cortical representations in visual, orthographic and semantic information between English and Mandarin, we tested whether model weights estimated in one language could accurately predict BOLD responses to stimuli in the other language (Figure 4.4). If the representation is shared between languages, then a model estimated with BOLD responses to one language will generalize well to the BOLD responses to a different language. Based on this logic, we first estimated the BOLD residuals for each language by removing the BOLD responses that could be explained by the joint model using language statistics (words/letters/characters rate, Materials and Methods) from this

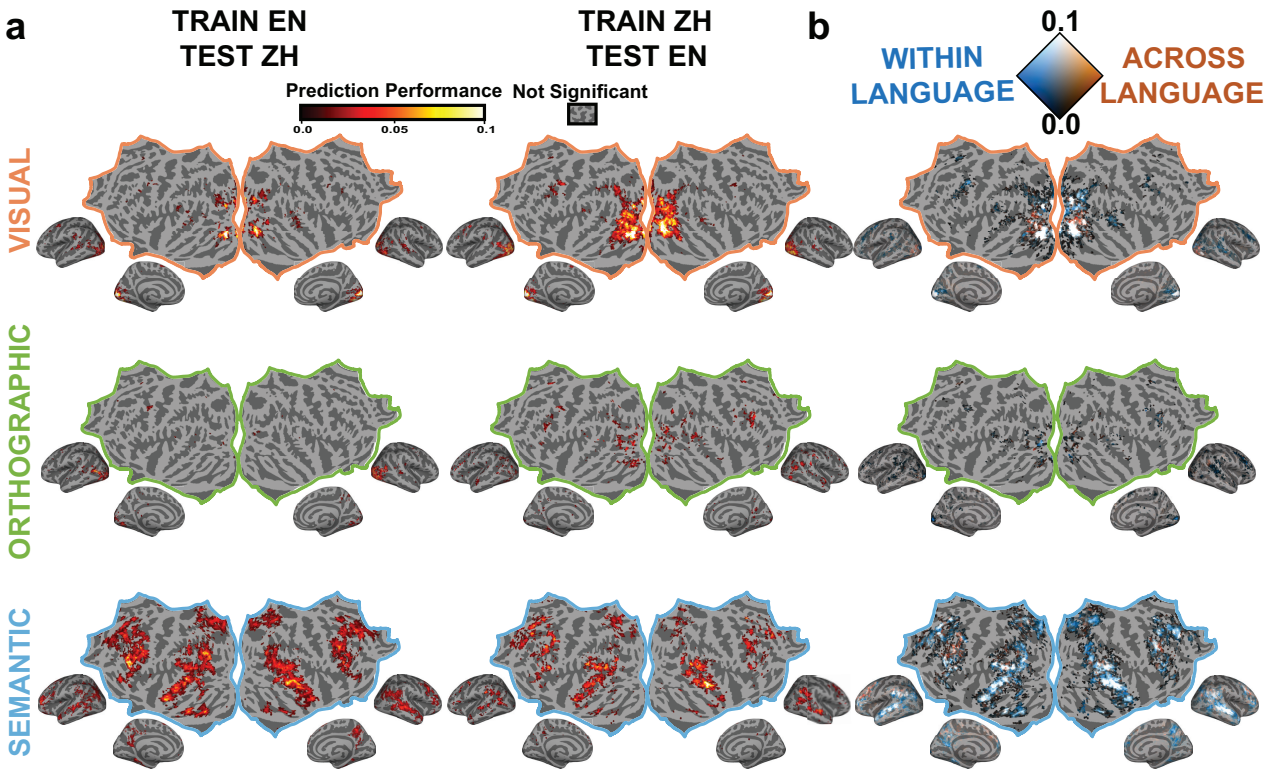


Figure 4.4: **Across-language prediction accuracy.** To investigate if the visual, orthographic and semantic information are shared across languages for Mandarin-English bilinguals, BOLD response residuals for each subject and each language were obtained by regressing out any BOLD response explained by the mere presence and absence of reading stimuli. Afterwards, visual, orthographic and semantic features are used to create a joint model with the BOLD residual for each subject and language separately. Estimated model weights from each feature space for one language (e.g. English) were used to predict held-out data for the other language (e.g. Mandarin). Prediction accuracy of each feature space was first computed as the coefficient of determination (R^2) between predicted response from the feature space and recorded BOLD responses from the other language, and then divided by EV to correct for the differences of SNR between languages. (**a**) Group-averaged across-language prediction accuracy is shown on the flattened cortical surface of the standard fsAverage brain. The flatmaps in the first column show the model weights estimated from visual (first row), orthographic (second row), and semantic (third row) feature spaces for English predict responses to Mandarin, and vice versa are shown in the second column. The flatmaps show that many voxels are well predicted across languages in the visual and semantic feature spaces, while very limited voxels are well predicted across languages in the orthographic feature space. (**b**) For group-averaged results, the mean of across-language prediction accuracy and mean of within-language prediction accuracy are shown on the flattened cortical surface of the standard fsAverage brain with a 2D colormap. Vertices that are well-predicted across-language are shown in blue. Vertices that are well-predicted within-language are shown in orange. Vertices that are well-predicted both across- and within-language are shown in white. For visual and semantic feature spaces, many vertices are white throughout bilateral visual, temporal, parietal, and prefrontal cortices. For orthographic features, most vertices are in the visual cortex and sPMv and are blue.

language. Afterwards, we used visual, orthographic and semantic features to create a joint model with the BOLD residual for each participant and language separately. Estimated model weights from each feature space for one language (e.g. English) were used to predict held-out data for the other language (e.g. Mandarin). Prediction accuracy of each feature space was first computed as the coefficient of determination (R^2) between predicted response from the feature space and recorded BOLD responses from the other language, and then divided by EV to correct for the differences of SNR between languages.

Figure 4.4a shows the group-averaged across-language prediction accuracy projected on the flattened cortical surface of the standard fsAverage brain (results for individual participants are similar and are shown in Figure C.8, Figure C.9, Figure C.10). The flatmaps show that many voxels are well predicted across languages in the visual and semantic feature spaces, while very limited voxels are well predicted across languages in the orthographic feature space. This result indicates that for Mandarin-English bilinguals, the visual and semantic representations are shared across languages, while the orthographic representations are not.

To directly compare vertex-wise across language prediction accuracy to within-language prediction accuracy, we used a 2D colormap to visualize the average across-language prediction accuracy and the average within-language prediction accuracy on the flattened cortical surface of the standard fsAverage brain (Figure 4.4b). For visual and semantic feature spaces, many vertices are white throughout bilateral visual, temporal, parietal, and prefrontal cortices. For orthographic features, most vertices are in the visual cortex and sPMv and are blue. This result further supported the previous observation that for Mandarin-English bilinguals, the visual and semantic information are represented in the same cortical regions, while the orthographic representations are not.

Visual feature tuning of reading in English versus Mandarin

To characterize the differences between languages in how specific visual properties are represented, we looked at the direction, spatial frequency and temporal frequency tuning of the vertices that are significantly predicted by visual features in both English and Mandarin. These vertices are referred to as visually selective vertices. The visual feature tuning for each vertex is given by its 2772-dimensional vector of estimated visual feature space model weights, one weight value for each of 2772 spatio-temporal Gabor filters. The filters are created at multiple spatial and temporal frequencies, directions of motion, and x-y positions. For each vertex, the visual feature tuning describes the optimal visual feature that predicts its BOLD response. Furthermore, to determine whether the difference between languages in the visual property tuning stem from the language difference (English versus Mandarin) or from the participants' language experience (English monolinguals versus Mandarin-English bilinguals), we fit the same joint models with five features on the BOLD response from nine English monolingual participants who read the same English narrative stories.

Figure 4.5 shows the group-averaged optimal direction (first column), spatial frequency (second column) and temporal frequency (third column) tuning of each vertex for English

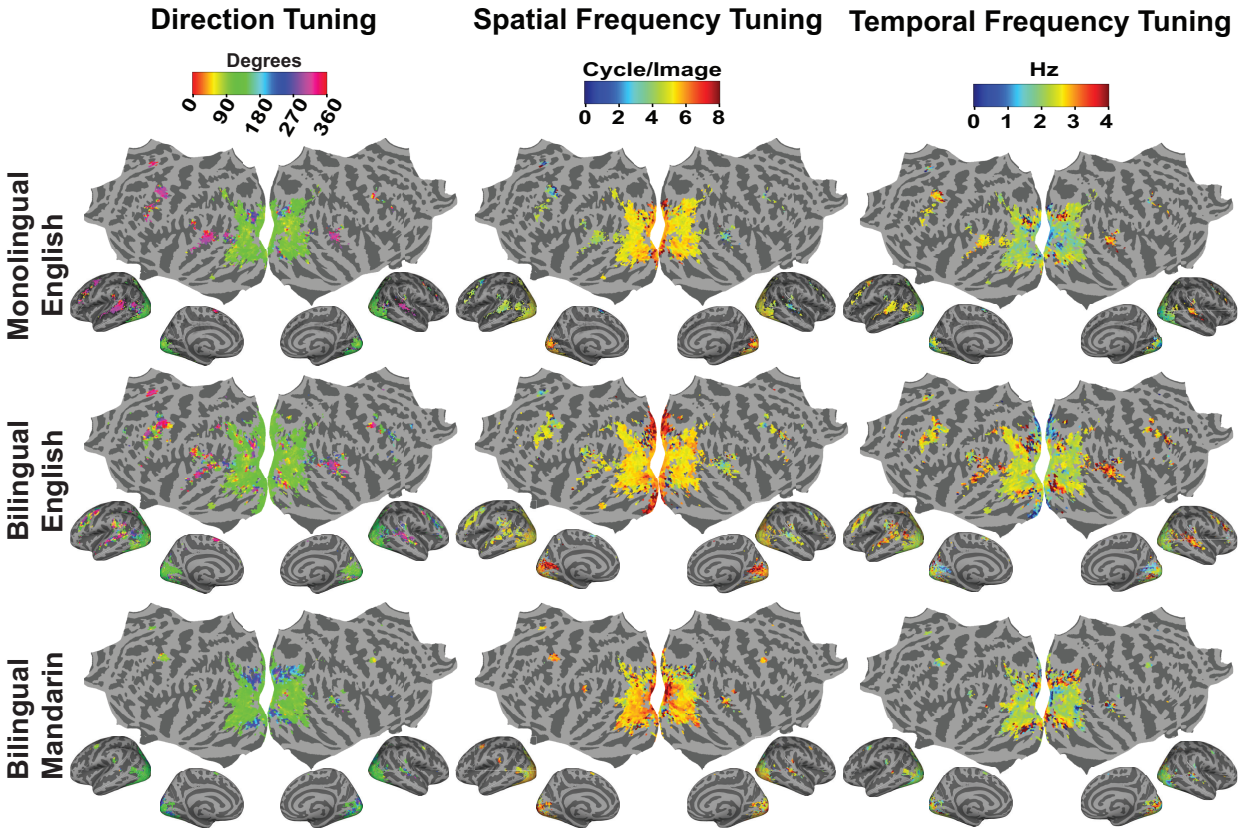


Figure 4.5: **Visual feature tuning.** To compare the similarities and differences of the selectivities of visual properties for each visually selective vertex between English and Mandarin, the visual tuning for each vertex for English monolinguals (first row) and Mandarin-English bilinguals (second and third row, within subject comparison) is examined. Group-averaged optimal direction (first column), spatial frequency (second column) and temporal frequency (third column) tuning of each vertex are shown on the flattened cortical surface of the fsAverage brain. (**a**) Direction tuning. For English (both monolingual and bilingual), vertices in the visual cortex are mostly tuned to directions around 90 degrees, while vertices in peripheral visual cortex, STS, sPMv are mostly tuned to directions between 310 and 340 degrees. For Mandarin, vertices in visual cortex and language related areas are mostly tuned to directions around 90 degrees, while vertices in the peripheral visual cortex are mostly tuned to directions between 210 and 270 degrees. (**b**) Spatial frequency tuning. For English (both monolingual and bilingual), vertices in the visual cortex are tuned to the spatial frequency similarly ranging between 5 and 8 cycles per image. In STS and sPMv, vertices in bilinguals are tuned to higher spatial frequency in bilinguals than that in monolinguals. For Mandarin, vertices are tuned to higher spatial frequencies than that in English. (**c**) Temporal frequency tuning. For English, vertices in the visual cortex and STS are tuned to higher temporal frequencies for bilinguals than monolinguals, while vertices in sPMv are tuned to similar temporal frequencies. For bilinguals, vertices in sPMv are tuned to lower temporal frequencies in Mandarin than English.

monolinguals (first row) and Mandarin-English bilinguals (second and third row) projected on the flattened cortical surface of the fsAverage brain (results for individual participants are similar and are shown in Figure C.11). For English (both monolingual and bilingual speakers), vertices in the visual cortex are mostly tuned to directions around 90 degrees, while voxels in peripheral visual cortex, STS, sPMv are mostly tuned to directions between 310 and 340 degrees. For Mandarin, vertices in visual cortex and language related areas are mostly tuned to directions around 90 degrees, while vertices in the peripheral visual cortex are mostly tuned to directions between 210 and 270 degrees. This result indicates that when the cerebral cortex extracts meaning from visual stimuli, the horizontal directions correlating with word length is very important for English comprehension, while it is less important for Mandarin comprehension. In addition, for English (both monolingual and bilingual), vertices in the visual cortex are tuned to the spatial frequency similarly ranging between 5 and 8 cycles per image. In STS and sPMv, vertices in bilinguals are tuned to higher spatial frequency in bilinguals than that in monolinguals. For Mandarin, vertices are tuned to higher spatial frequencies than that in English. This result indicates that when the cerebral cortex extracts meaning from visual stimuli, Mandarin-English bilinguals pay more attention to finer details of the visual stimuli in both English and Mandarin comprehension than English monolinguals. It also suggests that in general, Mandarin comprehension requires attending to finer details than English comprehension.

Lastly, for English, vertices in the visual cortex and STS are tuned to higher temporal frequencies for bilinguals than that for monolinguals, while vertices in sPMv are tuned to similar temporal frequencies. This result suggests that Mandarin-English bilinguals are more sensitive to the temporal variations in the visual representations of English than English monolinguals. In addition, for bilinguals, vertices in sPMv are tuned to lower temporal frequencies in Mandarin than that in English. This result indicates that when the cerebral cortex extracts meaning from visual stimuli for Mandarin-English bilinguals, it integrates information from a longer temporal window for Mandarin than for English.

4.4 Discussion

We systematically compared cerebral cortical representations of visual, orthographic, and semantic representations between English and Mandarin in fluent Mandarin-English bilinguals. Our results provide unique evidence of how the cerebral cortex performs reading comprehensions in different languages in bilinguals' brains. First, similar cerebral cortical regions, including voxels throughout occipital, temporal, parietal, and prefrontal cortices, are involved in natural language processing (reading comprehension) in both languages (Figure 4.2a,c, Figure C.1), while the activation level varies across different cortical areas between languages. Second, the cerebral cortex is more involved in representing primary (L1) than secondary language (L2) (Figure 4.2b,d, Figure C.2). Third, for extracting semantic meaning from visual stimuli, the cerebral cortex is more involved in visual representations in L2 than L1, while more involved in semantic processing in L1 than L2 (Figure 4.3, Figure C.3,

Figure C.4, Figure C.5 and Figure C.6). Fourthly, the visual and semantic information are represented in the similar cerebral cortical areas, while orthographic information are represented in different cortical areas (Figure 4.4, Figure C.8, Figure C.9 and Figure C.10). Also, the visual and semantic model weights are highly corrected between languages and generalize well between languages, while the orthographic model weights are not generalized well between languages. Lastly, although visual information is represented in the same cortical regions between languages, the specific visual properties tuning, including direction, spatial frequency and temporal frequency tuning, show systematic differences between languages, as well as between monolingual and bilingual speakers (Figure 4.5, Figure C.11, Figure C.12 and Figure C.13).

The first finding from this experiment is that for English-Mandarin bilinguals, the same and a wide range of cortical regions are reliably involved in processing both languages including bilateral occipital, temporal, parietal and prefrontal cortices. This result is in congruent with previous studies showing that primary (L1) and secondary language (L2) reading share the same set of brain regions (reviewed in Perani and Abutalebi, 2005). In addition, the primary visual cortex (V1) exhibits a higher SNR to L2 (English) than L1 (Mandarin). This could be explained by the fact that the length of English words are longer and more variable than Mandarin characters, which elicits more responses in the peripheral visual areas. In addition, cortical areas, including V4, LO1, LO2, pITG, pFFG, and FFA, exhibit higher SNR to L1 than L2. These areas are critically involved in orthographic processing (Dehaene et al., 2002; Larsson & Heeger, 2006; Roe et al., 2012; Silson et al., 2013; Zhan et al., 2023). The result makes sense as Mandarin is a more complicated writing system than English, which requires more orthographic analysis and representations.

The second finding of this study is that although the cortical regions involved in representing different languages are the same in bilingual brains and repeatable signals (SNR) are comparable across most cortical areas, the cerebral cortex is more involved in encoding primary (L1) than secondary language (L2). This contrast strengthens the necessity and importance of our methodology to build explicit models to test what feature is encoded in each voxel for each participant. Replying solely on comparing the brain activations from different stimuli could lead to misleading conclusions. One plausible explanation of this result is the difference in language competence between L1 and L2 among our participants. All participants were born and raised in China, began learning English at ages ranging from 2 to 11, and lived in an English-speaking country for between 5 to 12 years. The correlation analysis shows that English standard scores are significantly positively correlated with the joint model prediction accuracy for English, indicating that despite all participants being fluent in both English and Mandarin (all over 85% of standard scores), their English competence does not match with their Mandarin proficiency. On the one hand, our study contributes to the growing body of research suggesting that language competence, rather than age of acquisition or duration of language immersion, influence how languages are processed in the brain Perani and Abutalebi, 2005). On the other hand, this reveals a limitation of our study: not being able to recruit bilinguals who grew up in English speaking countries and are also highly fluent in Mandarin. Future studies should examine these populations.

The third finding from this experiment is that bilingual brains exhibit more visual representations in L2 than L1, while more semantic representations in L1 than L2. One plausible explanation is that for bilingual brains where L2 proficiency is lower than L1, visual processing demands more effort in L2 than L1. Compared to L1, L2 comprehension requires more effort to recognize the visual input, and more cognitive effort to extract semantic meaning from the visual input (Freeman & Marian, 2022). At the semantic level, compared to L2, L1 lexicons are associated with a broader range of semantic senses and quicker to be accessed (Brysbaert et al., 2017). Furthermore, this result reveals one main advantage of our research methodology. By collecting brain responses over long duration using natural narrative stories with rich contextual information, and modeling these responses using rich and interpretable features, this methodology allows us to discern subtle differences in language proficiency that remain undetected through other methods.

The fourth finding from this experiment is that different cortical areas are involved in orthographic representation between two languages with distinct writing systems (logographic Mandarin and alphabetic English) in bilingual brains. This finding aligns with the previous studies (Das et al., 2011; Kumar et al., 2010; Meschyan & Hernandez, 2006), such as Zhan and colleagues (Zhan et al., 2023) showed that different VWFA patches are found for languages with different writing systems. In addition, our result showed that the orthographic representations are not shared across languages in the Mandarin-English bilingual brains. Previous studies showed features of the orthographic organization are different between different writing systems (G. Ding et al., 2003; Liu & Cao, 2016; C. A. Perfetti & Liu, 2005). For example, Tian and colleagues (Tian et al., 2020) showed that for processing alphabetic languages, more word-like stimuli, such as real words and pseudo-words, induced higher activation in the middle and anterior left posterior ventral occipitotemporal area (vOT) than false fonts. However, for processing logographic languages, false characters induced stronger responses compared to real and pseudo characters in the bilateral middle vOT. However, our experiment identified more cortical areas involved in orthographic analysis and showed differences in the representation across languages. Previous studies most likely did not identify these many cortical areas and the differences between languages because of the stimuli and methodological differences. We collected a relatively large amount of fMRI data (2 hours for each language) for each participant using natural narrative stories. In contrast, prior studies collected a small amount of fMRI data per participant from many participants, using strictly controlled stimuli, such as real words/characters, pseudo-characters/words, false characters/words/fonts, consonant letter strings, stroke combinations, etc. On the one hand, these stimuli are very different from natural reading comprehension, thus they elicit very limited brain activities and the task the brain is performing is more aligned with word recognition than natural reading comprehension. On the other hand, pseudo and false words/characters are artificially generated stimuli, so it's very hard to disentangle the confounding effects, such as attention and new exposure, from the genuine orthographic processing.

Lastly, our study also finds that language experience modulates the visual representations in the cerebral cortex, which is consistent with previous studies showing that early

exposure to a language continues to influence the neural processing of subsequently learned language (Pierce et al., 2015). Comparing the visual feature tuning in monolingual brains with bilingual brains shows that bilingual brains are tuned to higher spatial frequencies in processing Mandarin than English, which is also higher than monolingual English speakers. One plausible explanation underlying this result is because Chinese reading comprehension demands greater attention to fine structural details than English. This requirement stems from the complex visual-spatial layouts of Chinese characters, each representing a morpheme or a syllable, unlike alphabetic systems where simpler structures prevail. Prior studies have shown that sentence comprehension in Chinese involves significant reanalysis and coordination across different syntactic levels due to its intricate character composition (Yang et al., 2021). In addition, bilingual brains are tuned to slower temporal frequencies in processing Mandarin than English, which is also slower than monolingual English speakers. This is consistent with psychophysics tests of reading speed showing on average, people could read in Mandarin $158 +/- 19$ words/min, which is slower than English of $228 +/- 19$ words/min (Trauzettel-Klosinski et al., 2012).

4.5 Materials and Methods

Participants

Structural and functional brain data were collected from six Chinese-English bilingual participants: $S1$ (female, age 29), $S2$ (male, age 25), $S3$ (female, age 25), $S4$ (male, age 25), $S5$ (female age 24), $S6$ (male, age: 26). All participants were fluent in both Mandarin Chinese (native) and English (non-native). All the participants were borned and raised in China. On average, the participants were first exposed to English at the age of $7+/-4$, and spent $8+/-3$ years in a country where English is spoken. At the time of the experiment, all participants used English in their everyday life and had an English proficiency score $87\% + / - 12$ as measured by the TOEFL. All participants primarily used Chinese in interactions with family, English in interactions at school or work, and a mix of the two languages in interactions with friends. Each subject's English and Chinese bilingual capabilities were measured using Language Experience and Proficiency Questionnaire (LEAP-Q, Marian et al., 2007) and Language History Questionnaire (LHQ3, P. Li et al., 2020).

All participants were healthy and had normal or corrected-to-normal vision. Each subject's handedness was evaluated by the Edinburgh handedness inventory [[cite]]. All subjects were healthy and had normal or corrected-to-normal vision and self-reported normal hearing. The use of human subjects in this study was approved by the UC Berkeley Committee for the Protection of Human Subjects. A written statement of informed consent has been obtained from each subject.

Stimuli

The stimuli in these experiments consisted of eleven 10− to 15− min stories that were pre-recorded from the Public Radio Exchange (PRX) radio show "The Moth Radio Hour", which has been used in previous studies of our lab (C. Chen et al., 2024; de Heer et al., 2017; Deniz et al., 2019, 2023; Gong et al., 2023; Huth et al., 2016). They are engaging stories performed verbally in English that capture the attention of the subjects. The audio recording of each story was manually transcribed, and the written transcription was aligned to the audio recording (see Deniz et al., 2019 for details). To construct matched Chinese stimuli, each of the English stories was translated into Chinese by a professional translation service. To obtain word presentation times that correspond to natural speech, each translated story was performed out by two professional voice actors (one male and one female), and then the written translations were aligned to these audio recordings. The stimuli were split into separate model estimation and model validation sets. The model estimation stimulus-set consisted of ten 10− to 15− min stories played once each. The length of each scan was tailored to the story and also included 10 seconds of silence both before the story and 14 seconds after the story. Each subject read the same 10 stories in both English and Simplified Mandarin. The model validation stimulus-set consisted of a single 10-min story told that was read four times for each subject in each language in order to estimate voxel response reliability and noise ceiling. This resulted in 3737 (English)/ (Chinese) time points (sampled at TR) for the training dataset and 291 (English)/(Chinese) time points for the validation dataset.

Stimulus presentation format

The words of each story were presented one-by-one at the center of the screen using a Rapid Serial Visual Presentation (RSVP) procedure (Buchweitz et al., 2009; Forster, 1970). Each word was presented for a duration equal to the duration of that word in the spoken version of the story.

The Python pygame library was used to display text on a gray background at 34 horizontal, and 27 vertical degrees of visual angle. Black letters were used to present the words. For English stimuli, words were presented at average 6 (min = 1, max = 16) horizontal and 3 vertical degrees of visual angle. For Chinese stimuli, words were presented at average 6 (min = 3, max = 12) horizontal and 3 vertical degrees of visual angle. A white fixation cross was present at the center of the display throughout the experiment. Participants were instructed to fixate on a center cross while reading the story. Participants' eye movements were monitored at 60 Hz throughout the scanning sessions using a custom-built camera system equipped with an infrared source (Avotec) and the View-Point EyeTracker software suite (Arrington Research). The eye tracker was calibrated at the end of each run of data acquisition. All participants read all the stories in both English and Chinese. Story presentation order was balanced between languages and across participants. To verify comprehension and attention, the participants were asked at the end of each session outside the scanner to

recount the contents of each story. All participants were able to accurately summarize the contents of each narrative.

MRI data collection and preprocessing

The experiment contains two two-hour and two three-hour scanning sessions. The first session collects structural MRI data and functional MRI data using generic localizers to delineate regions-of-interest (ROI) for each subject (V1, V2, V3, V4, FFA, PPA, OPA, RSC, AC, STG, Broca's area, and sPMv) (see Deniz et al., 2019 for details). For the remaining two three-hour sessions, Functional MRI data was collected on different days. Each scanning session consisted of seven functional runs. Two of these runs presented the test story. The remaining five runs presented five different training stories.

MRI data were collected on a 3T Siemens TIM Trio scanner at the UC Berkeley Brain Imaging Center, using a 32- channel Siemens volume coil. Functional scans were collected using a gradient echo-EPI sequence with repetition time (TR) = 2.0045 s, echo time (TE) = 31 ms, flip angle = 70 degrees, voxel size = 2.24 × 2.24 × 4.1 mm, matrix size = 100 ×100, and field of view = 224 × 224 mm. 32 axial slices were prescribed to cover the entire cortex. A custom-modified bipolar water excitation radiofrequency(RF) pulse was used to avoid signals from fat tissue. Anatomical data were collected using a T1- weighted MP-RAGE45 sequence on the same 3T scanner. Each functional run was motion-corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL 5.0 (Jenkinson et al., 2012). All volumes in the run were then averaged to obtain a high quality template volume. FLIRT was also used to automatically align the template volume for each run to the overall template, which was chosen to be the template for the first functional run for each subject. The overall template is obtained within each participant over all runs. These automatic alignments were manually checked and adjusted for accuracy. The cross-run transformation matrix was then concatenated to the motion-correction transformation matrices obtained using MCFLIRT, and the concatenated transformation was used to resample the original data directly into the overall template space. This overall template space is also obtained within each participant across all runs. The noise of the data was estimated using six significant principal components derived from the template volume and this was subtracted from the signal (Behzadi et al., 2007). Afterwards, the response was scaled to have unit variance (i.e. z-scored). Ten TRs from the beginning and the end of each narrative were discarded in order to account for the 10 seconds of silence at the beginning and end of each scan and to account for non-stationarity in brain responses at the beginning and end of each scan.

Cortical surface reconstruction and visualization.

Cortical surface meshes were generated from the T1-weighted anatomical scans using Freesurfer software (Dale et al., 1999). Before surface reconstruction, anatomical surface segmentations were carefully checked by hand and corrected using Blender software and

pycortex (Gao et al., 2015) (http://pycortex.org). Relaxation cuts were made into the surface of each hemisphere. Blender and pycortex were used to remove the surface crossing the corpus callosum (medial wall). The calcarine sulcus cut was made at the horizontal meridian in V1 using retinotopic mapping data as a guide. Functional images were aligned to the cortical surface using pycortex. Functional data were projected onto the surface for visualization and analysis using the line-nearest scheme in pycortex. This projection scheme samples the functional data at 64 evenly spaced intervals between the inner (white matter) and outer (pial) surfaces of the cortex and then averages together the samples. Samples are taken using nearest-neighbor interpolation, wherein each sample is given the value of its enclosing voxel.

Localizers for known ROIs.

Known ROIs were localized separately in each participant using standard techniques (Hansen et al, 2007) For all participants, ROIs were defined using a visual category localizer, and a retinotopic localizer. For details of localizer experiments please refer to (Deniz et al., 2019; Huth et al., 2016).

Voxelwise Modeling

Voxelwise modeling (VM) was used to model BOLD response (Deniz et al., 2019; Gong et al., 2023; Huth et al., 2012, 2016; la Tour et al., 2022; Nishimoto et al., 2011). A single joint VM model that included all feature spaces was estimated for each voxel in each subject separately using banded ridge regression (la Tour et al., 2022; Nunez-Elizalde et al., 2019). Banded ridge regression assigns a different regularization parameter for each feature space to reduce bias caused by correlations between feature spaces. Mathematically, the m delayed feature spaces $F_i(X), i \in 1, 2, ..., m$ (each of dimension p) were concatenated to form a feature matrix $F'(X)$ (dimension $(m \times p) \times n$).Then banded ridge regression was used to estimate a mapping B (dimension $v \times (\sum_{i=1}^{f} p)$) from $F'(X)$ to the matrix of voxel responses Y (dimension $v \times n$). B is estimated according to $\hat{B} =_B ||Y - BF'(X)||_2^2 + \lambda||CB||_2^2$. The diagonal matrix C of regularization hyperparameters for each feature space and each voxel is optimized over 10-fold cross-validation. The features spaces were language statistics features, including letter rate (1 parameter), word rate (1 parameter) and phoneme rate (1 parameter); phoneme features, including single phonemes (39 parameters), diphones (858 dimension) and triphones (4841 dimension); motion-energy features (2772 parameters); orthographic similarity features including letter orthographic (52 parameters) and word orthographic (2978 parameters); semantics (300 parameters) and syntactic features (33 parameters).

Feature Spaces

Language statistics feature construction.

To account for the variable speech rate, variable word length during visual presentation and the BOLD response simply elicited from presence versus absence of the reading stimuli,

letter rate, word rate and phoneme rate feature spaces were constructed by counting the number of letters, words and phonemes during the acquisition of each fMRI volume (2.0045s).

Motion-energy feature construction

A spatiotemporal Gabor pyramid was used to extract low-level visual features from the sequence of word frames (Adelson & Bergen, 1985; Watson & Ahumada, 1985). The word frames were first cropped to $95x500$ pixels (20 horizontal, and 5 vertical degrees of visual angle) to include all the words. The word frames were then converted to the CIE L*A*B* color space (McLaren, 1976) and the color information was discarded. The spatiotemporal Gabor pyramid consisted of a total of 2772 three-dimensional Gabor filter pairs of orthogonal quadrature spanning a square grid that covered the screen. The filters consisted of two spatial and one temporal dimension and were created using five spatial frequencies (0, 2, 4, 6, and 8 cycles/image), three temporal frequencies (0, 2 and 4 Hz), and twenty-four directions of motion (at a 15 degree interval from 0 to 360 degree). Each of the filters was convolved with the sequence of word frames. The resulting filter activations were squared and summed for each quadrature pair, resulting in a 2772-dimensional feature vector for each word frame. The output was down-sampled to the functional image acquisition rate (2.0045 s) using sinc interpolation (Oliphant, 2007). For more details, see Nishimoto et al., 2011.

Orthographic similarity feature construction.

A letter pixel-overlay feature was constructed to capture the distinctiveness of each English letter to represent its orthographic information. Distinctiveness of a letter is a function of the average overlap of that letter with every other letter in the alphabet (Wiley & Rapp, 2019). In order to calculate the distinctiveness of each letter, we first generated images of each letter with the font "Arial" and image size of $100x100$ pixels. Each image was then converted to binary arrays with background (0) and stroke (1). The overlap of two letters is then simply the percentage of pixels that contain strokes for both letters. In order to find the maximum overlap between two letters, we kept one letter fixed and shifted the other letter over to find the position of the maximum overlap. In sum, the letter pixel-overlay feature for each letter is a 52-dimension vector that measures the maximum overlap between this letter with all 52 English letters (26 upper case and 26 lower case).

$$distinctiveness_i = 1 - \frac{1}{N-1}\sum_{j!=1}^{N} overlap(i,j) \tag{4.1}$$

With the same line of thought, we also created a word pixel-overlay feature to capture the orthographic information of each English word or Chinese character. We first generated images (4680) of all the unique English words (2978) and Chinese characters (1882) in the stimuli. Then, we calculated the maximum overlap between all the English words and Chinese characters. In sum, the word pixel-overlay feature for each word/character is a

4680-dimension vector that measures the maximum overlap between this word/character with all the other words/characters.

Semantic feature construction.

To capture semantic content of stimulus narratives, each word of stimulus narrative was projected to a 300-dimensional word embedding space (Bojanowski et al., 2017). The word embeddings are from the FastText model that considers both the whole word and sub-word components to generate meaningful embeddings for out-of-vocabulary words by aggregating the vectors of its sub-word units.

Feature space preprocessing

Before down-sampling to the fMRI acquisition rate, the phoneme, orthographic and semantic feature spaces were represented as unevenly sampled impulse trains. A three-lobe Lanczos filter with cutoff frequency set to the fMRI Nyquist rate (0.249 Hz) was used to resample these impulse trains at evenly spaced time points corresponding to the middle of each fMRI volume. Afterwards, each feature channel was z-scored within each story (training and testing features were z-scored independently) by subtracting the mean and dividing by the standard deviation. This was done to match the features to the fMRI responses, which were also z-scored within each story. In addition, 10 TRs from the beginning and 10 TRs at the end of each story were discarded before VM.

Regularization hyperparameter selection

Ten-fold cross-validation was used to find the optimal regularization hyperparameters for each feature space and each voxel. Hyperparameter candidates were chosen with a random search procedure (Bergstra & Bengio, 2012): 1000 normalized hyperparameter candidates were randomly sampled from a dirichlet distribution and were then scaled by 20 log-spaced values ranging from 10^{-10} to 10^{10}. The regularization hyperparameters for each feature space and voxel were selected as the hyperparameters that produced the minimum squared error (L2) loss between the predicted voxel responses and the recorded voxel responses ($_{hyperparameters}||\hat{y} - y||_2^2$). Regularization hyperparameters were chosen separately for each participant and language. This hyperparameter search was performed using the Himalaya Python package (la Tour et al., 2022).

Model estimation and evaluation

The selected regularization hyperparameters were used to estimate model weights that map from the semantic feature space to voxel BOLD response. Model weights were estimated separately for each voxel in each language and participant. The model weight for each voxel

and language reflects the semantic tuning of the voxel in that language, the types of words that elicit the highest response in each voxel.

The test dataset was not used to select hyperparameters or to estimate regression weights. The prediction accuracy R^2 of the feature spaces was computed per voxel as the coefficient of determination between the predicted voxel responses and the recorded voxel responses. To determine which voxels represent semantic information in each language, prediction accuracy was computed for within-language predictions (train and test on the same language). To determine how well model weights estimated for one language generalize to the other language, prediction accuracy was also computed for across-language predictions (train on one language, and test on a different language).

Group-level prediction accuracy

Group-level prediction accuracy was computed by first computing prediction accuracy for each participant in the participant's native brain space, and then projecting individual participant results into a standard fsAverage space. Average prediction accuracy across six participants was computed for each fsAverage vertex.

Generalization to new participants

To ensure generalization to new participants, two important steps were performed. First, the entire analysis was performed at the individual participant level – group-averaged results are shown only as summary statistics. Second, before any analyses were performed, two out of the six participants were set aside as held-out participants. The data for these participants were not analyzed until the entire data analysis and interpretation pipeline was finalized (Popham et al., 2021). Analysis of the two held-out participants was performed immediately prior to manuscript submission.

Stepwise regression procedure

To remove confounds from stimulus correlations between semantics and low-level sensory stimulus features, a stepwise regression procedure was used. First banded ridge regression was used to jointly estimate encoding models that predict BOLD responses from low-level stimulus features: visual spatial and motion features, word count, phonemes, diphones, triphones, pixel overlap, letter count (for English only), and character count (for Chinese only). Only data from the train narratives was used to estimate models. Then, the low-level models were used to predict BOLD responses to the train and test narratives. The predicted BOLD responses $\hat{Y}_{lowlevel,train}, \hat{Y}_{lowlevel,test}$ were subtracted from the true BOLD responses Y_{train}, Y_{test}. The residual BOLD responses $Y_{train} - \hat{Y}_{lowlevel,train}, Y_{test} - \hat{Y}_{lowlevel,test}$ were zscored. During z-scoring, for each voxel separately the mean response across time was subtracted and the remaining response was scaled to have unit variance. The zscored residual

BOLD responses were used to estimate encoding models that predict BOLD responses from lexical semantic stimulus features.

Principal Component Analysis

We used principal components analysis (PCA) to recover a low-dimensional visual space from the estimated visual model weights. We first selected only the significantly predicted voxels in each subject according to the average bootstrap correlation (for the selected regularization parameter value) obtained during model estimation. This was done to avoid including noise from poorly modelled voxels. Then we removed temporal information from the voxel-wise model weights by averaging across the four delays for each feature. Finally, we applied PCA to the visual weights, yielding 2772 principal components (PCs).

Chapter 5

Conclusion

In this dissertation, we make progress in understanding the cerebral cortical representation of language in both speech and reading, through three fMRI experiments using naturalistic stimuli and voxelwise encoding models. In Chapter 2, we analyzed BOLD activity from eleven participants when they listened to engaging autobiographical stories for over two hours. We found that the cerebral cortex segments continuous speech into diphones at the phonemic level to facilitate semantic comprehension. This segmentation aligns with phonetic theory, which suggests that phonemes vary acoustically when produced in combinations (Öhman, 1966). This variability complicates the direct transformation from temporal-spectral acoustic features to single phoneme identification, but is mitigated by diphone segmentation. Given the relatively modest number of diphones used in languages (e.g. approximately 860 in English), it is plausible that the brain optimizes its phonemic representation by directly operating on spectral-temporal features to identify diphones. However, the study has limitations that future research should address. First, we have not thoroughly investigated segmentation based on other speech features, such as syllables, envelope peaks, etc. Second, our models have only considered static segmentation, and future models should benefit from incorporating adaptive segmentation using both acoustic and linguistic features. Lastly, our study lacks direct evidence of diphone specificity, such as a cortical map organized by diphone but not triphone structures. Therefore, additional explorations on the exact nature of the cortical representation for phonemic segments is needed.

The experiments in Chapter 2 demonstrate the feasibility of our methodology for studying language. Therefore, Chapter 3 extends this investigation into the cerebral cortical representation of reading comprehension. We analyzed BOLD activity when nine human participants read engaging narrative stories for over two hours. Contrary to traditional views that visual, orthographic, and semantic representation are segregated within the cerebral cortex, our results show that the brain represents multiple levels of information simultaneously in different cortical areas. The discrepancy from previous studies likely results from methodological differences, such as our use of unsmoothed data and log naturalistic stimuli. Additionally, we found that the human cerebral cortex represents English letters orthographically based on their curvature and intersections, which are the key features in object recognition within

the ventral visual pathways. This supports the neuronal recycling hypothesis that the brain repurposes existing neural circuits for new functions such as reading, adapting objects and face recognition systems for process orthographic information (Dehaene et al., 2004). Despite these insights, this study failed to interpret the low-dimensional space where word orthographic features are represented. Therefore, future research is needed to interpret this space, and understand how these features integrate with visual, phonemic and semantic processing.

The work in Chapter 3 naturally raises the question: how are languages represented in the bilinguals' brain? We explored this question in Chapter 4 by collecting BOLD activity when six Mandarin-English bilinguals read engaging narrative stories for over four hours in both English and Mandarin. We found that similar cerebral cortical regions are involved in reading comprehension across languages, suggesting a shared language processing network. This is consistent with literature suggesting that primary (L1) and secondary language (L2) utilize the same brain regions for reading (Perani & Abutalebi, 2005). Additionally, our study expands previous findings by identifying additional cortical areas involved in orthographic processing, likely due to our use of prolonged, naturalistic stimuli which different significantly from the more limited tasks typically used in traditional experiments. Furthermore, we found that orthographic information is processed differently in distinct cortical areas across languages. This result is consistent with previous studies showing that different writing systems elicit different cortical activation patterns (G. Ding et al., 2003; Liu & Cao, 2016; C. A. Perfetti & Liu, 2005). This variability suggests the adaptive capabilities of bilingual brains to understand diverse orthographic systems.

In summary, these experiments are very different from traditional methods to study language representation in the brain with strictly controlled stimuli. These results provide a new perspective on understanding how different aspects of language are represented in the brain. This is a bird's-eye view of the visual, orthographic and semantic representations in the brain, where we see the large scale patterns. These experiments were an attempt to a holistic, total picture of the whole language representation stream, from sound or graphs to meaning.

Bibliography

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Josa a*, *2*(2), 284–299.

Al-Zubaidi, A., Bräuer, S., Holdgraf, C. R., Schepers, I. M., & Rieger, J. W. (2022). Sub-lexical cues affect degraded speech processing: Insights from fmri. *Cerebral cortex communications*, *3*(1), tgac007.

Ansaldo, A. I., Marcotte, K., Scherer, L., & Raboyeau, G. (2008). Language therapy and bilingual aphasia: Clinical implications of psycholinguistic and neuroimaging research. *Journal of Neurolinguistics*, *21*(6), 539–557.

Baker, C. I., Liu, J., Wald, L. L., Kwong, K. K., Benner, T., & Kanwisher, N. (2007). Visual word processing and experiential origins of functional selectivity in human extrastriate cortex. *Proceedings of the National Academy of Sciences*, *104*(21), 9087–9092.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R., et al. (2006). Top-down facilitation of visual recognition. *Proceedings of the national academy of sciences*, *103*(2), 449–454.

Behrmann, M., & Plaut, D. C. (2014). Bilateral hemispheric processing of words and faces: Evidence from word impairments in prosopagnosia and face impairments in pure alexia. *Cerebral Cortex*, *24*(4), 1102–1118.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (compcor) for bold and perfusion based fmri. *Neuroimage*, *37*(1), 90–101.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, *19*(12), 2767–2796.

Binkofski, F. C., Klann, J., & Caspers, S. (2016). On the neuroanatomy and functional role of the inferior parietal lobule and intraparietal sulcus. In *Neurobiology of language* (pp. 35–47). Elsevier.

Boets, B., Op de Beeck, H. P., Vandermosten, M., Scott, S. K., Gillebert, C. R., Mantini, D., Bulthé, J., Sunaert, S., Wouters, J., & Ghesquière, P. (2013). Intact but less accessible phonetic representations in adults with dyslexia. *Science*, *342*(6163), 1251–1254.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135–146.

Bookheimer, S. (2002). Functional mri of language: New. *Annu. Rev. Neurosci*, *25*, 151–88.

Booth, J. R., Burman, D. D., Meyer, J. R., Gitelman, D. R., Parrish, T. B., & Mesulam, M. M. (2002). Modality independence of word comprehension. *Human brain mapping*, *16*(4), 251–261.

Bornkessel-Schlesewsky, I., Schlesewsky, M., Small, S. L., & Rauschecker, J. P. (2015). Neurobiological roots of language in primate audition: Common computational properties. *Trends in cognitive sciences*, *19*(3), 142–150.

Bosse, M.-L., Tainturier, M. J., & Valdois, S. (2007). Developmental dyslexia: The visual attention span deficit hypothesis. *Cognition*, *104*(2), 198–230.

Bouhali, F., de Schotten, M. T., Pinel, P., Poupon, C., Mangin, J.-F., Dehaene, S., & Cohen, L. (2014). Anatomical connections of the visual word form area. *Journal of Neuroscience*, *34*(46), 15402–15414.

Bracci, S., & Op de Beeck, H. P. (2023). Understanding human object vision: A picture is worth a thousand representations. *Annual review of psychology*, *74*, 113–135.

Brant-Zawadzki, M., Gillan, G. D., & Nitz, W. R. (1992). Mp rage: A three-dimensional, t1-weighted, gradient-echo sequence–initial experience in the brain. *Radiology*, *182*(3), 769–775.

Brodbeck, C., & Simon, J. Z. (2020). Continuous speech processing. *Current Opinion in Physiology*, *18*, 25–31.

Brysbaert, M., Lagrou, E., & Stevens, M. (2017). Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. *Bilingualism: Language and Cognition*, *20*(3), 530–548.

Buchweitz, A., Mason, R. A., Tomitch, L., & Just, M. A. (2009). Brain activation for reading and listening comprehension: An fmri study of modality effects and individual differences in language comprehension. *Psychology & neuroscience*, *2*, 111–123.

Burton, M. W. (2001). The role of inferior frontal cortex in phonological processing. *Cognitive Science*, *25*(5), 695–709.

Callan, D. E., Jones, J. A., & Callan, A. (2014). Multisensory and modality specific processing of visual speech in different regions of the premotor cortex. *Frontiers in psychology*, *5*, 82762.

Cargnelutti, E., Tomasino, B., & Fabbro, F. (2019). Language brain representation in bilinguals with different age of appropriation and proficiency of the second language: A meta-analysis of functional imaging studies. *Frontiers in human neuroscience*, *13*, 154.

Carter, B. T., Foster, B., Muncy, N. M., & Luke, S. G. (2019). Linguistic networks associated with lexical, semantic and syntactic predictability in reading: A fixation-related fmri study. *NeuroImage*, *189*, 224–240.

Carter, R. (2019). *The human brain book: An illustrated guide to its structure, function, and disorders*. Penguin.

Chee, M. W., Soon, C. S., Lee, H. L., & Pallier, C. (2004). Left insula activation: A marker for language attainment in bilinguals. *Proceedings of the National Academy of Sciences*, *101*(42), 15265–15270.

Chen, C., Dupré la Tour, T., Gallant, J. L., Klein, D., & Deniz, F. (2024). The cortical representation of language timescales is shared between reading and listening. *Communications Biology*, *7*(1), 284.

Chen, L., Wassermann, D., Abrams, D. A., Kochalka, J., Gallardo-Diez, G., & Menon, V. (2019). The visual word form area (vwfa) is part of both language and attention circuitry. *Nature communications*, *10*(1), 5601.

Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M.-A., & Michel, F. (2000). The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, *123*(2), 291–307.

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

Comstock, L. (2024). The role of research design in the reproducibility of l1 and l2 language networks: A review of bilingual neuroimaging meta-analyses. *Brain and Language*, *249*, 105377.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *Neuroimage*, *9*(2), 179–194.

Das, T., Padakannaya, P., Pugh, K. R., & Singh, N. C. (2011). Neuroimaging reveals dual routes to reading in simultaneous proficient readers of two orthographies. *NeuroImage*, *54*(2), 1476–1487.

Decramer, T., Premereur, E., Uytterhoeven, M., Van Paesschen, W., van Loon, J., Janssen, P., & Theys, T. (2019). Single-cell selectivity and functional architecture of human lateral occipital complex. *PLoS biology*, *17*(9), e3000280.

Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The neuronal recycling hypothesis.

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in cognitive sciences*, *15*(6), 254–262.

Dehaene, S., Jobert, A., Naccache, L., Ciuciu, P., Poline, J.-B., Le Bihan, D., & Cohen, L. (2004). Letter binding and invariant recognition of masked words: Behavioral and neuroimaging evidence. *Psychological science*, *15*(5), 307–313.

Dehaene, S., Le Clec'H, G., Poline, J.-B., Le Bihan, D., & Cohen, L. (2002). The visual word form area: A prelexical representation of visual words in the fusiform gyrus. *Neuroreport*, *13*(3), 321–325.

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Filho, G. N., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, J., & Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *science*, *330*(6009), 1359–1364.

Dehaene-Lambertz, G., Monzalvo, K., & Dehaene, S. (2018). The emergence of the visual word form: Longitudinal evolution of category-specific ventral visual areas during reading acquisition. *PLoS biology*, *16*(3), e2004103.

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, *37*(27), 6539–6557.

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, *39*(39), 7722–7736.

Deniz, F., Tseng, C., Wehbe, L., la Tour, T. D., & Gallant, J. L. (2023). Semantic representations during language comprehension are affected by context. *Journal of Neuroscience*, *43*(17), 3144–3158.

DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences*, *109*(8), E505–E514.

Di Liberto, G. M., Wong, D., Melnik, G. A., & de Cheveigné, A. (2019). Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. *Neuroimage*, *196*, 237–247.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Ding, G., Perry, C., Peng, D., Ma, L., Li, D., Xu, S., Luo, Q., Xu, D., & Yang, J. (2003). Neural mechanisms underlying semantic and orthographic processing in chinese–english bilinguals. *NeuroReport*, *14*(12), 1557–1562.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, *19*(1), 158–164.

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in human neuroscience*, *8*, 311.

Fabbro, F. (2001). The bilingual brain: Bilingual aphasia. *Brain and language*, *79*(2), 201–210.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and brain sciences*, *39*, e229.

Forster, K. I. (1970). Visual perception of rapidly presented word sequences of varying complexity. *Perception & psychophysics*, *8*, 215–221.

Freeman, M. R., & Marian, V. (2022). Visual word recognition in bilinguals: Eye-tracking evidence that l2 proficiency impacts access of l1 phonotactics. *Studies in second language acquisition*, *44*(3), 759–787.

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological reviews*.

Friederici, A. D., Chomsky, N., Berwick, R. C., Moro, A., & Bolhuis, J. J. (2017). Language, mind and brain. *Nature human behaviour*, *1*(10), 713–722.

Gao, J. S., Huth, A. G., Lescroart, M. D., & Gallant, J. L. (2015). Pycortex: An interactive surface visualizer for fmri. *Frontiers in neuroinformatics*, *9*, 23.

Ghuman, A. S., & Martin, A. (2019). Dynamic neural representations: An inferential challenge for fmri. *Trends in Cognitive Sciences*, *23*(7), 534–536.

Gifford III, G. W., MacLean, K. A., Hauser, M. D., & Cohen, Y. E. (2005). The neurophysiology of functionally meaningful categories: Macaque ventrolateral prefrontal cortex plays a critical role in spontaneous categorization of species-specific vocalizations. *Journal of cognitive neuroscience*, *17*(9), 1471–1482.

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, *31*(3-4), 305–320.

Gong, X. L., Huth, A. G., Deniz, F., Johnson, K., Gallant, J. L., & Theunissen, F. E. (2023). Phonemic segmentation of narrative speech in human cerebral cortex. *Nature communications*, *14*(1), 4309.

Grainger, J., & Ziegler, J. C. (2011). A dual-route approach to orthographic processing. *Frontiers in psychology*, *2*, 54.

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), 536–548.

Grosjean, F. (1996). Gating. *Language and cognitive processes*, *11*(6), 597–604.

Grosjean, F. (2021). *Life as a bilingual: Knowing and using two or more languages*. Cambridge University Press.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS biology*, *11*(12), e1001752.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics*, *31*(3-4), 423–445.

Hagoort, P., & Indefrey, P. (2014). The neurobiology of language beyond single words. *Annual review of neuroscience*, *37*, 347–362.

Hamilton, L. S., Oganian, Y., Hall, J., & Chang, E. F. (2021). Parallel and distributed encoding of speech across human auditory cortex. *Cell*, *184*(18), 4626–4639.

Hasson, U., Levy, I., Behrmann, M., Hendler, T., & Malach, R. (2002). Eccentricity bias as an organizing principle for human high-order object areas. *Neuron*, *34*(3), 479–490.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119.

Henson, R., & Friston, K. (2007). Convolution models for fmri. *Statistical parametric mapping: The analysis of functional brain images*, 178–192.

Herbet, G., Zemmoura, I., & Duffau, H. (2018). Functional anatomy of the inferior longitudinal fasciculus: From historical reports to current hypotheses. *Frontiers in neuroanatomy*, *12*, 77.

Hesse, J. K., & Tsao, D. Y. (2016). Consistency of border-ownership cells across artificial stimuli, natural stimuli, and stimuli with ambiguous contours. *Journal of Neuroscience*, *36*(44), 11338–11349.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature reviews neuroscience*, *13*(2), 135–145.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews neuroscience*, *8*(5), 393–402.

Hickok, G., & Poeppel, D. (2016). Neural basis of speech perception. *Neurobiology of language*, 299–310.

Hsu, A., Borst, A., & Theunissen, F. E. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network: Computation in Neural Systems*, *15*(2), 91–109.

Huang, C., Lorusso, M. L., Luo, Z., & Zhao, J. (2019). Developmental differences in the relationship between visual attention span and chinese reading fluency. *Frontiers in psychology*, *10*, 475862.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, *160*(1), 106.

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., & Chang, E. F. (2016). Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *Journal of Neuroscience*, *36*(6), 2014–2026.

Humphries, C., Love, T., Swinney, D., & Hickok, G. (2005). Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. *Human brain mapping*, *26*(2), 128–138.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *Neuroimage*, *62*(2), 782–790.

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical image analysis*, *5*(2), 143–156.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644.

Klimovich-Gray, A., Tyler, L. K., Randall, B., Kocagoncu, E., Devereux, B., & Marslen-Wilson, W. D. (2019). Balancing prediction and sensory input in speech comprehension: The spatiotemporal dynamics of word recognition in context. *Journal of Neuroscience*, *39*(3), 519–527.

Ko, H.-k., & von der Heydt, R. (2018). Figure-ground organization in the visual cortex: Does meaning matter? *Journal of neurophysiology*, *119*(1), 160–176.

Kourtzi, Z., & Connor, C. E. (2011). Neural representations for object perception: Structure, category, and adaptive coding. *Annual review of neuroscience*, *34*, 45–67.

Kumar, U., Das, T., Bapi, R. S., Padakannaya, P., Joshi, R. M., & Singh, N. C. (2010). Reading different orthographies: An fmri study of phrase reading in hindi–english bilinguals. *Reading and Writing*, *23*, 239–255.

Larsson, J., & Heeger, D. J. (2006). Two retinotopic visual areas in human lateral occipital cortex. *Journal of neuroscience*, *26*(51), 13128–13142.

la Tour, T. D., Eickenberg, M., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *bioRxiv*.

Lawrence, S. J., Zamboni, E., Vernon, R. J., Gouws, A. D., Wade, A. R., & Morland, A. B. (2023). The emergence of tuning to global shape properties of radial frequency patterns in the ventral visual pathway. *Journal of Neuroscience*, *43*(29), 5378–5390.

Lenzo, K., et al. (2007). The cmu pronouncing dictionary. *Carnegie Melon University*, *313*, 74.

Leonard, M. K., Bouchard, K. E., Tang, C., & Chang, E. F. (2015). Dynamic encoding of speech sequence probability in human temporal cortex. *Journal of Neuroscience*, *35*(18), 7203–7214.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906–2915.

Lescroart, M. D., & Gallant, J. L. (2019). Human scene-selective areas represent 3d configurations of surfaces. *Neuron*, *101*(1), 178–192.

Li, H., Zhang, J., & Ding, G. (2021). Reading across writing systems: A meta-analysis of the neural correlates for first and second language reading. *Bilingualism: Language and cognition*, *24*(3), 537–548.

Li, P., Zhang, F., Yu, A., & Zhao, X. (2020). Language history questionnaire (lhq3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition*, *23*(5), 938–944.

Liu, H., & Cao, F. (2016). L1 and l2 processing in the bilingual brain: A meta-analysis of neuroimaging studies. *Brain and language*, *159*, 60–73.

Lobier, M., Dubois, M., & Valdois, S. (2013). The role of visual processing speed in reading speed development. *PloS one*, *8*(4), e58097.

Lobier, M., Zoubrinetzky, R., & Valdois, S. (2012). The visual attention span deficit in dyslexia is visual and not verbal. *cortex*, *48*(6), 768–773.

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, *115*(38), E9015–E9024.

Lukatela, G., & Turvey, M. T. (1994). Visual lexical access is initially phonological: 2. evidence from phonological priming by homophones and pseudohomophones. *Journal of Experimental Psychology: General*, *123*(4), 331.

Malik-Moraleda, S., Mahowald, K., Conway, B. R., & Gibson, E. (2023). Concepts are restructured during language contact: The birth of blue and other color concepts in tsimane'-spanish bilinguals. *Psychological Science*, *34*(12), 1350–1362.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (leap-q): Assessing language profiles in bilinguals and multilinguals.

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human perception and performance*, *15*(3), 576.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1-2), 71–102.

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in cognitive sciences*, *7*(7), 293–299.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological review*, *88*(5), 375.

McLaren, K. (1976). Xiii—the development of the cie 1976 (l* a* b*) uniform colour space and colour-difference formula. *Journal of the Society of Dyers and Colourists*, *92*(9), 338–341.

McNealy, K., Mazziotta, J. C., & Dapretto, M. (2006). Cracking the language code: Neural mechanisms underlying speech parsing. *Journal of Neuroscience*, *26*(29), 7629–7639.

Meschyan, G., & Hernandez, A. E. (2006). Impact of language proficiency and orthographic transparency on bilingual word reading: An fmri investigation. *NeuroImage*, *29*(4), 1135–1140.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010.

Mooijman, S., Schoonen, R., Roelofs, A., & Ruiter, M. B. (2022). Executive control in bilingual aphasia: A systematic review. *Bilingualism: Language and cognition*, *25*(1), 13–28.

Mumford, J. A., & Nichols, T. E. (2008). Power calculation for group fmri studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage*, *39*(1), 261–268.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, *21*(19), 1641–1646.

Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage*, *197*, 482–492.

Oganian, Y., & Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Science advances*, *5*(11), eaay6279.

Öhman, S. E. (1966). Coarticulation in vcv utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, *39*(1), 151–168.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Oliphant, T. E. (2007). Python for scientific computing. *Computing in science & engineering*, *9*(3), 10–20.

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

Perani, D., & Abutalebi, J. (2005). The neural basis of first and second language processing. *Current opinion in neurobiology*, *15*(2), 202–206.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific studies of reading*, *11*(4), 357–383.

Perfetti, C. A., & Liu, Y. (2005). Orthography to phonology and meaning: Comparisons across and within writing systems. *Reading and Writing*, *18*, 193–210.

Petrides, M. (2016). The ventrolateral frontal region. In *Neurobiology of language* (pp. 25–33). Elsevier.

Piai, V., Anderson, K. L., Lin, J. J., Dewar, C., Parvizi, J., Dronkers, N. F., & Knight, R. T. (2016). Direct brain recordings reveal hippocampal rhythm underpinnings of language processing. *Proceedings of the National Academy of Sciences*, *113*(40), 11366–11371.

Pierce, L. J., Chen, J.-K., Delcenserie, A., Genesee, F., & Klein, D. (2015). Past experience shapes ongoing neural patterns for language. *Nature Communications*, *6*(1), 10073.

Poeppel, D., Emmorey, K., Hickok, G., & Pylkkänen, L. (2012). Towards a new neurobiology of language. *Journal of Neuroscience*, *32*(41), 14125–14131.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature reviews neuroscience*, *18*(2), 115–126.

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature neuroscience*, *24*(11), 1628–1636.

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*, *1191*(1), 62–88.

Price, C. J. (2012). A review and synthesis of the first 20 years of pet and fmri studies of heard speech, spoken language and reading. *Neuroimage*, *62*(2), 816–847.

Puce, A., Allison, T., Asgari, M., Gore, J. C., & McCarthy, G. (1996). Differential sensitivity of human visual cortex to faces, letterstrings, and textures: A functional magnetic resonance imaging study. *Journal of neuroscience*, *16*(16), 5205–5215.

Räsänen, O., Doyle, G., & Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, *171*, 130–150.

Roe, A. W., Chelazzi, L., Connor, C. E., Conway, B. R., Fujita, I., Gallant, J. L., Lu, H., & Vanduffel, W. (2012). Toward a unified theory of visual area v4. *Neuron*, *74*(1), 12–29.

Saygin, Z. M., Osher, D. E., Norton, E. S., Youssoufian, D. A., Beach, S. D., Feather, J., Gaab, N., Gabrieli, J. D., & Kanwisher, N. (2016). Connectivity precedes function in the development of the visual word form area. *Nature neuroscience*, *19*(9), 1250–1255.

Shen, Y., & Del Tufo, S. N. (2022). The influence of orthographic depth on multilinguals' neural networks. *Neuropsychologia*, *164*, 108095.

Silson, E. H., McKeefry, D. J., Rodgers, J., Gouws, A. D., Hymers, M., & Morland, A. B. (2013). Specialized and independent processing of orientation and shape in visual field maps lo1 and lo2. *Nature neuroscience*, *16*(3), 267–269.

Sulpizio, S., Del Maschio, N., Fedeli, D., & Abutalebi, J. (2020). Bilingual language processing: A meta-analysis of functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, *108*, 834–853.

Szwed, M., Qiao, E., Jobert, A., Dehaene, S., & Cohen, L. (2014). Effects of literacy in early visual and occipitotemporal areas of chinese and french readers. *Journal of cognitive neuroscience*, *26*(3), 459–475.

Tian, M., Li, H., Chu, M., & Ding, G. (2020). Functional organization of the ventral occipitotemporal regions for chinese orthographic processing. *Journal of Neurolinguistics*, *55*, 100909.

Toga, A. W., & Thompson, P. M. (2003). Mapping brain asymmetry. *Nature Reviews Neuroscience*, *4*(1), 37–48.

Trauzettel-Klosinski, S., Dietz, K., Group, I. S., et al. (2012). Standardized assessment of reading performance: The new international reading speed texts irest. *Investigative ophthalmology & visual science*, *53*(9), 5452–5461.

Trecca, F., Tylén, K., Fusaroli, R., Johansson, C., & Christiansen, M. H. (2019). Top-down information is more important in noisy situations: Exploring the role of pragmatic, semantic, and syntactic information in language processing.

Turkeltaub, P. E., & Coslett, H. B. (2010). Localization of sublexical speech perception components. *Brain and language*, *114*(1), 1–15.

Ungerleider, S. K., & G, L. (2000). Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, *23*(1), 315–341.

Vaden, K. I., Halpin, H., & Hickok, G. S. (2009). Irvine phonotactic online dictionary, version 2.0.[data file].

Valdois, S. (2022). The visual-attention span deficit in developmental dyslexia: Review of evidence for a visual-attention-based deficit. *Dyslexia*, *28*(4), 397–415.

Van de Putte, E., De Baene, W., Brass, M., & Duyck, W. (2017). Neural overlap of l1 and l2 semantic representations in speech: A decoding approach. *NeuroImage*, *162*, 106–116.

Van Orden, G. C. (1987). A rows is a rose: Spelling, sound, and reading. *Memory & cognition*, *15*, 181–198.

Venezia, J. H., Thurman, S. M., Richards, V. M., & Hickok, G. (2019). Hierarchy of speech-driven spectrotemporal receptive fields in human auditory cortex. *Neuroimage*, *186*, 647–666.

Watson, A. B., & Ahumada, A. J. (1985). Model of human visual-motion sensing. *JOSA A*, *2*(2), 322–342.

Wiley, R. W., & Rapp, B. (2019). From complexity to distinctiveness: The effect of expertise on letter perception. *Psychonomic Bulletin & Review*, *26*, 974–984.

Wu, X., Kong, X., Vatansever, D., Liu, Z., Zhang, K., Sahakian, B. J., Robbins, T. W., Feng, J., Thompson, P., & Zhang, J. (2022). Dynamic changes in brain lateralization correlate with human cognitive performance. *PLoS biology*, *20*(3), e3001560.

Xu, M., Baldauf, D., Chang, C. Q., Desimone, R., & Tan, L. H. (2017). Distinct distributed patterns of neural activity are associated with two languages in the bilingual brain. *Science advances*, *3*(7), e1603309.

Yang, S., Cai, Y., Xie, W., & Jiang, M. (2021). Semantic and syntactic processing during comprehension: Erp evidence from chinese qing structure. *Frontiers in Human Neuroscience*, *15*, 701923.

Zhan, M., Pallier, C., Agrawal, A., Dehaene, S., & Cohen, L. (2023). Does the visual word form area split in bilingual readers? a millimeter-scale 7-t fmri study. *Science advances*, *9*(14), eadf6140.

Zhang, J., Li, H., Zhang, M., Wang, Z., Ao, X., Jian, J., Wei, N., Liu, H., Ding, G., & Meng, X. (2023). Functional preference of the left inferior parietal lobule to second language reading. *NeuroImage*, *270*, 119989.

Zhou, H., Friedman, H. S., & Von Der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *Journal of Neuroscience*, *20*(17), 6594–6611.

Zhuang, J., Randall, B., Stamatakis, E. A., Marslen-Wilson, W. D., & Tyler, L. K. (2011). The interaction of lexical semantics and cohort competition in spoken word recognition: An fmri study. *Journal of Cognitive Neuroscience*, *23*(12), 3778–3790.

Appendix A

Supplements for "Phonemic segmentation"

A.1 Supplementary Figures

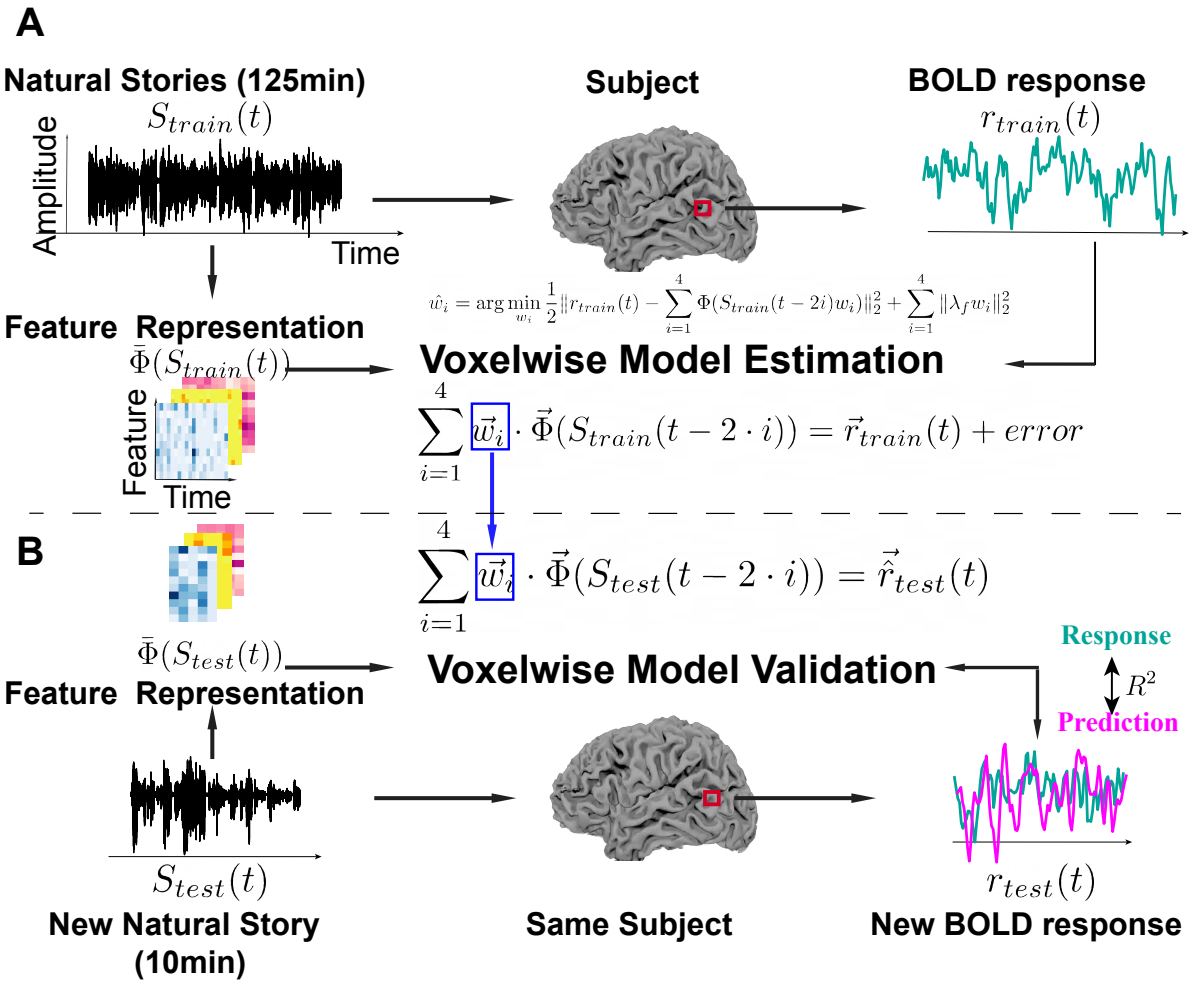
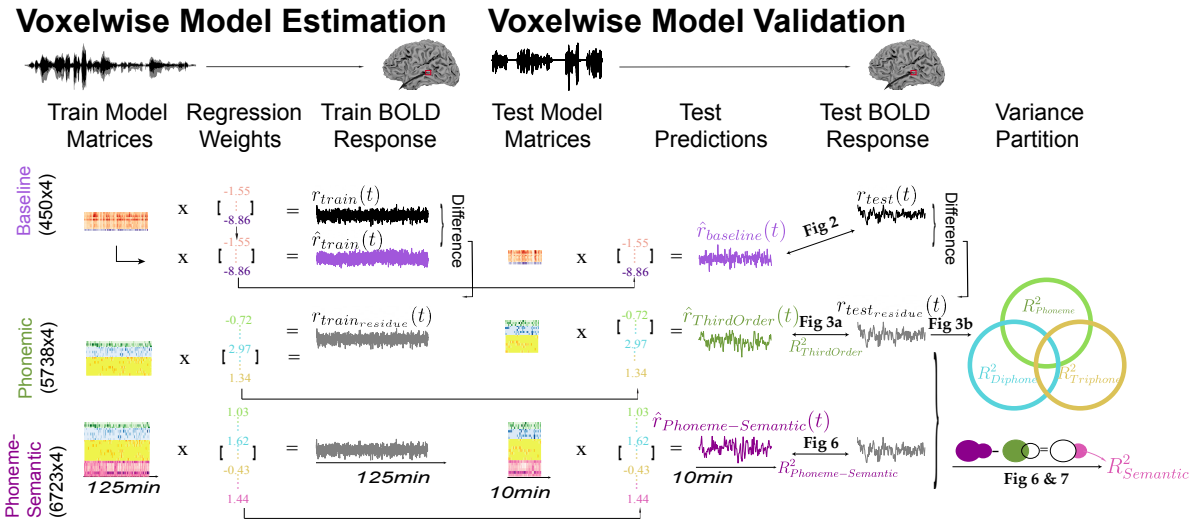
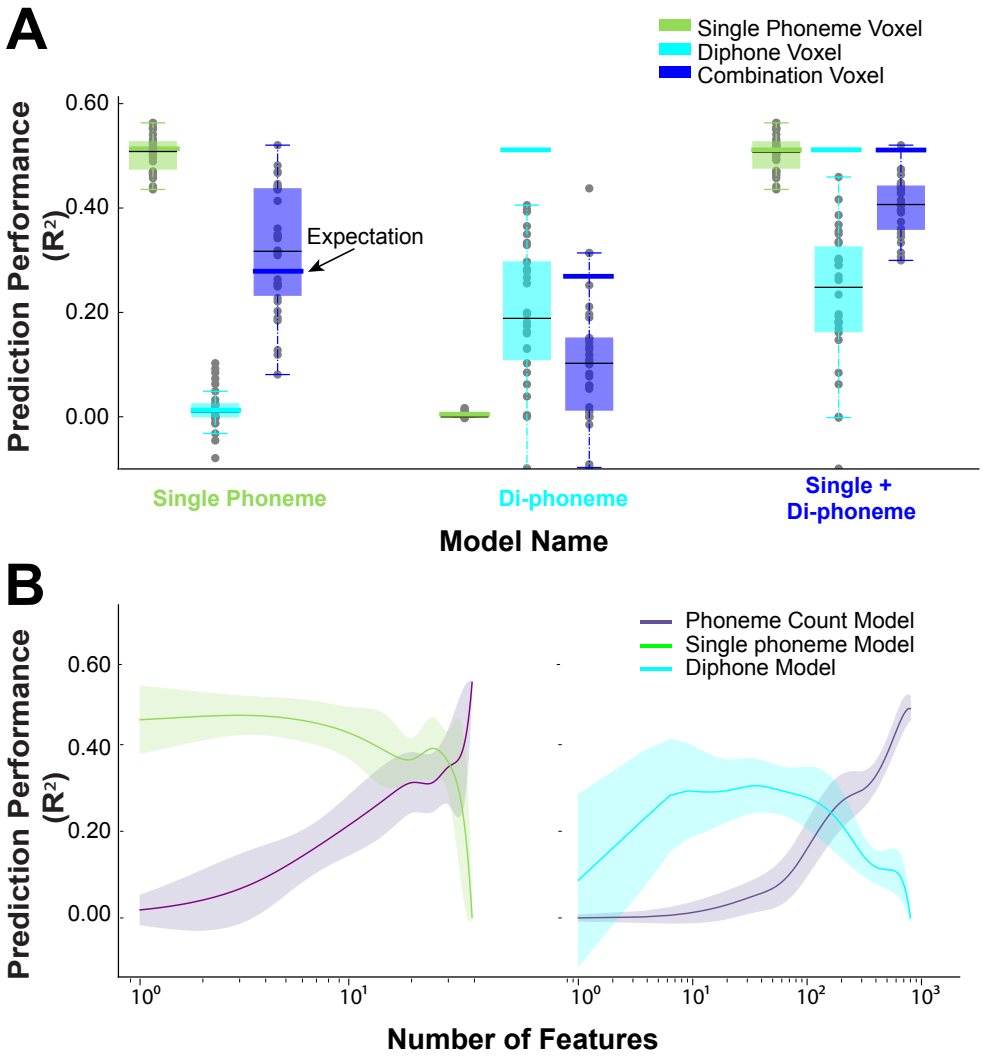


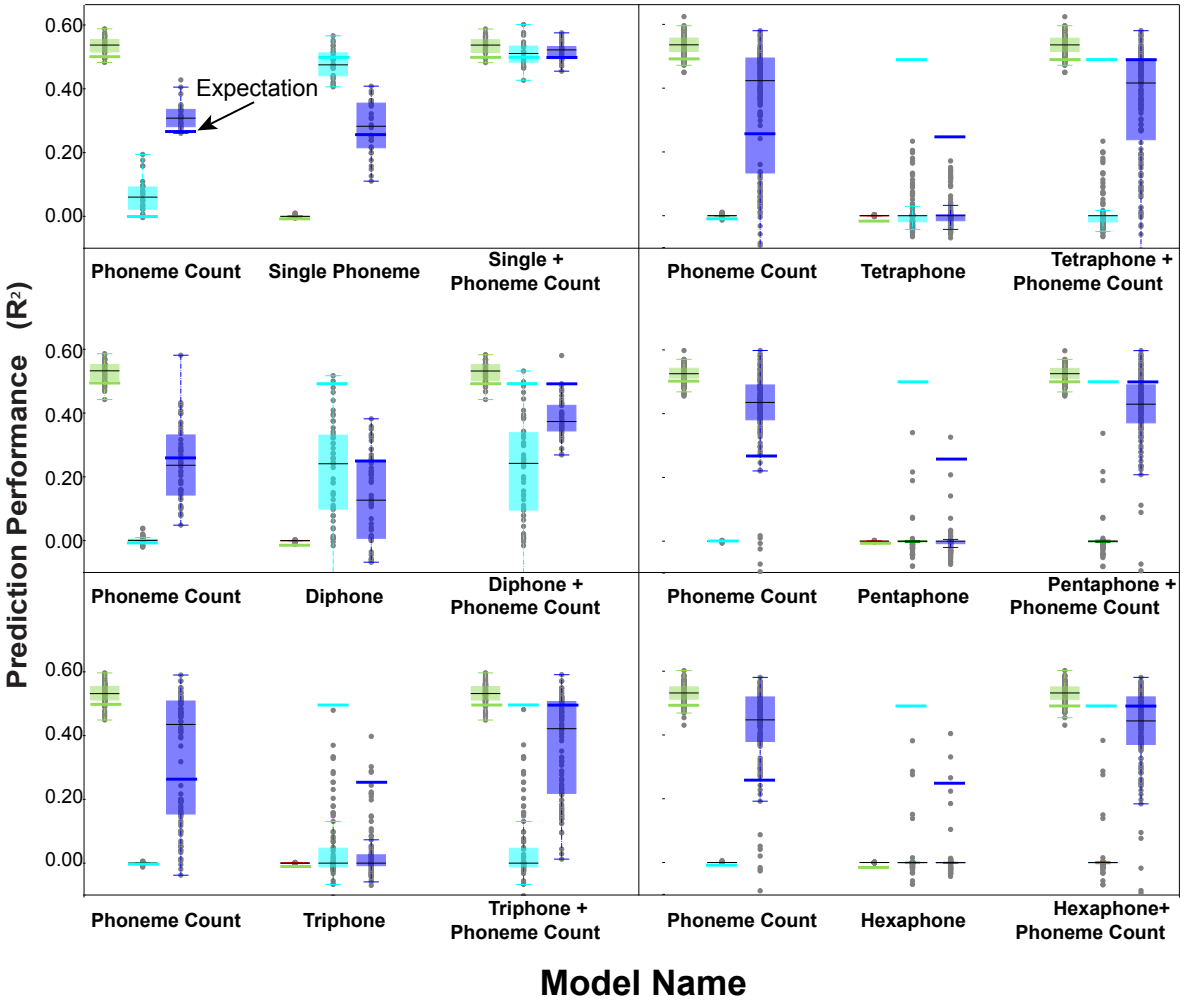
Figure A.1: **A. Voxelwise modeling (VM) estimation:** Regularized linear regression was used to predict BOLD activity in individual voxels. Regression weights ($\bar{w}$) were estimated for each voxel and each feature space. The linear regression used the stimulus features in the four time windows preceding the BOLD activity at time t; the time windows are $2sec$ long, corresponding to the scanner TR (0.5Hz). B. **Voxelwise Model Validation:** The quality of the models' predictions was assessed using one $10-min$ story not included during model estimation (validation story). The BOLD (green) response to this validation story (green) was compared to the prediction (magenta) to calculate the cross-validated R^2.



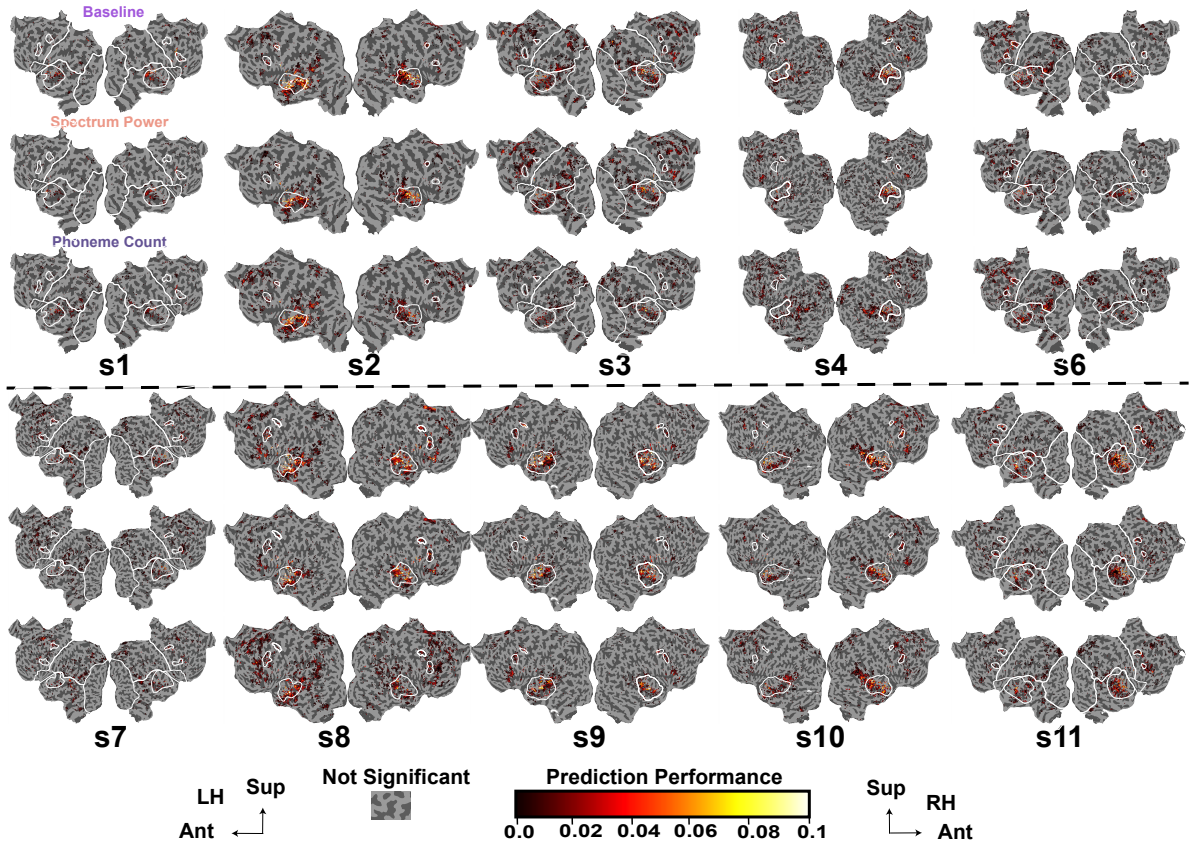
Supplementary Figure A.2: **Variance Partitioning.** Variance partitioning was used to obtain the BOLD response uniquely explained by different phonemic and semantic features. Before fitting models that capture particular phonemes or semantic meanings, the fraction of the BOLD response predicted by the baseline model was subtracted, in order to eliminate the BOLD responses explained by the mere presence versus absence of auditory speech stimuli. We then fitted a joint phonemic model with all phonemic features (single phonemes, diphones and triphones) and a joint phonemic-semantic model with all the phonemic and semantic features. Afterwards, we used variance partitioning to obtain unique variance explained by each phonemic feature and all the possible combinations of these phonemic features (single phonemes + diphones, diphones + triphones, single phonemes + triphone, single phonemes + diphones + triphones). The unique variance explained by semantic feature is obtained by subtracting from the variance explained of phonemic-semantic model from that of phonemic models.



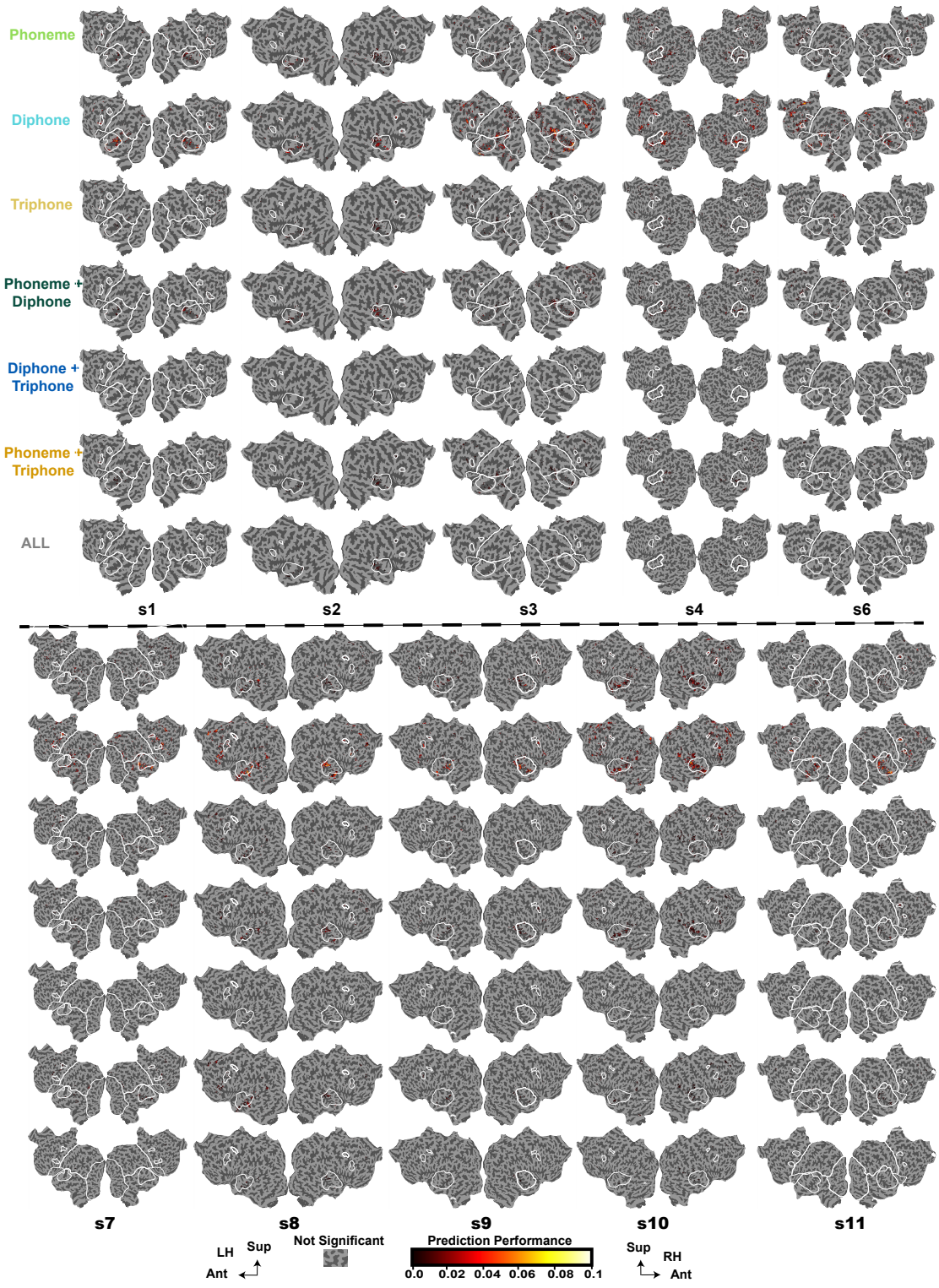
Supplementary Figure A.3: **Validation by simulations: effect of voxel sensitivity. Panel A.** The coefficient of determination, R^2, was calculated for model voxels sensitive either to three single phonemes (green), three diphones (cyan) or the combination of three single and three single diphones (blue). For each voxel (n=1), we repeated this procedure 30 times (Y=30 independent experiments) to generate a distribution of simulated results. The three groups on the x-axis correspond to the VMs based on single phonemes features, diphone features or both. The solid bold colored lines (labeled as Expectation) show the ground truth. **Panel B.** The simulation for single phoneme (left panel) and diphone (right panel) voxels was repeated by systematically varying the rate of phonemic units (x-axis) with non-zero weights; from highly sensitive model voxels that respond 1 phonemic unit to less sensitive units that respond to 10s or 100s of phones and diphones. To assess the sensitivity of our method (given our recording time of 3737 TRs), we compared the prediction obtained from the correct phonemic identity VM (single phoneme VM, red line for left plot and diphone VM, green line for right plot) to the VM based solely on the phoneme count (purple line). Error bars are SEM obtained by repeating the simulation by using different random samples of phonemic units with non-zero weights.



Supplementary Figure A.4: **Validation by simulation: sensitivity to longer combinations of phonemes.** To explore the limits of phonemic combinations that could be recovered from our dataset, we also simulated responses to higher order combinations. In this figure (similarly to Figure Supplementary Figure 3A), each subpanel shows the results obtained applying nested VM to sets of three putative voxels: one sensitive to the phoneme count, one to different phonemic combinations (single phoneme, diphone, triphone, tetraphone, pentaphone, and hexaphone) and one to the mixture of both. For each voxel (n=1), we repeated this procedure 30 times for single phoneme simulation, 50 times for diphones, and 150 times for triphone, tetraphone, pentaphone and hexaphone based VMs to obtain the distribution of the prediction performance. As shown in the plot, given the measured SNR and about 2 hours of data, one cannot recover the signals of putative voxels sensitive for phonemic combinations beyond the triphone.



Supplementary Figure A.5: **Prediction performance of auditory processing on flat maps for the remaining** 10 **subjects.** The cortical flatmaps on the first rows show the significant prediction performance of the Baseline VM model for all the subjects in the study (Subject 5 is shown on the main Figure 2). The Baseline VM uses the time varying power spectrum and the phoneme count as features. The second and third row show the additive contribution of the spectrum and phoneme count features to this auditory Baseline VM.



Supplementary Figure A.6: **Prediction performance of phonemic processing on flat maps for the remaining** 10 **subjects.** Each column of the cortical flatmaps shows the significant prediction performance of the unique contributions from single phoneme, diphone, triphone, phoneme + diphone, diphone + triphone, phoneme + triphone and single phoneme + diphone + triphone features using variance partitioning for each subject. (Subject 5 is shown on the main Figure 3).

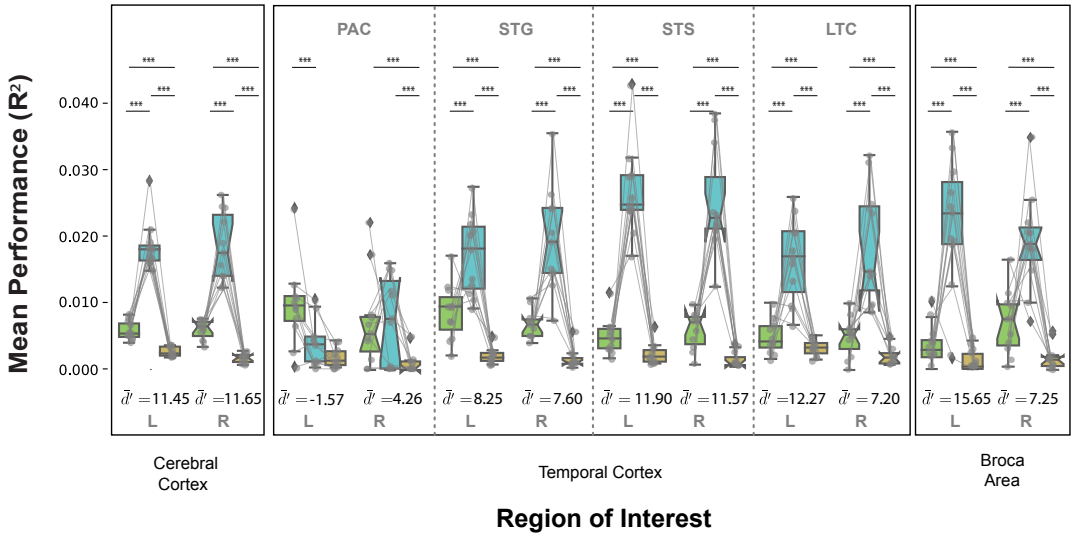


Figure A.7: **Phonemic processing: hemispheric analysis.** In order to examine whether the significant prediction performance from single phoneme, diphone and triphone features varied significantly between left (box) and right (notched box) hemispheres across regions of interest, we performed a linear mixed-effect model with post-hoc Tukey test. The statistics are derived from the performance of significant phonemic voxels for each ROI (n $=0-1,973$ voxels) over 11 independent subjects. All tests are two sided and corrected for multiple comparisons. The exact p values can be found in the Laterality of the phonemic representation and segmentation. section. The boxplots are defined the same way as in Figure 4 in the manuscript. The dots show the data of individual subjects with grey dots (*p<0.05, **p<0.01 and ***p<0.001). In STG, STS and Broca's area, the diphone feature space explains significantly more variance than single phoneme or triphone features in the left hemisphere than that in the right hemisphere. It indicates that the diphone segmentation for phonemic processing is more prominent in the left hemisphere than the right hemisphere.

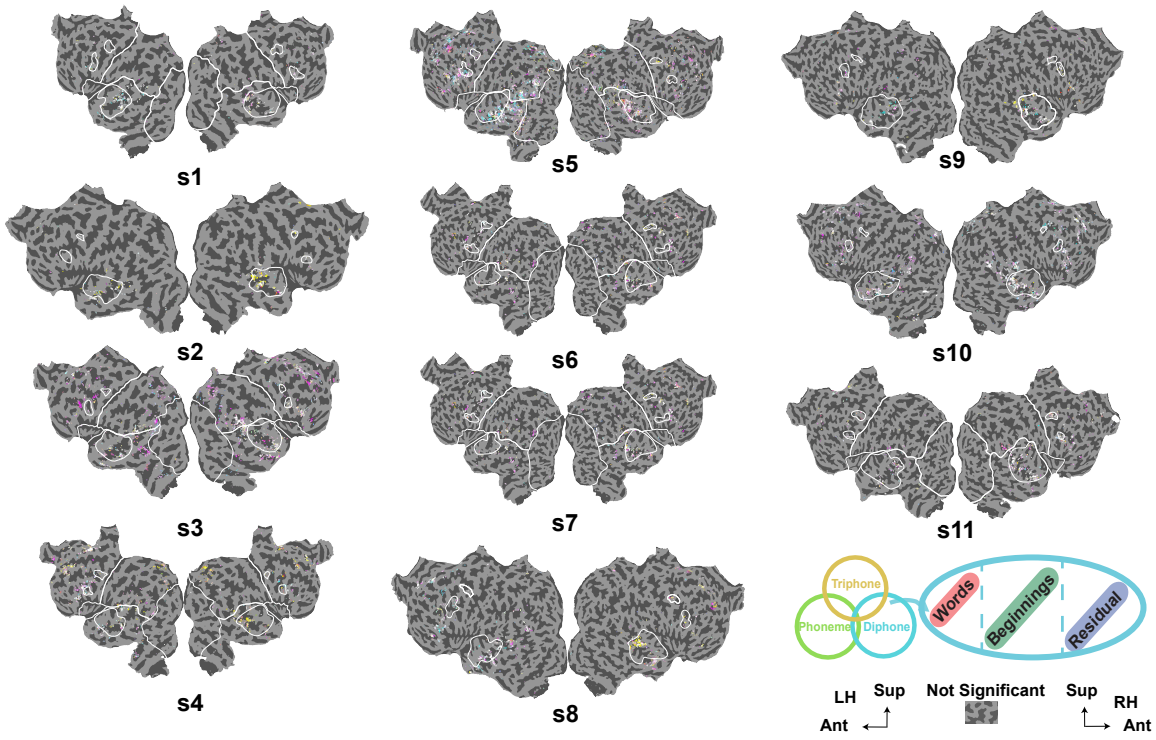
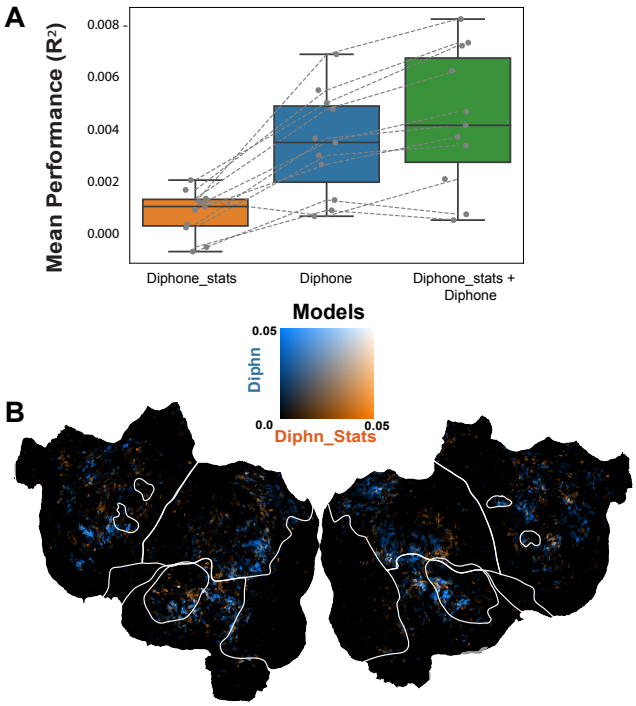
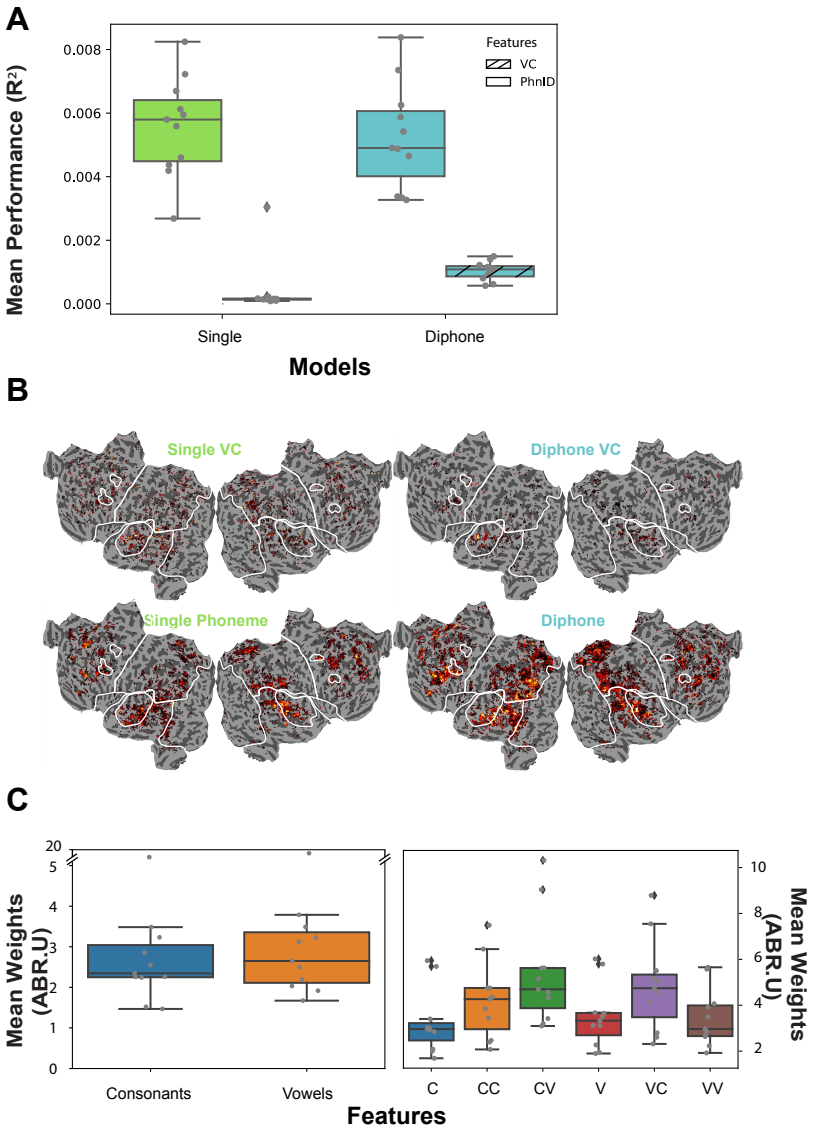


Figure A.8: **Prediction performance of distinct diphone categories for** 11 **subjects.** The cortical flatmaps show the significant prediction performance of the short words (red), word beginnings (green) and residuals (blue). The flatmaps did not show distinct cortical subregions within cortical areas with phonemic representations where the responses to short words were systematically higher than those to word beginnings of other diphones



Supplementary Figure A.9: **Modeling on diphoneme statistical properties.** In order to further explore if the cortical BOLD activities truly encode the content of diphones or merely the statistical properties of diphones, we built one diphone statistics model consisting of 8 phonological statistical features extracted from Irvine Phonotactic Online Dictionary (Vaden et al., 2009). These features describe the phonotactic probability of diphones (Supplementary Table 6). **Panel A** shows the average prediction performance of all the significant cerebral cortex voxels of 11 subjects (each dot represents one subject) from the diphone statistics model (orange), the diphone identity model (blue), and both models (green). It shows that the performance of the diphone statistics model is significantly lower than that of the diphone identity model (blue) ($t(1) = -30.19, p < 2.2^{10-16}$). All tests are two sided and corrected for multiple comparisons. The boxplots are defined the same way as in Figure 4 in the manuscript. **Panel B** shows the anatomical data for one example subject (S5). Blue voxels are better predicted by the diphone (content) model, and orange voxels are better predicted by the diphone statistics model. White voxels are equally well predicted by both models. This analysis shows that although diphone statistical properties can explain some of the variance of BOLD responses that could be captured by the diphone model, the actual diphone identities captured in the diphone model and not in the phonotactic probabilities yield significant additional explanatory power.



Supplementary Figure A.10: **Modeling on vowel and consonants.** In order to explore how the properties of vowel and consonants contribute to the cerebral cortical encoding of phonemes and phonemic combinations, we calculated the prediction performance based on the identity of consonants and vowels of single phonemes and diphones (Supplementary Methods). **Panel A** shows that average prediction performance for each subject (each dot) for each model. **Panel B** shows the additive prediction performance of each model for each voxel on a flattened cortical map of subject 5. It illustrates that the phoneme identity models (single phoneme and diphone features) significantly outperform the vowel and consonants features. **Panel C** shows the average (across voxels and subjects) of the absolute value of the weights for each regressor/channel of the single vc and diphone vc features. There is no significant difference in the contribution consonants and vowels towards the single phoneme encoding. In addition, "vc" and "cv" combinations contribute the most to the diphone encoding. For both panel A and C, the statistics are derived from the performance of significant voxels (n $=2,409-11,668$ voxels) over 11 independent subjects. The boxplots are defined the same way as in Figure 4 in the manuscript.

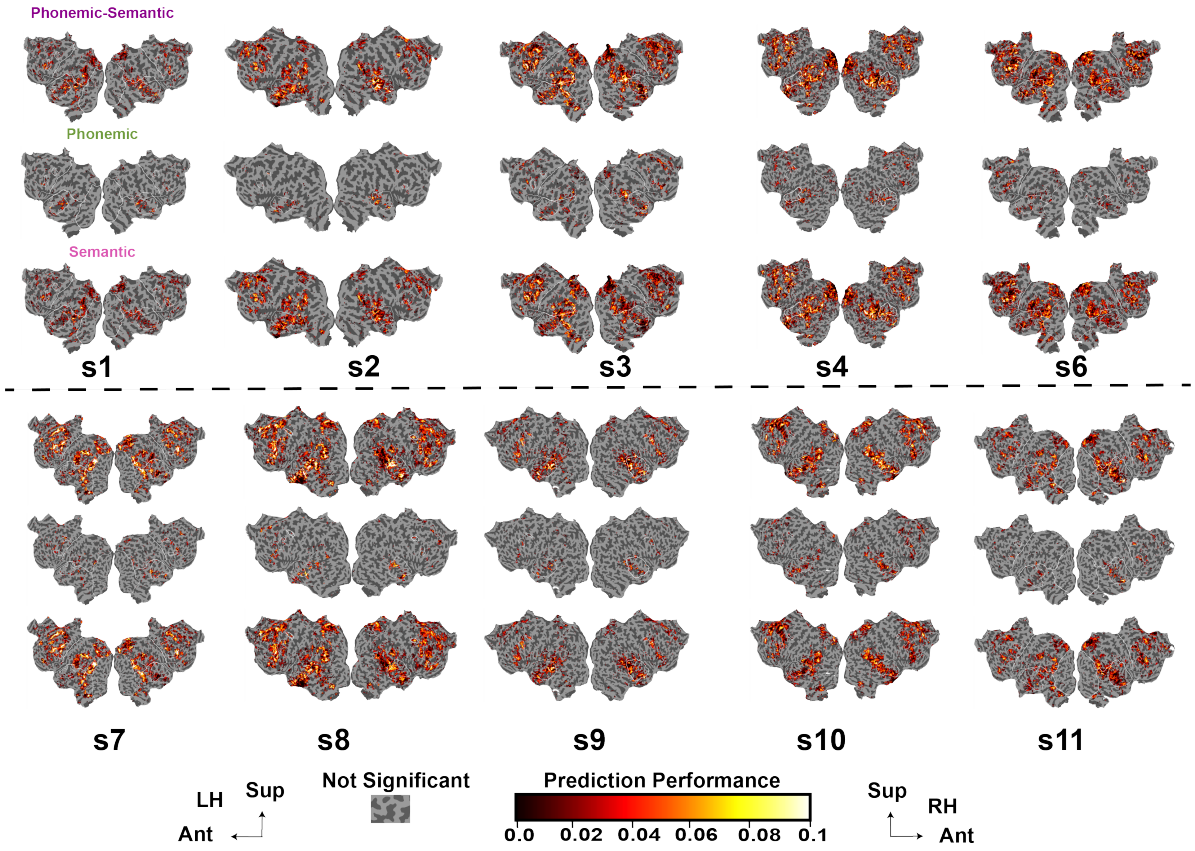
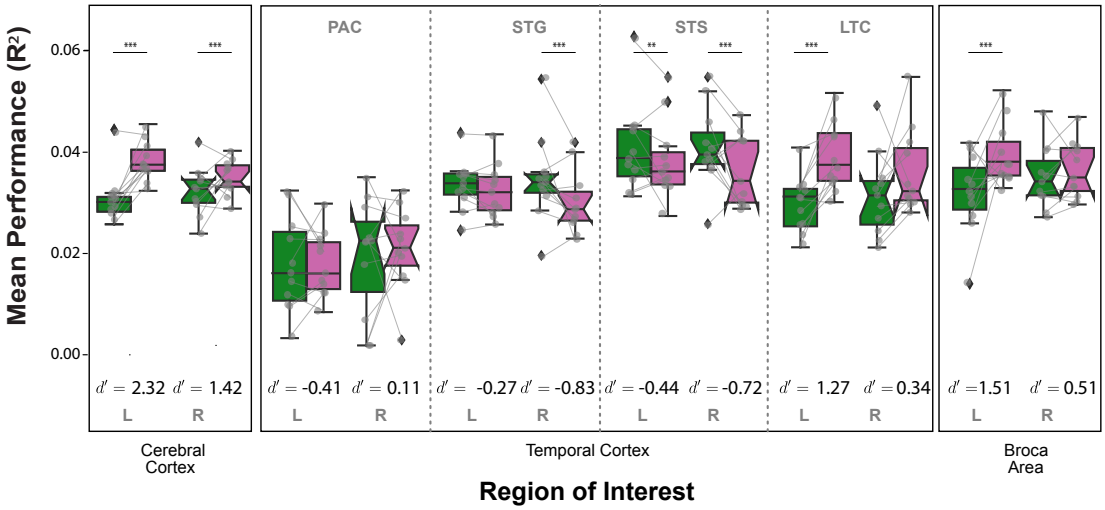


Figure A.11: **Prediction performance of phonemic-semantic cortical maps for the remaining** 10 **subjects.** The cortical flatmaps on the first row show the significant prediction performance of the Phonemic-Semantic VM model for all the subjects in the study (Subject 5 is shown on the main Figure 6). The Phonemic-Semantic VM uses the combination of the phonemic identities and semantic embeddings as features. The second and third row show the additive contribution of phonemic and semantic features to the Phonemic-Semantic VM.



Supplementary Figure A.12: **Semantic vs phonemic contributions: hemispheric analysis.** In order to examine whether the significant prediction performance from the phonemic model and semantic features are varied between left (box) and right (notched box) hemisphere across regions of interest, we performed a linear mixed-effect model with post-hoc Tukey test. The statistics are derived from the performance of significant phoneme-semantic voxels for each ROI (n $=56-19,511$ voxels) over 11 independent subjects. All tests are two sided and corrected for multiple comparisons. The exact p values can be found in the Laterality of phonemic versus semantic representations. section. The boxplots are defined the same way as in Figure 4 in the manuscript. In STG, phonemic model predicts significantly higher response in the right hemisphere, while it explains significantly more in STS of both hemispheres. In LTC and Broca's area, semantic features predict significantly higher response in the left hemisphere, while it explains significantly more variance of both cerebral cortical hemispheres. It indicates that the difference in semantic and phonemic processing is more prominent in the right hemisphere in STG and in the left hemiphere in LTC and Broca's area, while this difference is more balanced in the PAC and STS of both hemispheres.

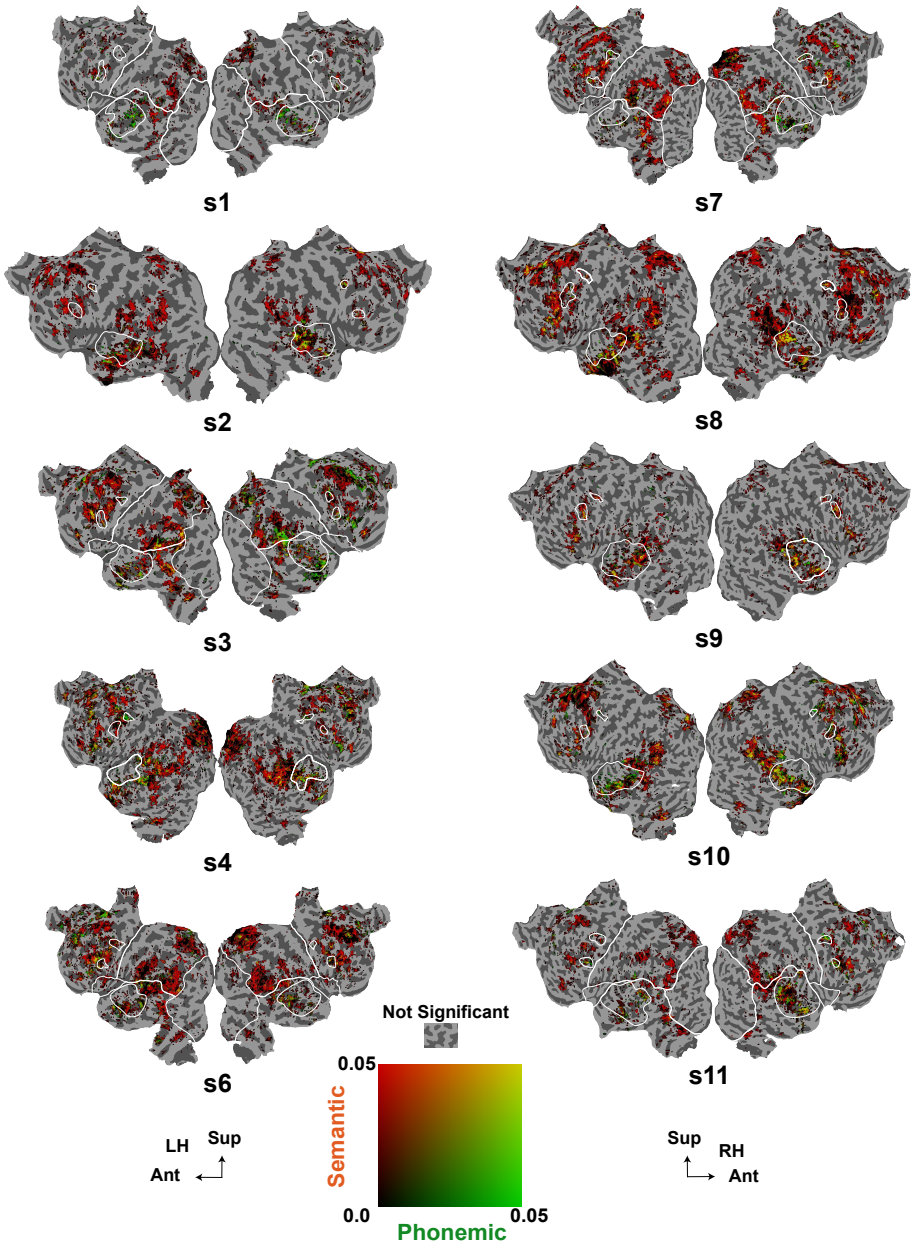
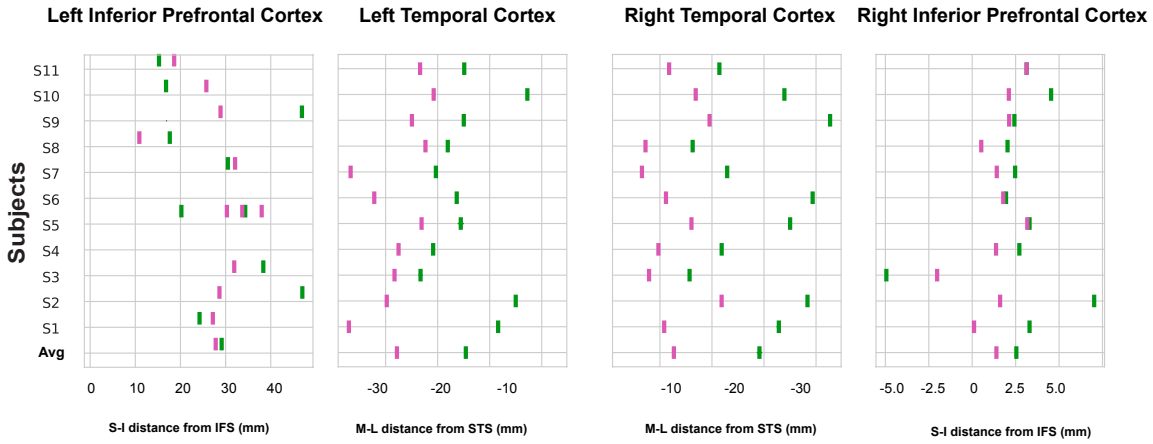


Figure A.13: **Flatmap of individual subject for prediction performance of phonemic versus semantic based VM.** The cortical flatmaps show the significant prediction performance of the phonemic VM versus semantic VM for all the subjects in the study (Subject 5 is shown on the main Figure 7). Green voxels are better predicted by phonemic model, while red voxels by semantic model. Yellow voxels share equally good prediction performance from phonemic and semantic models. These figures indicate two gradients of phonemic to semantic cortical representations. One in the temporal cortex, and another in the IPFC.



Supplementary Figure A.14: **Anatomical location of the center of mass for individual subjects.** To quantify the phonemic to semantic transition, the center of mass of voxels in temporal cortex and IPFC in both hemispheres with significant prediction performance from phonemic versus semantic features for each subject is shown. It indicates that there is a medial/lateral gradient for phonemic to semantic transition in the temporal cortex of both hemispheres.

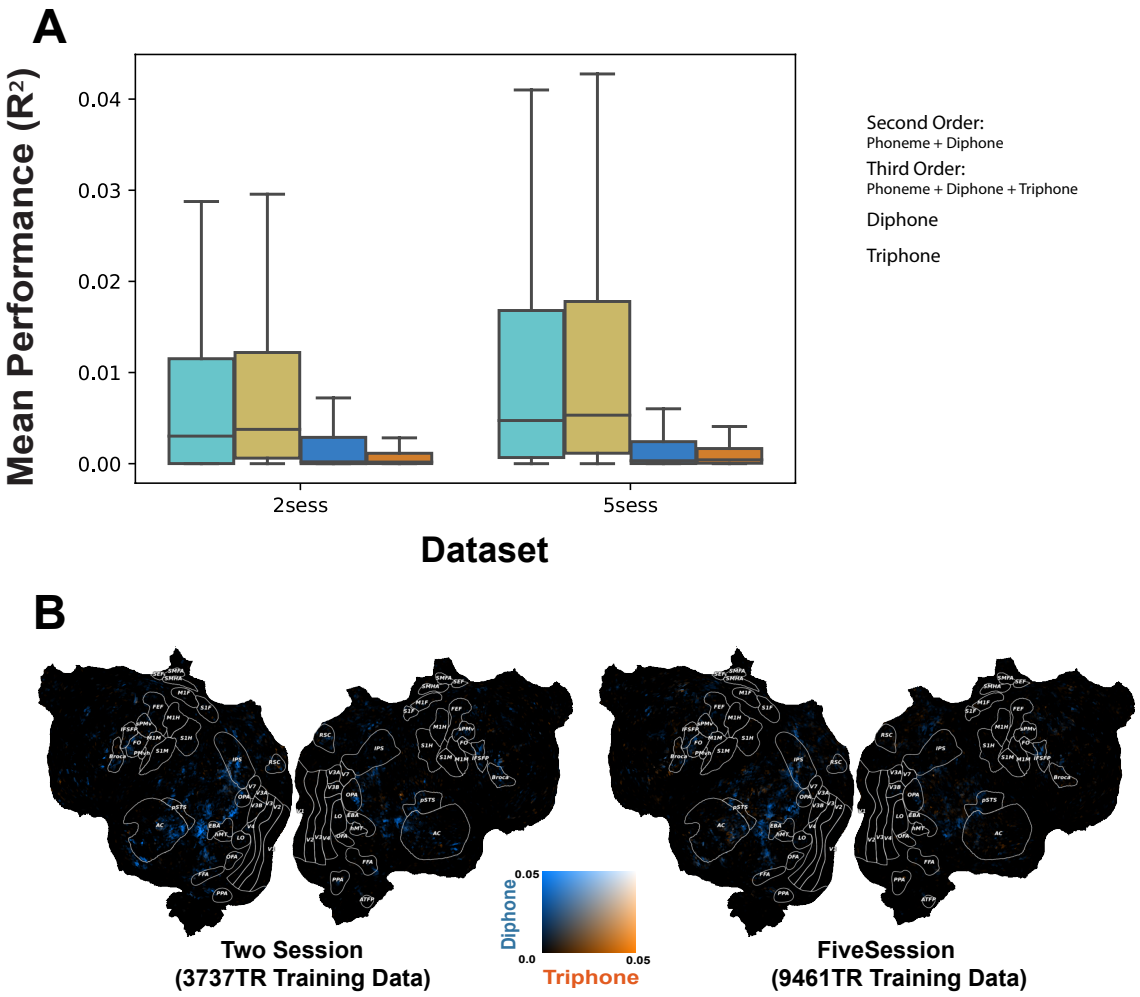


Figure A.15: **Prediction Performance from five sessions of data**. **Panel A** shows the average prediction performance (R^2) across significant voxels obtained from second order phonemic model (single phoneme + diphone: cyan), third order phonemic model (single phoneme + diphone + triphone: yellow), diphone only (blue) and triphone only (orange). The boxplots are defined the same way as in Figure 4 in the manuscript. It reveals that diphone features' prediction performance is significantly higher than the predictions obtained from triphone features (blue). This effect is observed both for analyses based on 2 sessions and 5 sessions of data collection. There is a small difference in terms of additional prediction obtained from the triphone features after taking into account diphone features when more data is used but our central result remains: the diphone features play a more important role than the triphone features to explain the bold response in phonemic cortical regions. **Panel B** shows the prediction performance of diphone features is dominant in both 2 sessions and 5 sessions data. The blue voxels are better explained by the diphone features, while the orange voxels are better explained by the triphone features. White voxels are equally well predicted by diphone and triphone features. The flatmap reveals more blue voxels than orange or white voxels irrespective of whether 2 or 5 sessions or data are used.

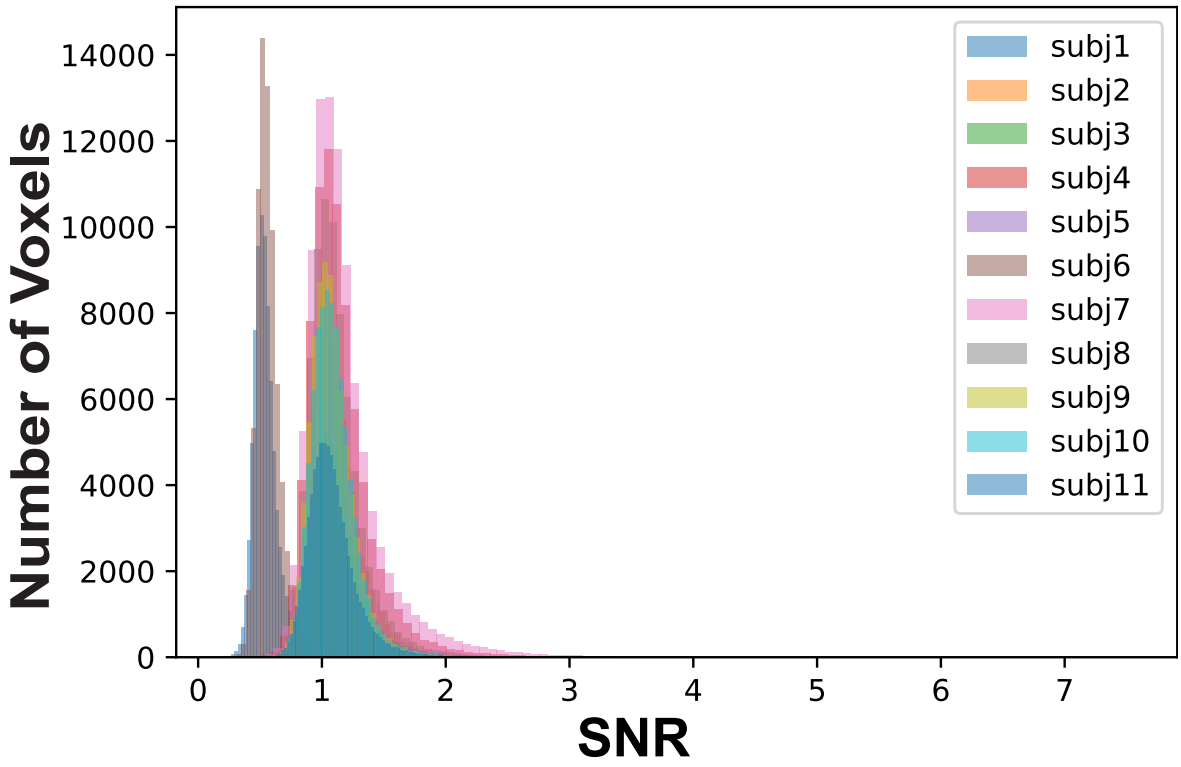


Figure A.16: **Signal-to-noise ratio of BOLD signal for each subject.** The SNR for each voxel of each subject is obtained from computing the coefficient of determination between two repeats of the BOLD responses collected when the subject was listening to the same story. The plot shows the histogram of SNR for each subject (each color represents one subject). It reveals that our simulation based on SNR=1 is representative to detect the phonemic feature sensitivity given our data size.

A.2 Supplementary Tables

Cerebral Cortex

subject	single	diphone	triphone	dprime
1	0.006	0.017	0.002	0.657
2	0.006	0.013	0.003	0.585
3	0.005	0.021	0.001	0.790
4	0.007	0.017	0.002	0.685
5	0.005	0.016	0.001	0.739
6	0.004	0.022	0.001	0.861
7	0.007	0.014	0.002	0.534
8	0.004	0.021	0.002	0.924
9	0.004	0.027	0.002	1.072
10	0.006	0.017	0.001	0.696
11	0.007	0.016	0.002	0.660
average	0.006	0.018	0.002	0.746
sem	0.000	0.001	0.000	0.047

Broca Area

subject	single	diphone	triphone	dprime
1	0.007	0.022	0.001	0.819
2	0.008	0.017	0.002	0.778
3	0.003	0.023	0.002	1.418
4	0.007	0.010	0.001	0.459
5	0.007	0.017	0.000	0.733
6	0.004	0.028	0.001	1.115
7	0.003	0.018	0.004	1.063
8	0.001	0.028	0.002	1.844
9	0.004	0.023	0.001	1.154
10	0.007	0.026	0.000	1.203
11	0.012	0.011	0.001	0.401
average	0.006	0.020	0.001	0.999
sem	0.001	0.002	0.000	0.127

Temporal Cortex

	PAC				STG				STS				LTC			
subject	single	diphone	triphone	dprime	single	diphone	triphone	dprime	single	diphone	triphone	dprime	single	diphone	triphone	dprime
1	0.015	0.000	0.000	-0.601	0.007	0.018	0.001	0.670	0.005	0.025	0.001	1.066	0.003	0.018	0.005	0.863
2	0.002	0.004	0.000	0.384	0.009	0.014	0.002	0.558	0.005	0.022	0.004	1.109	0.010	0.009	0.002	0.316
3	0.002	0.005	0.000	0.577	0.007	0.022	0.001	0.603	0.007	0.020	0.001	0.709	0.006	0.028	0.001	0.867
4	0.014	0.010	0.002	0.173	0.008	0.019	0.001	0.808	0.007	0.024	0.001	0.999	0.005	0.021	0.001	0.806
5	0.006	0.001	0.000	-0.252	0.004	0.020	0.001	0.824	0.006	0.017	0.001	0.750	0.005	0.010	0.002	0.512
6	0.003	0.015	0.002	0.848	0.006	0.018	0.000	0.611	0.003	0.031	0.001	1.233	0.003	0.021	0.002	0.938
7	0.017	0.005	0.001	-0.030	0.011	0.012	0.005	0.235	0.005	0.023	0.005	0.824	0.006	0.015	0.003	0.608
8	0.006	0.008	0.001	0.263	0.005	0.022	0.002	0.927	0.001	0.036	0.002	1.617	0.002	0.022	0.005	1.235
9	0.009	0.005	0.004	-0.137	0.005	0.029	0.002	0.962	0.003	0.040	0.002	1.512	0.004	0.024	0.002	1.217
10	0.009	0.010	0.001	0.287	0.007	0.022	0.001	0.800	0.006	0.027	0.001	1.042	0.005	0.024	0.001	1.219
11	0.009	0.008	0.004	0.141	0.011	0.013	0.002	0.477	0.010	0.020	0.001	0.757	0.005	0.011	0.002	0.635
average	0.008	0.006	0.001	0.150	0.007	0.019	0.002	0.680	0.005	0.026	0.002	1.056	0.005	0.018	0.002	0.838
sem	0.002	0.001	0.000	0.120	0.001	0.002	0.000	0.064	0.001	0.002	0.000	0.091	0.001	0.002	0.000	0.092

Table A.1: **Table of prediction performance of phonemic processing in each ROI for each subject and statistics based on subject averages.** The additive contribution of single phonemes, diphones and triphones to the prediction performance obtained with the Phonemic VM is shown for each subject for the entire cortex, Broca's area and the ROIs in the temporal cortex. The average and standard errors of estimated across subjects are shown in the last two rows. These data show that diphone prediction performance is significantly higher than single phoneme and triphones on average across cerebral cortex (11/11 subjects), Broca's area (9/11 subjects), STG (11/11 subjects), STS (11/11 subjects) and LTC (11/11 subjects). It indicates that the most important phoneme-related representations in the brain occur at the level of diphones. All tests are two sided. Yellow background: $p < 0.05$, Orange background: $p < 0.01$ and Red background: $p < 0.001$ obtained in a post-hoc pairwise t-test with Bonferroni correction.

Cerebral Cortex

subject	l_di_avg	l_sin_di	l_tri_di
1	1.589	0.614	2.565
2	1.399	0.294	2.504
3	2.220	1.229	3.211
4	1.574	0.745	2.403
5	1.901	0.715	3.087
6	2.145	1.029	3.260
7	1.430	0.463	2.396
8	2.041	1.128	2.953
9	2.405	1.781	3.029
10	1.684	0.465	2.903
11	1.665	0.774	2.557
average	1.823	0.840	2.806
sem	0.102	0.128	0.099

Broca's Area

subject	l_di_avg	l_sin_di	l_tri_di
1	1.367	0.431	2.303
2	1.445	1.099	1.792
3	2.649	2.996	2.303
4	0.714	-0.182	1.609
5	1.893	0.651	3.135
6	2.027	1.569	2.485
7	2.019	1.595	2.442
8	2.649	2.303	2.996
9	1.994	1.099	2.890
10	2.020	0.821	3.219
11	1.358	0.318	2.398
average	1.830	1.154	2.507
sem	0.174	0.277	0.157

Temporal Cortex

	PAC			STG			STS			LTC		
subject	l_single	l_diphone	l_triphone	l_single	l_diphone	l_triphone	l_single	l_diphone	l_triphone	l_single	l_diphone	l_triphone
1	-1.199	-2.398	-0.000	1.976	0.657	3.296	2.556	1.392	3.720	1.263	0.916	1.609
2	0.144	-0.405	0.693	1.472	0.359	2.584	2.447	1.407	3.486	0.795	-0.357	1.946
3	0.490	-0.405	1.386	1.979	0.609	3.350	2.013	0.721	3.305	2.497	1.433	3.561
4	0.450	-0.486	1.386	1.537	0.712	2.361	2.358	1.219	3.497	1.994	1.099	2.890
5	-0.602	-1.609	0.405	2.142	0.899	3.384	2.774	0.753	4.796	1.092	0.420	1.764
6	1.407	0.511	2.303	1.599	0.154	3.045	3.048	2.029	4.066	1.247	0.788	1.705
7	0.089	-1.075	1.253	0.670	-0.009	1.350	1.752	1.209	2.295	1.373	0.490	2.257
8	0.235	-0.916	1.386	1.756	1.237	2.335	3.314	2.621	4.007	1.966	1.764	2.169
9	0.405	-0.288	1.099	2.067	1.350	2.785	3.506	2.701	4.311	1.937	1.846	2.028
10	0.636	-0.336	1.609	1.899	0.819	2.979	2.400	1.216	3.584	1.893	1.435	2.351
11	0.182	-0.647	1.012	1.687	0.164	3.209	2.325	0.774	3.876	1.642	0.669	2.615
average	0.204	-0.732	1.139	1.710	0.632	2.789	2.590	1.459	3.722	1.609	0.955	2.263
sem	0.201	0.231	0.186	0.124	0.132	0.184	0.160	0.211	0.192	0.151	0.197	0.175

Table A.2: **Table of voxel counts and corresponding statistics of phonemic processing in each ROI for each subject based on subject averages with logistic regression.** In order to quantify the number of voxels that is best explained by single phoneme, diphone or triphone features, we assigned each cortical voxel to the best predictive feature. Then, we quantified the effect size of the difference in the average of the number of voxels best explained by each feature for each subject. This effect size is calculated as the average of the negative logits of the probability of being best explained by the single phone and triphone relative to the probability of being best diphone-based probability. The average and standard errors estimated across subjects are summarized as well. We performed a two-sided statistical analysis using mixed effect multinomial logistic regression which takes into account the varying number of voxels in each subject in its likelihood. This statistical analysis compares the actual probability of the number of voxels where the single phoneme, diphone and triphone contributions are higher to what is expected by chance given here by equal probability. We then use a likelihood ratio test comparing statistical mixed-effect model with fitted p_{single} and $p_{triphone}$ to the equal probability p_{single}=$p_{triphone}$=1/3. On average across the whole cerebral cortex, the additive prediction from the diphone features was consistently higher than that of the single phoneme or triphone features with $l = 1.802 + -0.172(2SE)$ ($\chi^2(2) = 1012.22, p < 2.2 \times 10^{-16}$, 11/11 subjects.Yellow background: $p < 0.05$, Orange: $p < 0.01$ and Red: $p < 0.001$ is obtained in a post-hoc pairwise t-test with Bonferroni correction). In the temporal cortex, the probabilities were estimated using a multinomial mixed effect statistical model with the subject as the random effect and the temporal cortex ROIs as a fixed effect (four levels: AC, STG, STS, and LTC). This statistical model was compared to the mixed effect multinomial model that did not include ROIs as fixed effect but did include the two non zero intercepts (yielding estimates for p_{single} and $p_{triphone}$ distinct from 1/3) and subject as a random effect. This interaction was significant ($\chi^2(6) = 227.88, p < 2.17 \times 10^{-46}$, 9/11 subjects) and showed a systematic increase in effect size from AC to STG and STS: the single phoneme, diphone and triphone-based models are equally good in AC ($l = 0.391 + -0.304(2SE)$, 4/11 subjects) but the diphone-based model becomes increasingly more and more dominant in STG ($l = 1.600 + -0.229(2SE)$, 11/11 subjects) and STS ($l = 2.423 + -0.223(2SE)$, 11/11 subjects) and less so in LTC ($l = 1.852 + -0.236(2SE)$, 9/11 subjects). This indicates that the additive contribution of the diphone feature is higher than the contribution of single phonemes and triphones in STG and STS but not in AC or LTC. In Broca's area, the additive prediction from the diphone feature was consistently higher than that of the single phoneme or triphone feature with $l = 1.790 + -0.305(2SE)$ ($\chi^2(2) = 287.76, p < 2.2 \times 10^{-16}$, 9/11 subjects). In sum, these results are in line with those obtained giving equal weighting to all subjects and with the results also obtained using the actual predicted value shown in Figure 4 of the main paper.

Diphone Category			
subject	Words	Beginning	Residual
1	0.036	0.027	0.013
2	0.063	0.024	0.009
3	0.057	0.022	0.015
4	0.049	0.021	0.014
5	0.050	0.019	0.009
6	0.035	0.017	0.011
7	0.060	0.020	0.016
8	0.092	0.033	0.019
9	0.068	0.021	0.012
10	0.040	0.019	0.015
11	0.073	0.022	0.012
average	0.057	0.022	0.013
sem	0.005	0.001	0.001

Table A.3: **Table of prediction performance of each diphone category for each subject.** The contribution to the prediction of short words is significantly higher than the beginning of words and diphone residuals for the subjects with red background in the table. The average and standard errors of estimated across subjects are shown in the last two rows. Yellow background: $p < 0.05$, Orange background: $p < 0.01$ and Red background: $p < 0.001$ obtained in a post-hoc pairwise two sided t-test with Bonferroni correction.

Cerebral Cortex

subject	phonemic	semantic	dprime
1	0.030	0.030	-0.002
2	0.032	0.035	0.080
3	0.032	0.034	0.083
4	0.029	0.036	0.209
5	0.025	0.035	0.368
6	0.030	0.033	0.100
7	0.029	0.039	0.295
8	0.032	0.042	0.289
9	0.043	0.040	-0.065
10	0.031	0.034	0.122
11	0.029	0.036	0.281
average	0.031	0.036	0.160
sem	0.001	0.001	0.042

Broca Area

subject	phonemic	semantic	dprime
1	0.036	0.032	-0.142
2	0.038	0.042	0.196
3	0.034	0.042	0.311
4	0.022	0.038	0.628
5	0.029	0.032	0.139
6	0.035	0.031	-0.159
7	0.029	0.037	0.255
8	0.035	0.037	0.064
9	0.034	0.038	0.135
10	0.039	0.049	0.349
11	0.027	0.033	0.303
average	0.033	0.037	0.189
sem	0.002	0.002	0.068

Temporal Cortex

	PAC			STG			STS			LTC		
subject	phonemic	semantic	dprime	phonemic	semantic	dprime	phonemic	semantic	dprime	phonemic	semantic	dprime
1	0.015	0.005	-0.639	0.032	0.027	-0.176	0.039	0.027	-0.420	0.029	0.038	0.338
2	0.008	0.017	0.741	0.034	0.033	-0.056	0.049	0.042	-0.243	0.028	0.032	0.181
3	0.009	0.012	0.223	0.036	0.025	-0.317	0.035	0.032	-0.094	0.041	0.041	-0.009
4	0.032	0.020	-0.600	0.035	0.030	-0.156	0.037	0.040	0.101	0.032	0.040	0.229
5	0.007	0.014	0.713	0.026	0.025	-0.046	0.027	0.032	0.229	0.020	0.034	0.562
6	0.022	0.028	0.235	0.026	0.025	-0.058	0.038	0.033	-0.155	0.029	0.036	0.226
7	0.025	0.019	-0.228	0.034	0.029	-0.135	0.042	0.030	-0.359	0.031	0.036	0.163
8	0.017	0.017	-0.006	0.032	0.026	-0.235	0.043	0.030	-0.497	0.032	0.053	0.530
9	0.022	0.022	-0.001	0.049	0.042	-0.194	0.058	0.050	-0.191	0.038	0.037	-0.025
10	0.023	0.029	0.182	0.035	0.038	0.110	0.040	0.046	0.177	0.038	0.032	-0.225
11	0.024	0.025	0.027	0.032	0.032	0.002	0.038	0.039	0.026	0.022	0.032	0.469
average	0.018	0.019	0.059	0.034	0.030	-0.115	0.041	0.036	-0.130	0.031	0.037	0.222
sem	0.002	0.002	0.134	0.002	0.002	0.036	0.002	0.002	0.073	0.002	0.002	0.074

Table A.4: **Table of prediction performance performance of Phonemic-Semantic processing in each ROI for each subject and statistics based on subject averages.** The table shows the mean prediction performance of each subject's BOLD response explained by the third order phonemic and semantic based VM for each ROI (whole cortex, AC, STG, STS, LTC, and Broca area). The average and standard errors of the mean performance estimated across subjects for each ROI are presented in the last two rows. These data show that semantic prediction performance is significantly higher than phonemic features on average across cerebral cortex (9/11 subjects), Broca's area (4/11 subjects), and LTC (5/11 subjects), while phonemic prediction performance is higher than the semantic features in STG (3/11 subjects) and STS (6/11 subjects). Yellow background: $p < 0.05$, Orange background: $p < 0.01$ and Red background: $p < 0.001$ obtained in a post-hoc pairwise two sided t-test with Bonferroni correction.

Count Data for Phonemic VS Semantic Comparisons

subject	l_phnSem_APC	l_phnSem_STG	l_phnSem_STS	l_phnSem_LTC	l_phnSem_BA	l_phnSem_CC
1	-0.588	-0.729	-1.200	0.833	-0.105	0.288
2	-0.182	-1.109	-0.736	0.226	0.693	0.270
3	-0.693	-0.726	-0.366	0.829	1.504	0.634
4	-0.480	-0.444	0.048	0.864	-0.542	0.742
5	0.596	-0.217	-0.557	0.654	0.087	1.049
6	0.642	-0.329	-1.128	1.995	0.288	1.213
7	0.359	0.134	0.654	1.363	0.606	1.336
8	0.074	0.231	0.366	0.338	0.211	0.459
9	0.288	0.660	0.928	1.295	2.451	1.514
10	0.457	0.171	0.134	1.596	-0.128	0.864
11	0.095	0.167	0.382	0.758	0.591	0.437
average	0.052	-0.199	-0.134	0.978	0.514	0.801
sem	0.143	0.159	0.215	0.161	0.253	0.130

Table A.5: **Table of voxel counts and corresponding statistics of Phonemic-Semantic processing in each ROI for each subject based on subject averages with logistic regression.** In order to quantify the number of voxels that is best explained by phonemic or semantic model, we assigned each cortical voxel to the best predictive feature. Then, we quantified the effect size of the difference by calculating the negative logit of the probability based on the number of voxels being best explained by the Phonemic features vs the additive contribution of the semantic feature in those significant voxels (winner-take-all count analysis). The average and standard errors of the logits estimated across subjects for each ROI are presented in the last two rows. Using subject mean and two SE, one could conclude that Semantic VM is more predictive in the whole cortex, in LTC and in Broca's area (just reaching the threshold of significance). To better model the statistical significance for this count data, we also used mixed-effect logistic regression, which incorporates the number of voxels in each subject in its likelihood. In logistic regression, we statistically assessed whether the number of voxels best explained by the phonemic vs semantic features are different from expectations based on the binomial distribution. The number of voxels best explained by semantic versus phonemic features was found to be significantly higher throughout the whole cortex with the logit equal to $l = 0.159 + -0.106(2SE)$ (Likelihood ratio test comparing statistical mixed-effect model with equal probability of semantic and third order phonemic voxels: $\chi^2(1) = 6.61, p = 1.0 \times 10^{-2}$). The number of voxels best explained by the semantic based VM was also higher in Broca's area: $l = 0.598 + -0.266(2SE)$ ($\chi^2(1) = 11.51, p = 6.9 \times 10^{-4}$). In order to further examine how different temporal cortical areas involved in the phonemic versus semantic processing, we used a generalized mixed-effect linear statistical model (glmer with family=binomial) with the temporals cortex ROIs (four levels: PAC, STG, STS, and LTC) as regressor, subject as the random effect and the fraction of voxels best explained by the semantic feature as the response variable. We found that the fraction of voxels best explained by the semantic based VM varied significantly across ROIs (likelihood ratio test with nested model that does not include ROIs, $\chi^2(3) = 583.80, p < 2.2 \times 10^{-16}$). Moreover, the number of voxels best predicted by the semantic feature is significantly higher in LTC $l = 0.244 + -0.068(2SE)$ ($\chi^2(1) = 19.21, p = 1.2 \times 10^{-5}$), but the differences are negligible in STG $l = -0.070 + -0.160(2SE)$ ($\chi^2(1) = 0.74, p = 3.9 \times 10^{-1}$) and STS $l = 0.293 + -0.240(2SE)$ ($\chi^2(1) = 4.78, p = 2.9 \times 10^{-2}$). These results are in line with those obtained giving equal weighting to all subjects and with the results also obtained using the actual predicted value shown in Figure 6.

Feature Name	Feature Definition
unsDPAV	Unstressed diphone probability average; vowel-stress ignored
unsFDPAV	unsDPAV weighted with SUBTLEXus word frequency
unsLDPAV	unsDPAV weighted with log SUBTLEXus word frequency
unsCDPAV	unsDPAV weighted with SUBTLEXus context count
strDPAV	Stressed diphone probability average; distinct stressed-vowels
strFDPAV	strDPAV weighted with SUBTLEXus word frequency
strLDPAV	strDPAV weighted with log SUBTLEXus word frequency
strCDPAV	strDPAV weighted with SUBTLEXus context count

Table A.6: **Table of definition of diphone phontactic probability features**. The table summarizes the name and content of each feature for modeling diphone phontactic probability. Phonotactic probabilities refer to the concurrence likelihood of the sequence of sounds that are present in a given word (**vitevitch2004web**; Vaden et al., 2009). Diphone probability average refers to the average likelihood of each diphone occurring in each position of a word. In these measures, the syllable stress placement of vowels can also be considered. For stressed calculations, identical vowel sounds are considered to be distinct phonemes depending on primary, secondary, or no-stress placement. In unstressed calculations, vowel sounds are considered to be single phoneme categories.

A.3 Supplementary Methods

Feature Construction

In order to explore how the properties of vowel and consonants contribute to the cerebral cortical encoding of phonemes and phonemic combinations, we created two features based on the identity of consonants and vowels of each phoneme.

Single-vc features uses a one-hot coding matrix (dimension: [number of TR, 2]) to encode if a single phoneme is a consonant or vowel. Diphone-vc features (dimension: [number of TR, 6]) encodes if a diphone belongs to any of these six consonant/vowel combinations: vowel-vowel (vv), vowel-consonant (vc), consonant-vowel (cv) and consonant-consonant (cc), vowel-blank (vb), consonant-blank (cb)

Afterwards, We then fitted four nested VMs: single vc model (using single-vc feature), first order vc model (using single-vc + single phoneme features), first order vc + diphone vc model (using single-vc + single phoneme + diphone-vc features) and second order vc model (using single-vc + single phoneme + diphone-vc + diphone features). The variance explained by the single-vc feature is obtained from the single vc model, while the additional variance explained by diphone-vc feature is obtained from subtracting the explainable variance of the first order vc model from the explainable variance of the first order vc + diphone vc model. In addition, the additional variance explained by the single phoneme identities beyond vowel and consonant categories is obtained from subtracting the explainable variance of the single vc model (using single-vc feature) from the first order vc model. Similarly, the variance explained by the diphone identities beyond the combination of vowel and consonants is obtained from subtracting the explainable variance of the first order vc + diphone vc model from the second order vc model.

Furthermore, In order to compare the relative contribution of vowels and consonants towards the phoneme encoding, we examined the modeling weights (coefficients) of single-vc (consonant or vowels) and diphone-vc (vv, vc, cv, cc, vb, cb) features. The weights of each voxel have been scaled by the prediction performance of this voxel in order to get rid of the random effects from noisy voxels.

A.4 Supplementary Notes

Laterality of the phonemic representation and segmentation.

We examined whether the phonemic representation was lateralized. First, we found that the average performance of the full *phonemic* was significantly different throughout the cerebral cortex across the right and left hemispheres (likelihood ratio test with the nested statistical model that does not include the hemisphere factor: $\chi^2(1) = 5.07, p = 2.4 \times 10^{-2}$), as well as in the LTC ($\chi^2(1) = 13.52, p = 2.4 \times 10^{-4}$). The prediction performance was not lateralized in AC ($\chi^2(1) = 0.27, p = 6.0 \times 10^{-1}$), STG ($\chi^2(1) = 0.26, p = 6.1 \times 10^{-1}$), STS($\chi^2(1) = 0.04, p = 8.5 \times 10^{-1}$) or Broca's area ($\chi^2(1) = 0.05, p = 8.2 \times 10^{-1}$).

We also tested whether the segmentation at the level of the diphone was comparable for the left and right hemispheres (Supplementary Fig 7). The unique contributions of the three levels of phonemic segmentation were significantly different between left and right hemispheres when the entire cortex is analyzed ($\chi^2(2) = 12.32, p = 2.1 \times 10^{-3}$), as well as for the temporal cortex ($\chi^2(12) = 78.74, p = 7.2 \times 10^{-12}$) and for Broca's area ($\chi^2(2) = 44.55, p = 2.1 \times 10^{-10}$). More specifically, within the temporal cortex, the additive contributions based on three levels of phonemic segmentation do not differ significantly between left and right STS ($\chi^2(2) = 4.23, p = 0.12$), but do differ significantly in left and right primary auditory cortex ($\chi^2(2) = 36.63, p = 1.1 \times 10^{-8}$), STG ($\chi^2(2) = 28.25, p = 7.3 \times 10^{-7}$), and LTC ($\chi^2(2) = 14.57, p = 6.9 \times 10^{-4}$). In STG, STS, LTC and Broca's area, the effect size d' was higher in the left versus right hemispheres. This indicates that the diphone segmentation for phonemic processing is more prominent in the left hemisphere than the right hemisphere.

Laterality of phonemic versus semantic representations.

We also examined whether the semantic representation was lateralized. First, we found that the average performance of the the semantic VM was significantly different throughout the cerebral cortex across the right and left hemispheres ($\chi^2(1) = 255.53, p < 2.2 \times 10^{-16}$), as well as in the STG ($\chi^2(1) = 9.57, p = 2.0 \times 10^{-3}$), STS ($\chi^2(1) = 5.76, p = 1.6 \times 10^{-2}$), and Broca's area ($\chi^2(1) = 5.09, p = 2.4 \times 10^{-2}$). The prediction performance was not lateralized in LTC ($\chi^2(1) = 0.86, p == 3.5 \times 10^{-1}$)

Next, we examined whether the relative predictive power of the phonemic versus semantic VM varied significantly across hemispheres (Supplementary Fig 12). We found that this interaction was statistically significant for the whole cortex (mixed-effect statistical models with subject as random factor: likelihood ratio test with the nested statistical model that does not include the interaction between hemisphere and models: $\chi^2(1) = 93.12, p < 2.2 \times 10^{-16}$) and in the temporal cortex (likelihood ratio test with the nested statistical model that does not include ROIs, features spaces, hemisphere and the interaction between ROIs and feature spaces: $\chi^2(7) = 32.32, p = 3.5 \times 10^{-5}$), but not for Broca's area ($\chi^2(1) = 1.76, p = 1.8 \times 10^{-1}$). In addition, within the temporal cortex, hemispheric differences in relative performance were not significant in primary AC ($\chi^2(1) = 0.98, p = 3.2 \times 10^{-1}$) and STS

($\chi^2(1) = 1.36, p = 2.4\times10^{-1}$), but significant in STG ($\chi^2(1) = 5.06, p = 2.5\times10^{-2}$), and LTC ($\chi^2(1) = 10.42, p = 1.2\times10^{-3}$). This indicates that the difference in semantic and phonemic processing is more prominent in the right hemisphere in STG and LTC, while this difference is more balanced in the PAC, STS and Broca's area of both hemispheres. Supplemental figure 12 shows the data for all subjects and the effect sizes across two hemispheres between the semantic and phonemic based VMs.

Appendix B

Supplements for "Representation of English reading"

B.1 Supplementary Figures

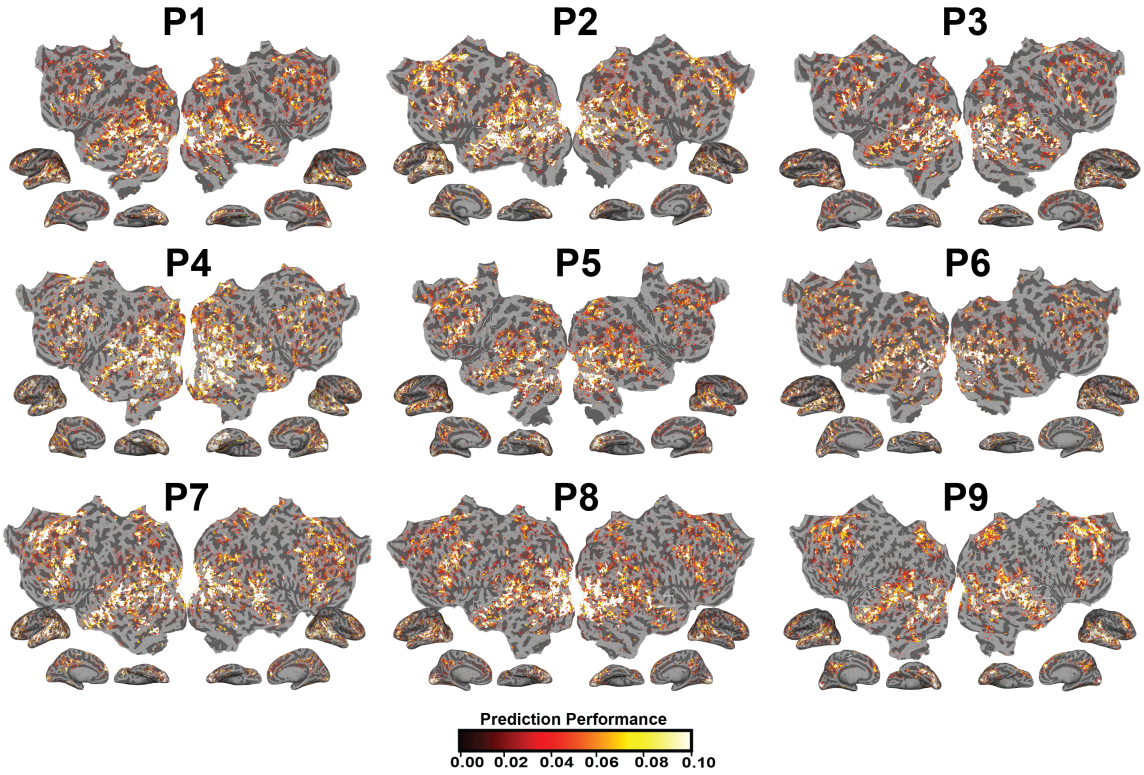


Figure B.1: **Joint prediction performance for individual participants.**

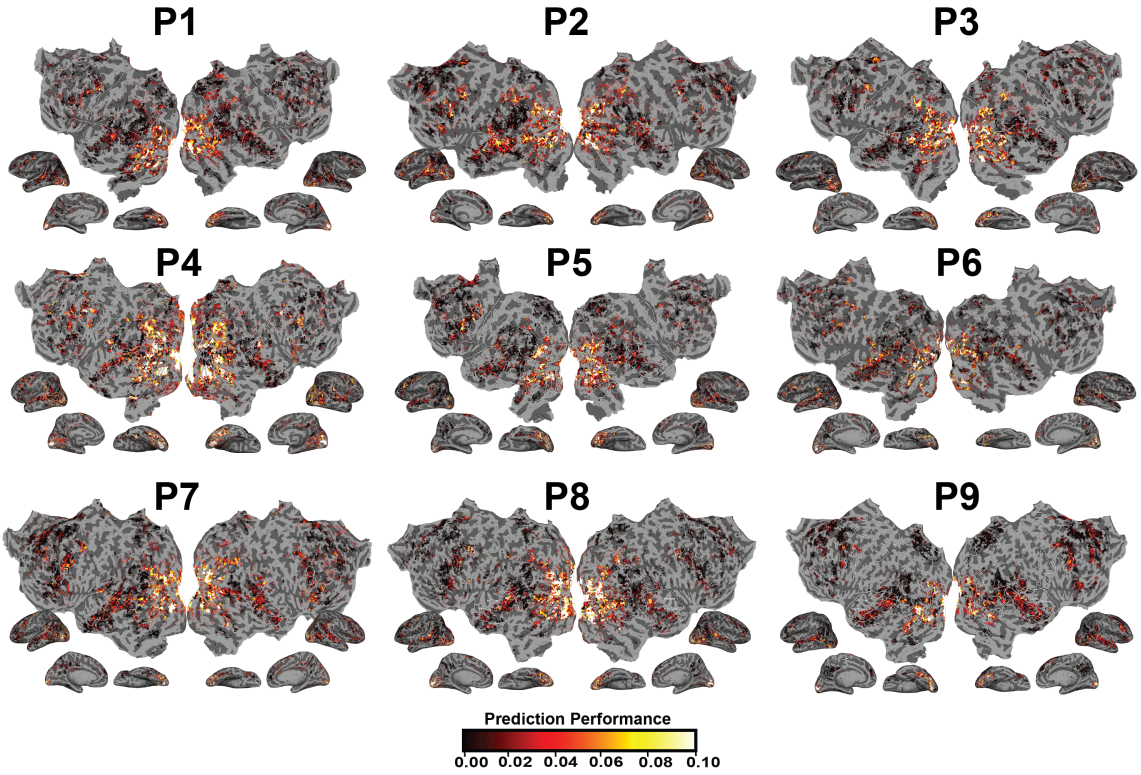


Figure B.2: **Split prediction performance of visual features for individual participants.**

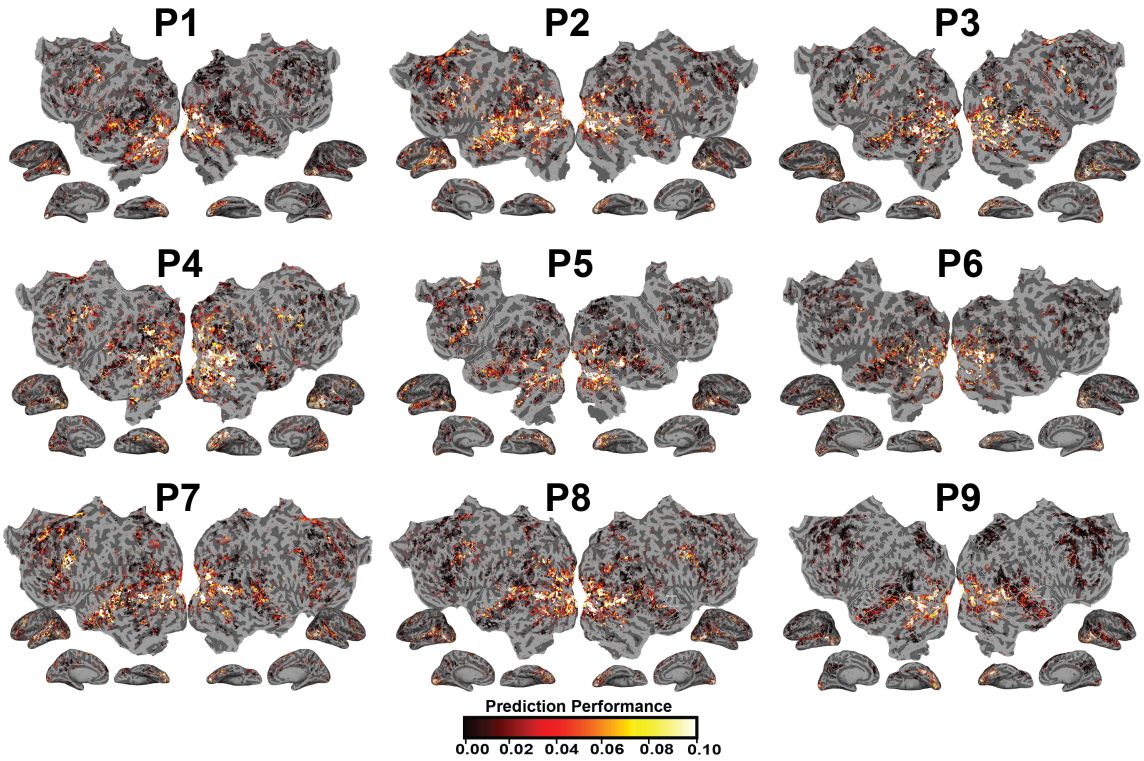


Figure B.3: **Split prediction performance of orthographic features for individual participants.**

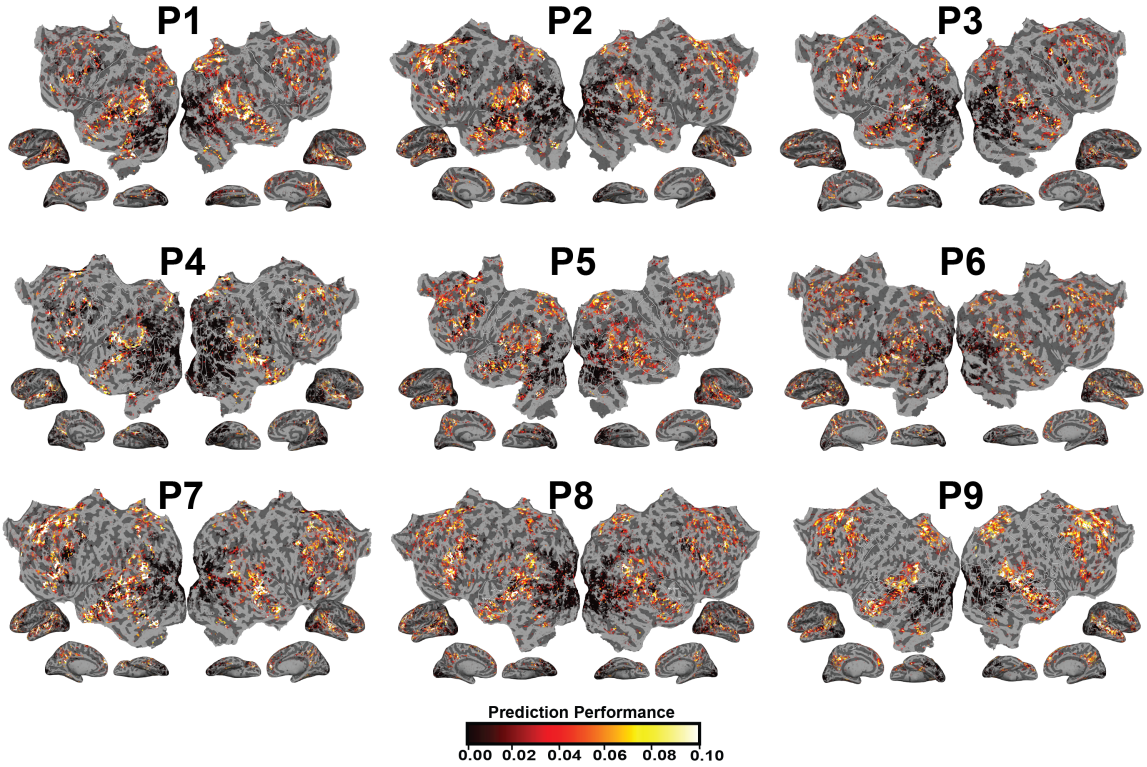


Figure B.4: **Split prediction performance of semantic features for individual participants.**

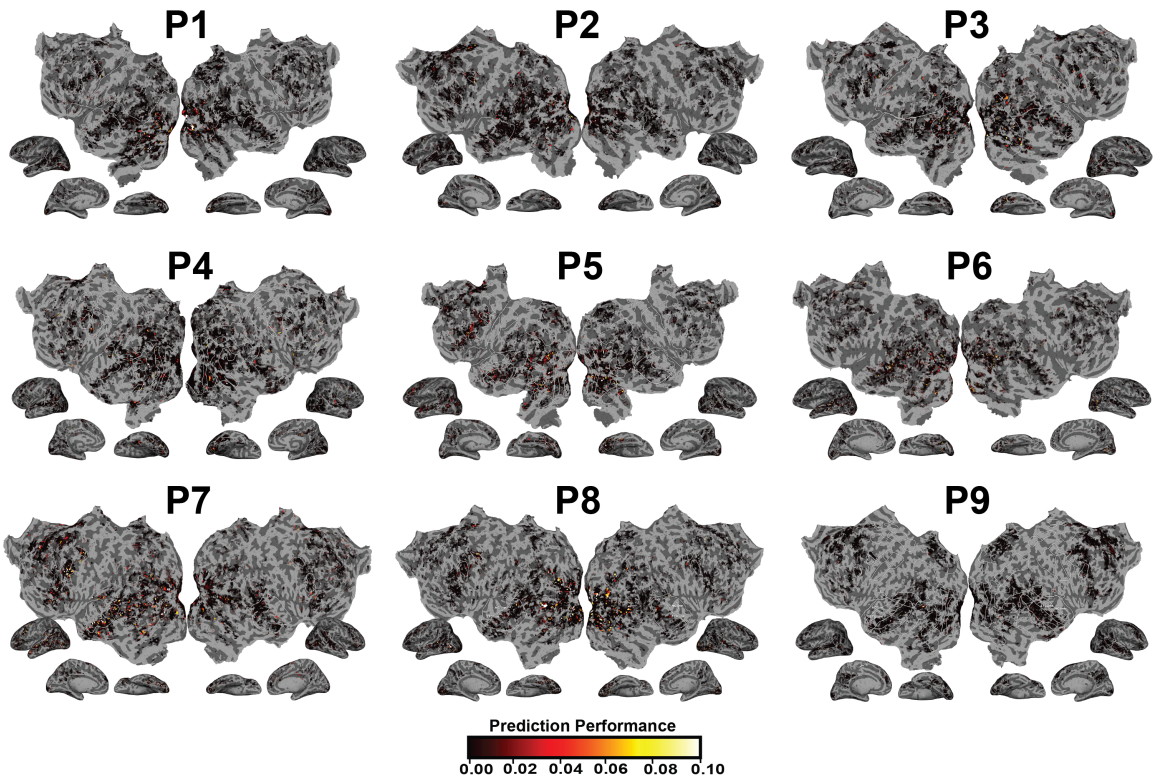


Figure B.5: **Split prediction performance of statistics features for individual participants.**

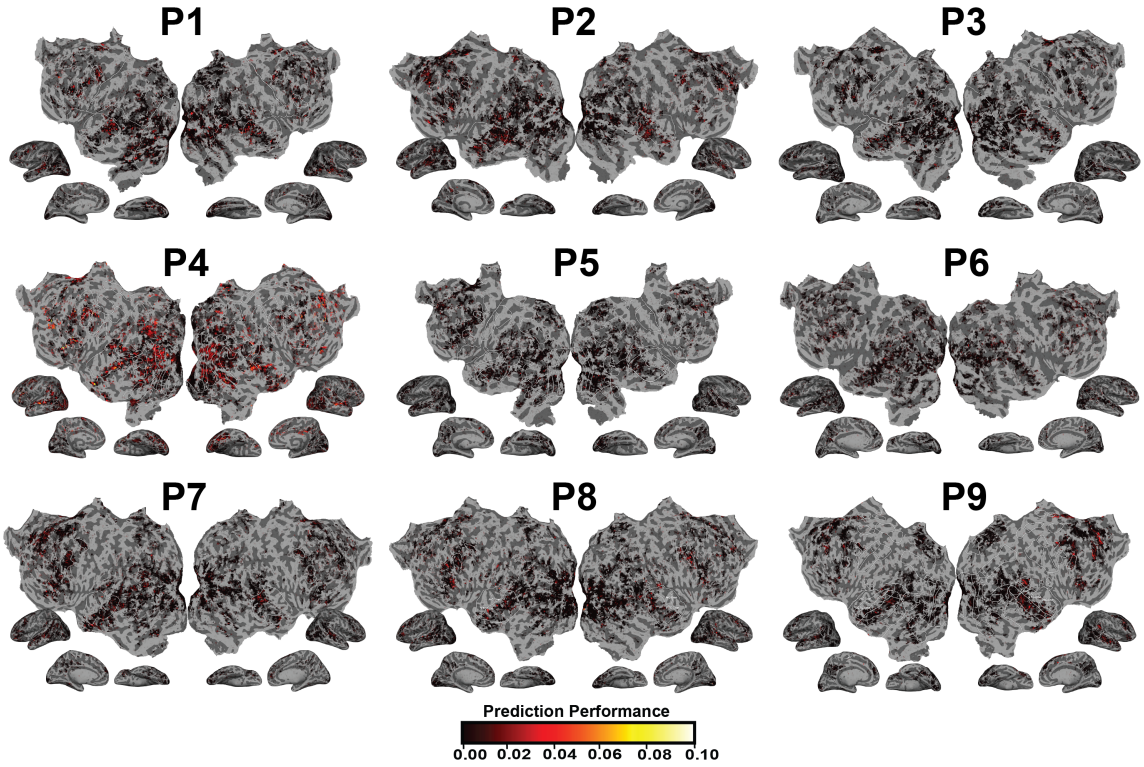


Figure B.6: **Split prediction performance of phonemes features for individual participants.**

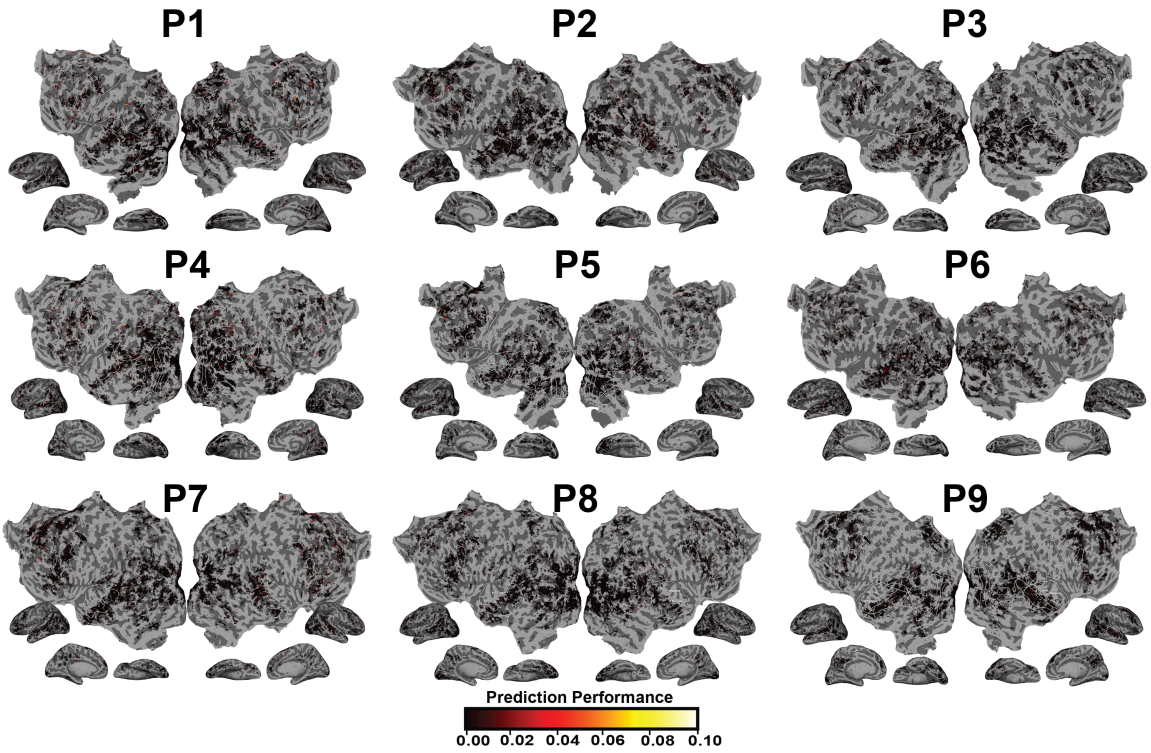


Figure B.7: **Split prediction performance of syntax features for individual participants.**

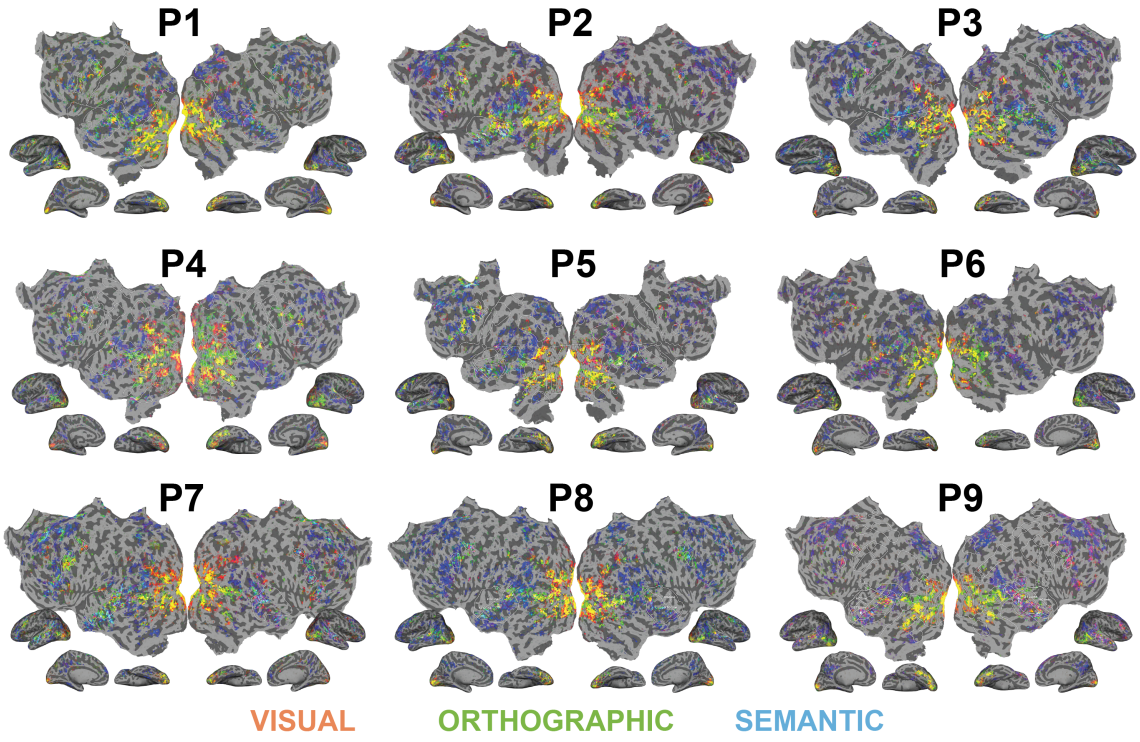


Figure B.8: **Prediction accuracy is shown for visual (orange), orthographic (green) and semantic (blue) feature spaces on the flattened cortical surface of individual participants.**

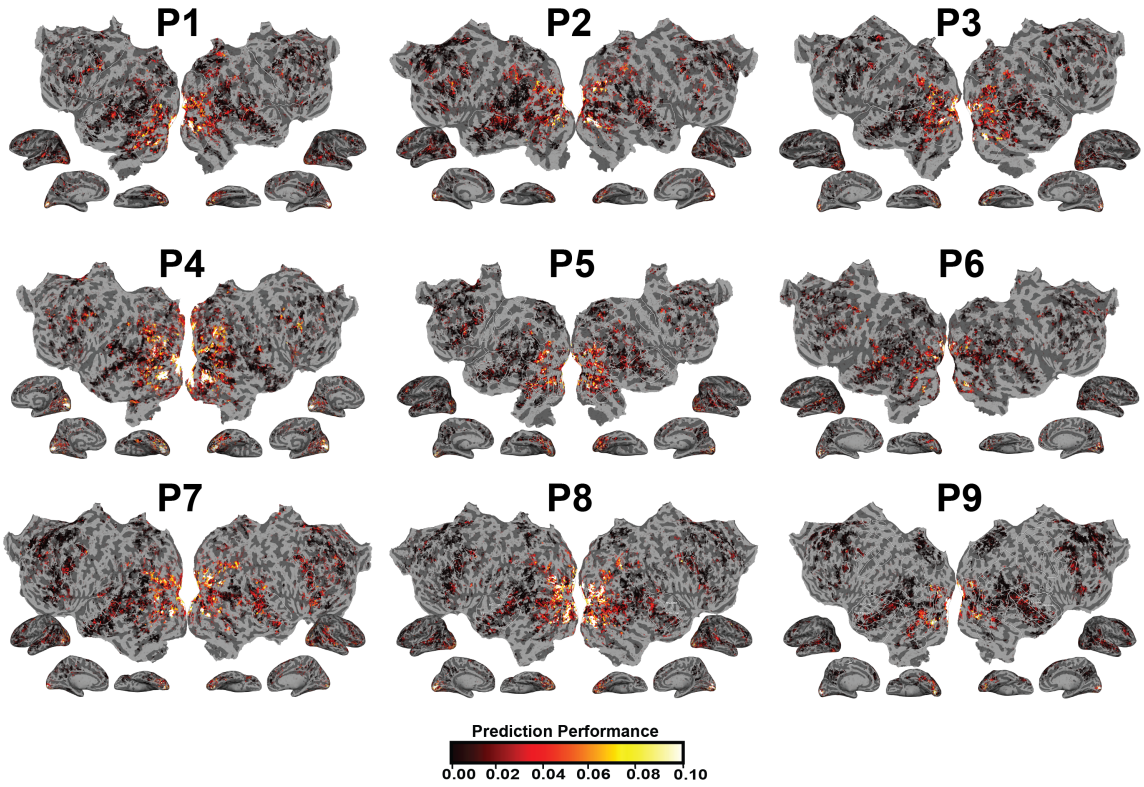


Figure B.9: **Unique variance of visual partition for individual participants.**

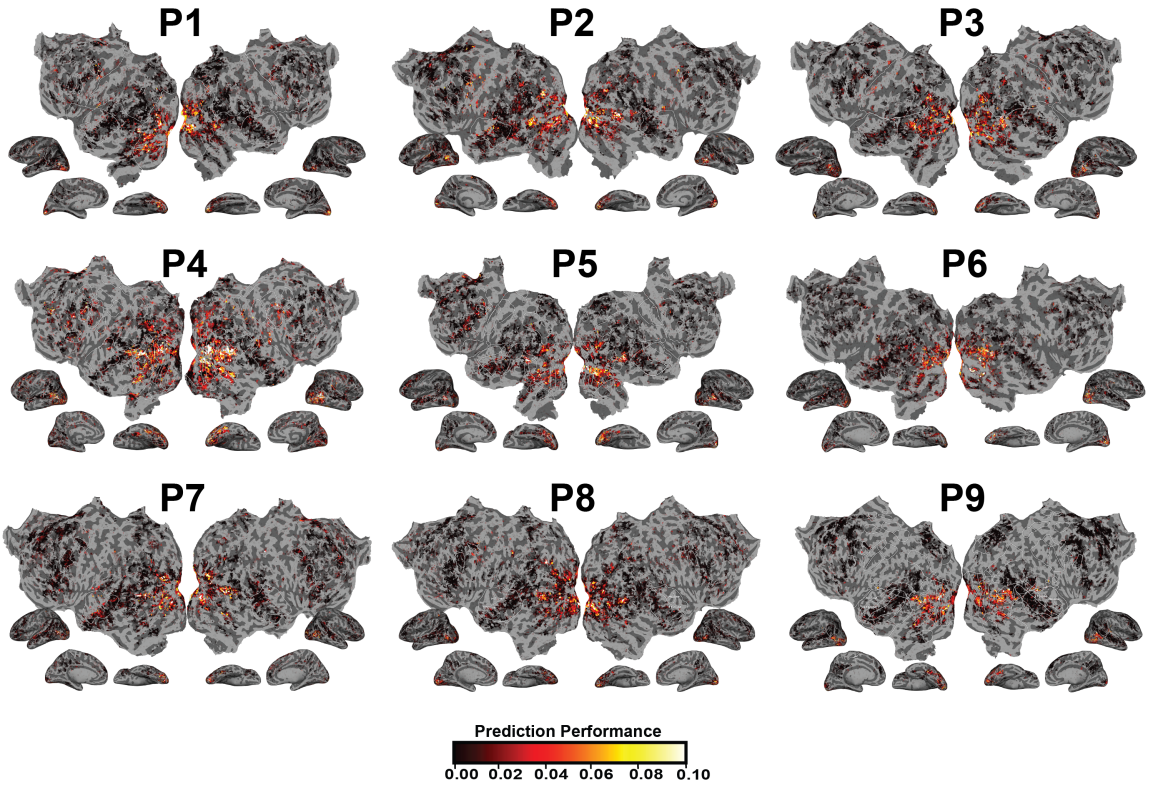


Figure B.10: **Unique variance of orthographic partition for individual participants.**

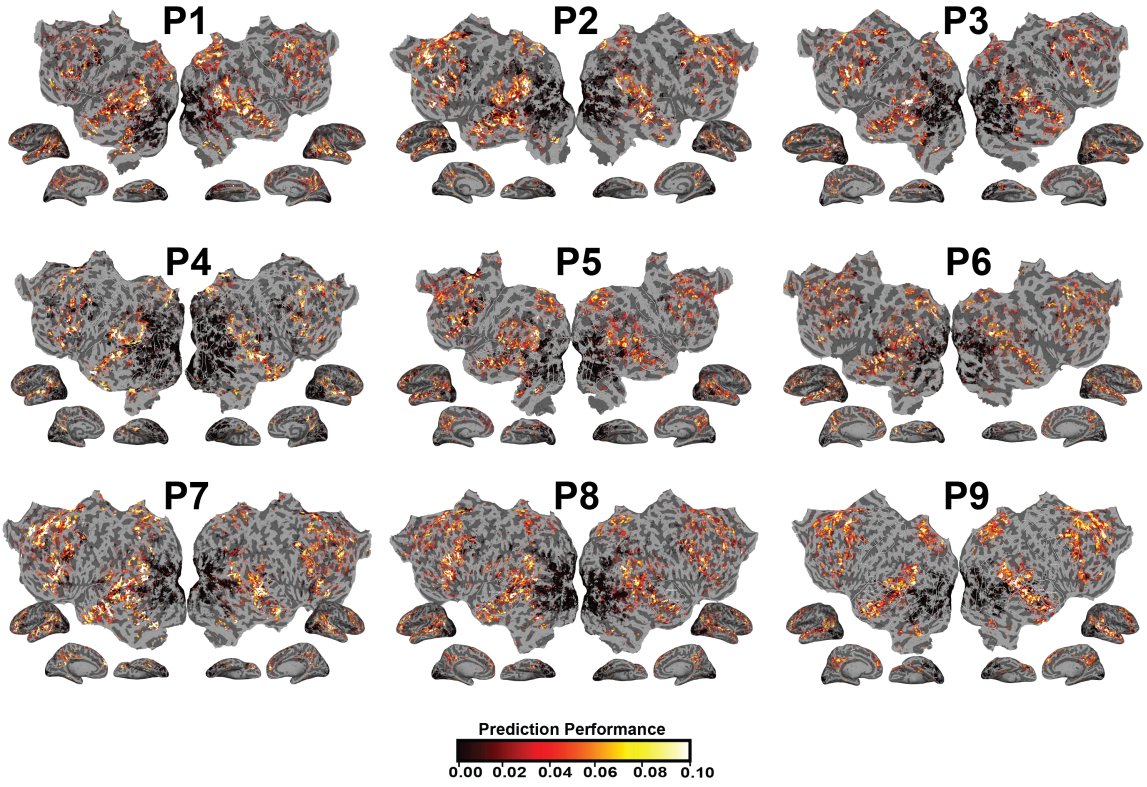


Figure B.11: **Unique variance of semantic partition for individual participants.**

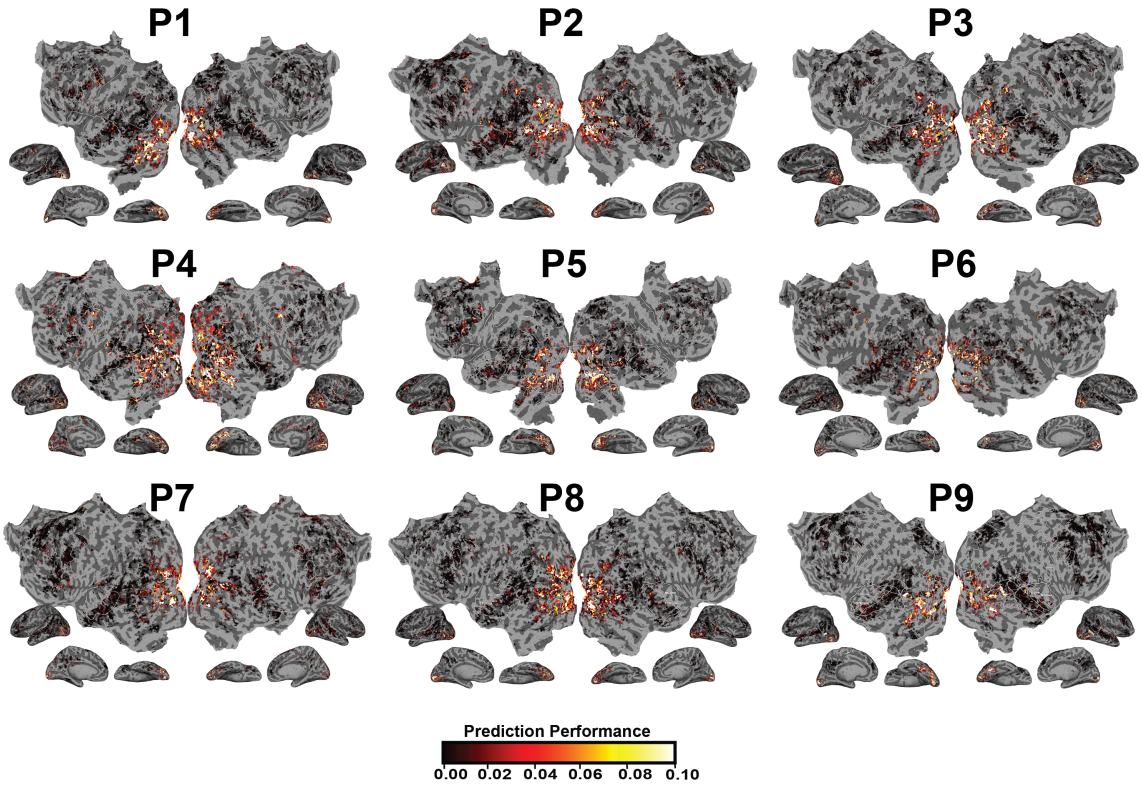


Figure B.12: **Unique variance of visual-orthographic partition for individual participants.**

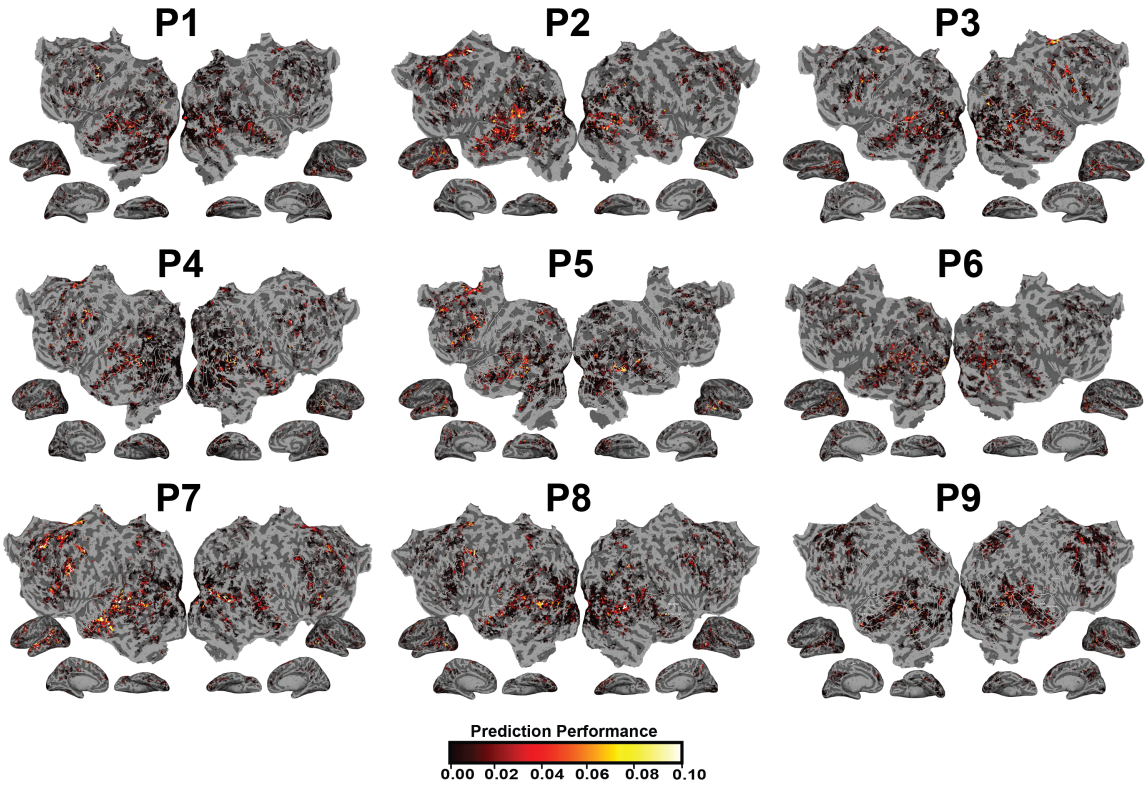


Figure B.13: **Unique variance of orthographic-semantic partition for individual participants.**

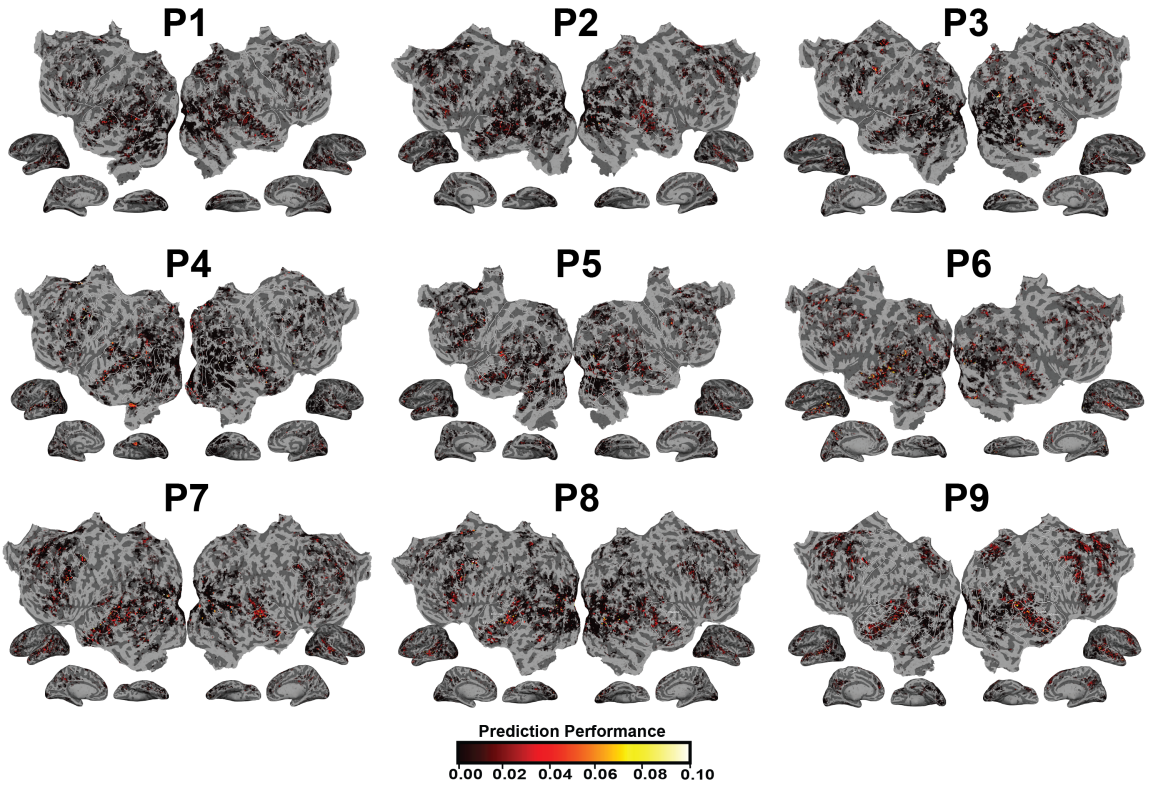


Figure B.14: **Unique variance of visual-semantic partition for individual participants.**

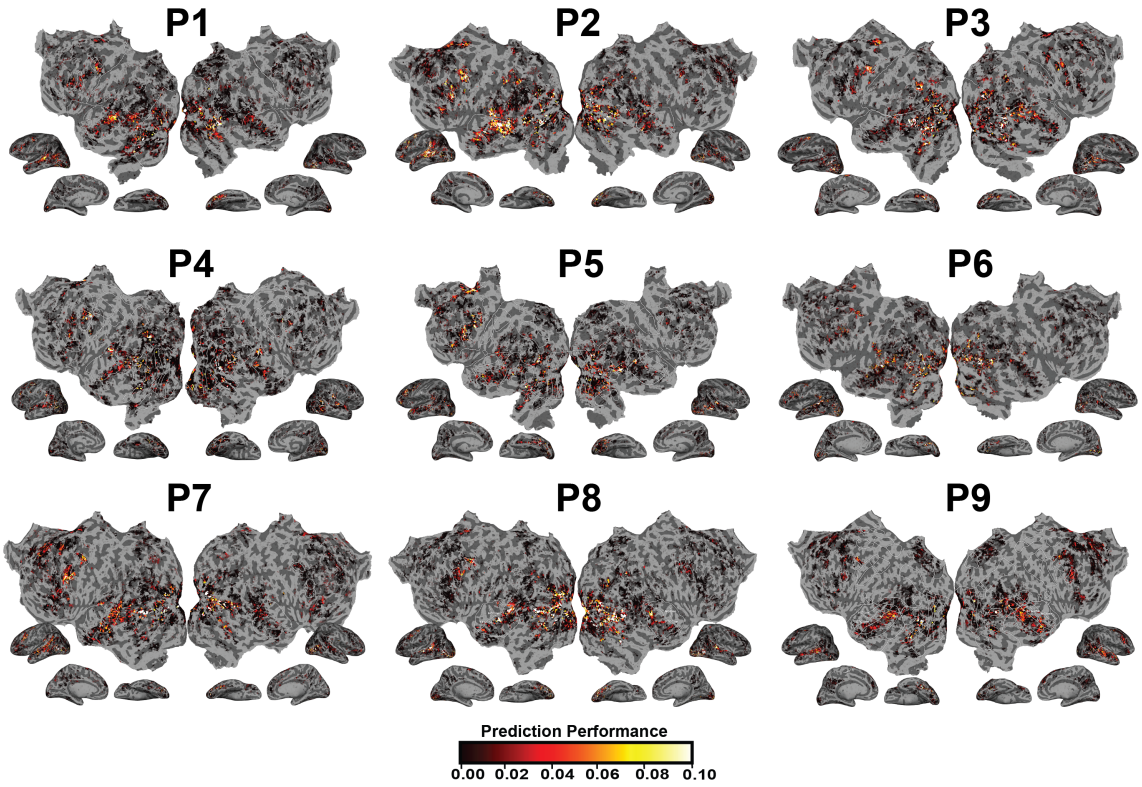


Figure B.15: **Unique variance of visual-orthographic-semantic partition for individual participants.**

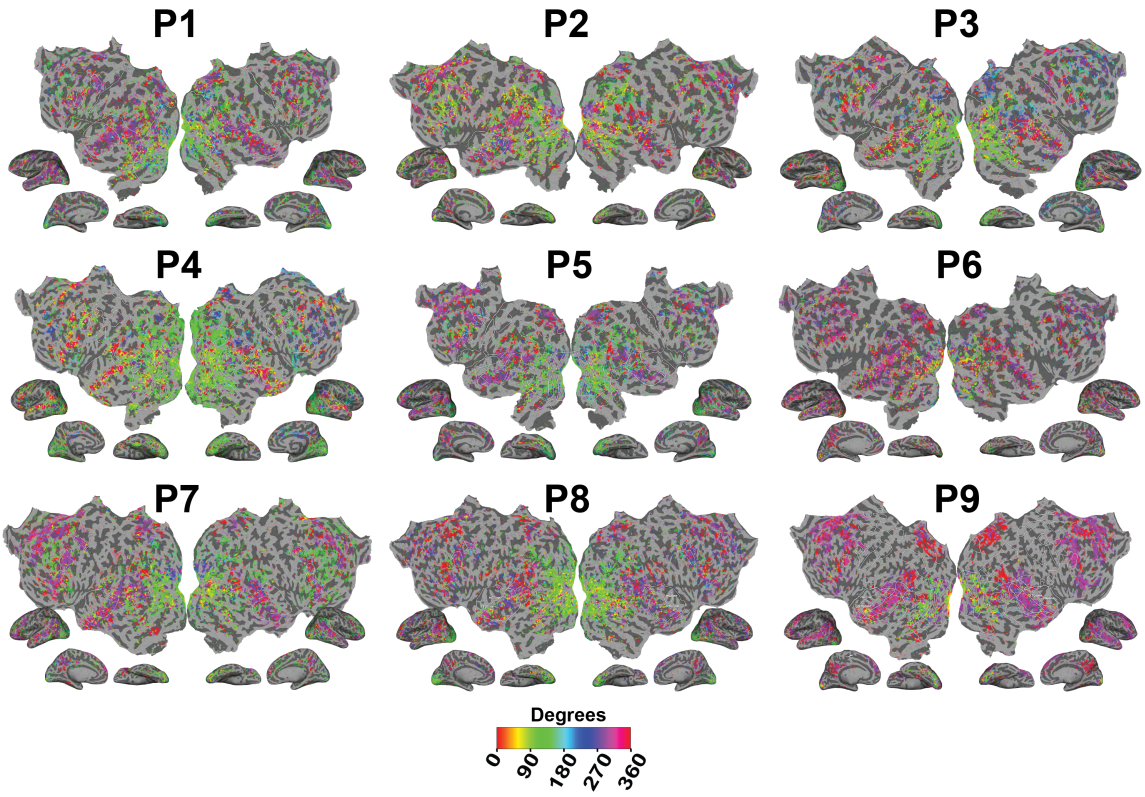


Figure B.16: **Visual features direction tuning for individual participants.**

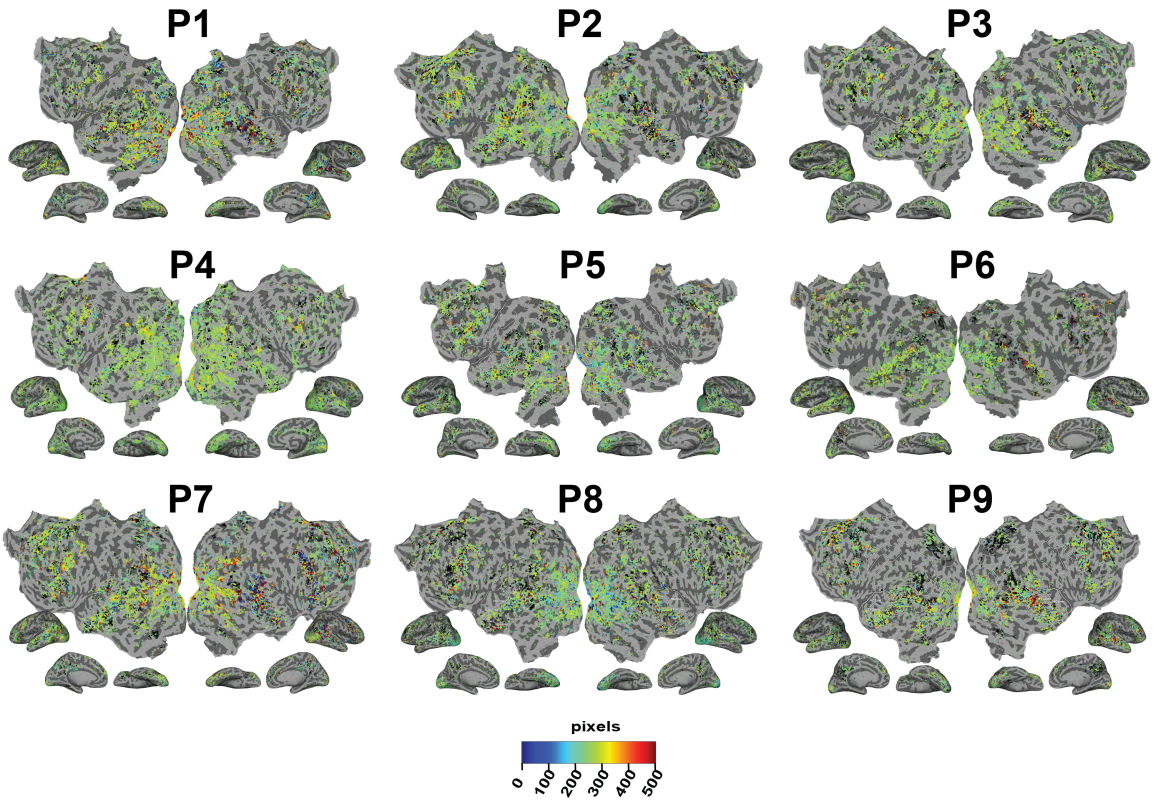


Figure B.17: **Visual features position tuning for individual participants.**

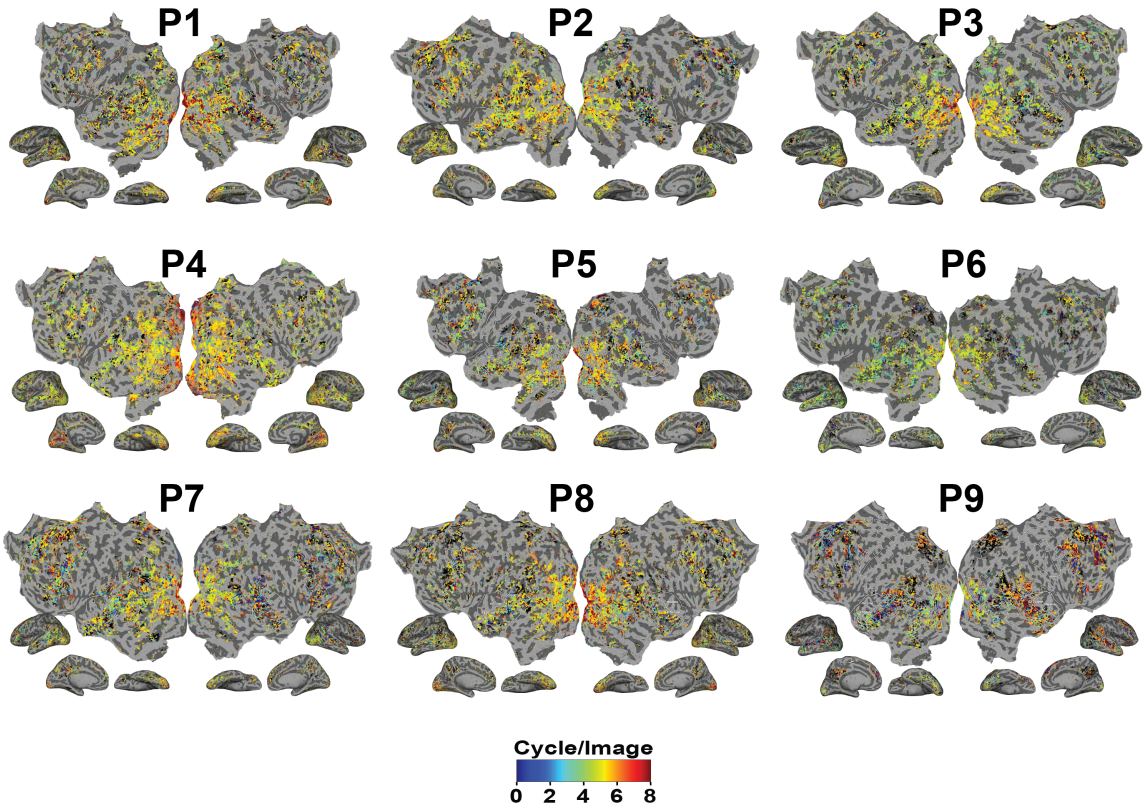


Figure B.18: **Visual features spatial frequency tuning for individual participants.**

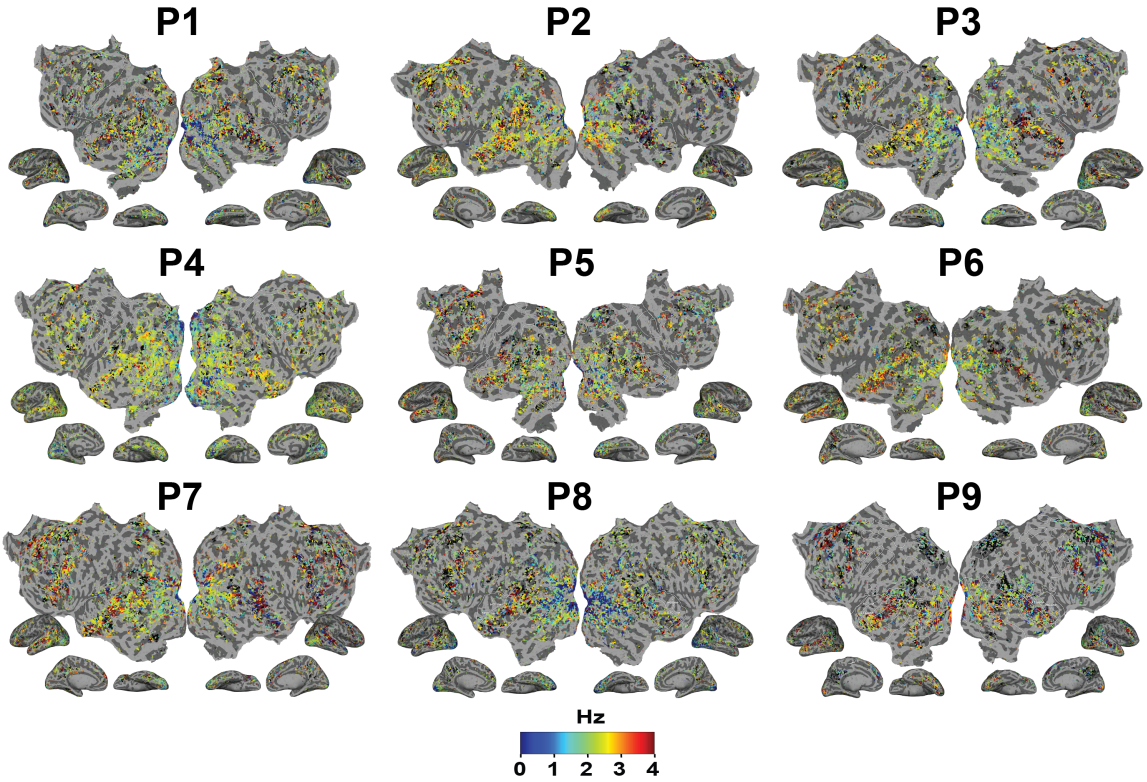


Figure B.19: **Visual features temporal frequency tuning for individual participants.**

Appendix C

Supplements for "Representation of Bilingual reading"

C.1 Supplementary Figures

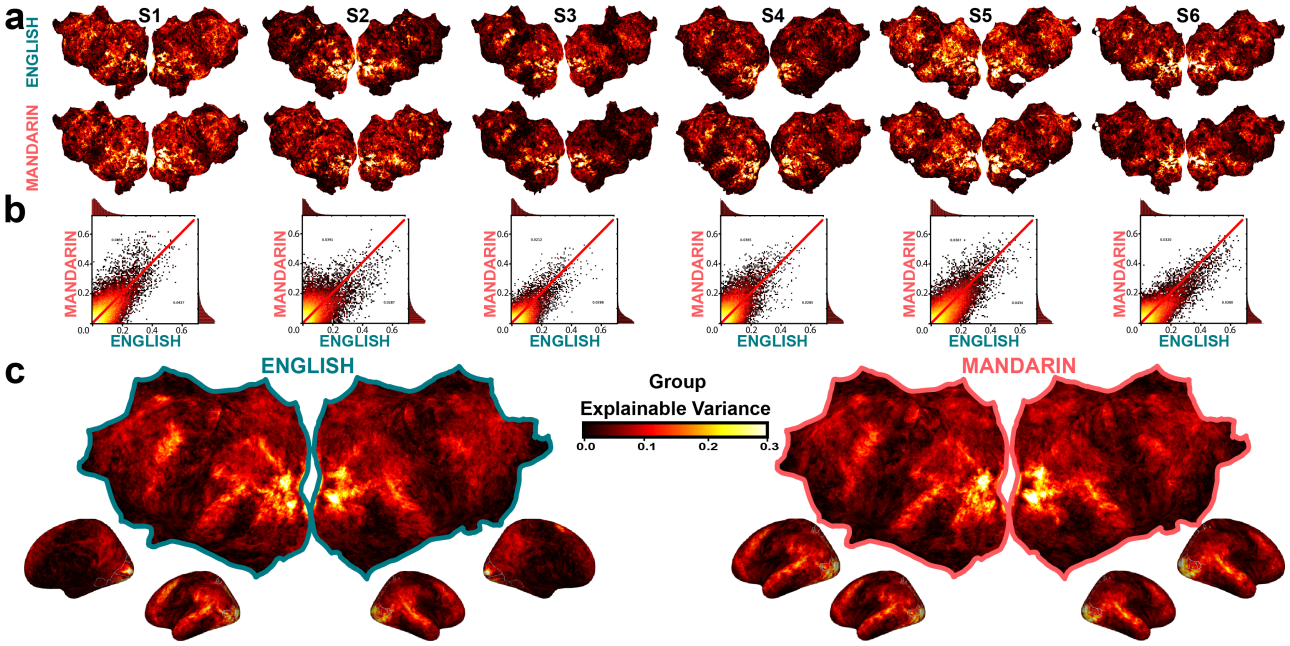


Figure C.1: **Explainable variance of English versus Mandarin.** Subjects read one validation story four times in both English and Mandarin, in order to estimate the signal-to-noise ratio (SNR) of each cortical voxel involved throughout all stages of reading comprehension. The coefficient of determination (R^2) of responses among four presentations for each voxel and each language is computed as explainable variance. a. The explainable variance (EV) is shown for each subject on the flattened cortical surface of the participant's native brain space. Well-explained voxels appear brighter. For each subject, bilateral occipital, temporal, parietal, and prefrontal cortices show high EV, indicating that large cortical regions are involved in reading comprehension in both English and Mandarin. b. The 2D histogram with density plot of each participant's EV for English (x-axis) versus Mandarin (y-axis) is shown. Points indicate all voxels. Brighter colors indicate a larger number of voxels in the corresponding bin. EV of voxels cluster along the diagonal line, indicating that the SNR of the recorded BOLD responses was comparable for each participant across both languages. c. Group-averaged explainable variance shown on the flattened cortical surface of the standard fsAverage brain. EV for each subject was projected into a standard fsAverage space, and then the average EV across subjects was computed for each fsAverage vertex per language separately. Group-averaged EV is shown for each language on the flattened cortical surface of the fsAverage brain. Across subjects and languages, EV is highest within the bilateral occipital, temporal, parietal, and prefrontal cortices.

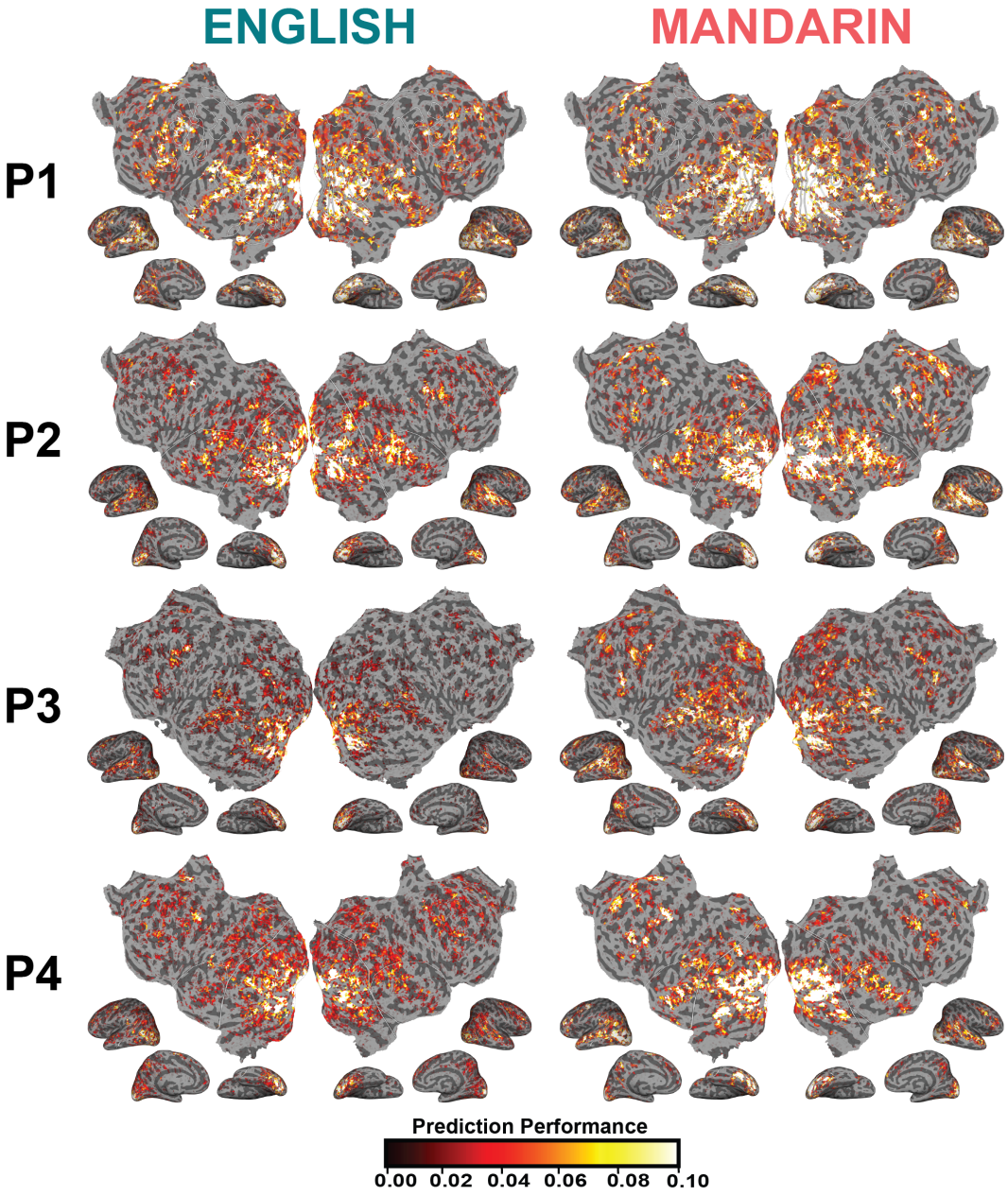


Figure C.2: **Joint prediction performance for individual participants per language per language.**

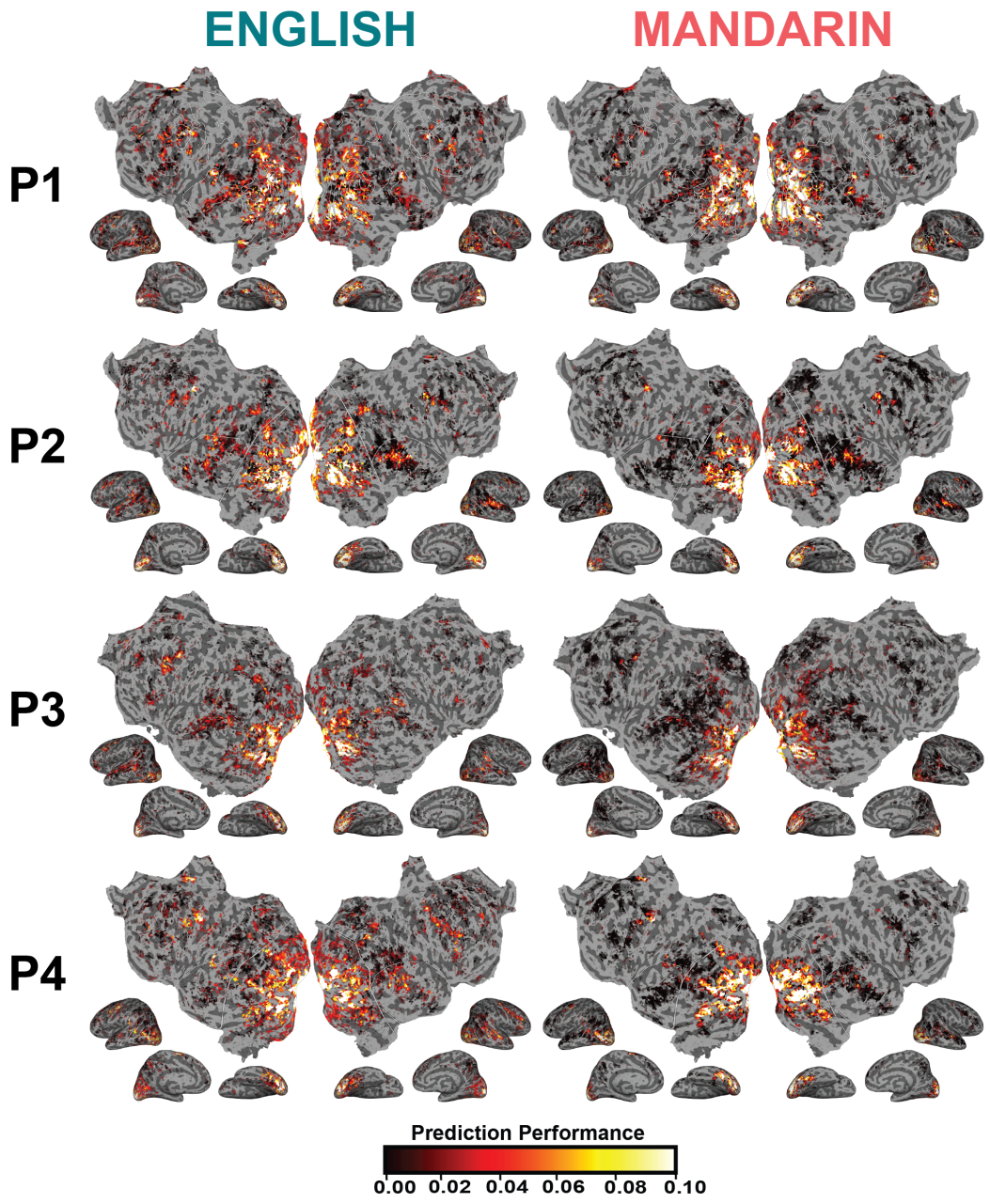


Figure C.3: **Split prediction performance of visual features for individual participants per language.**

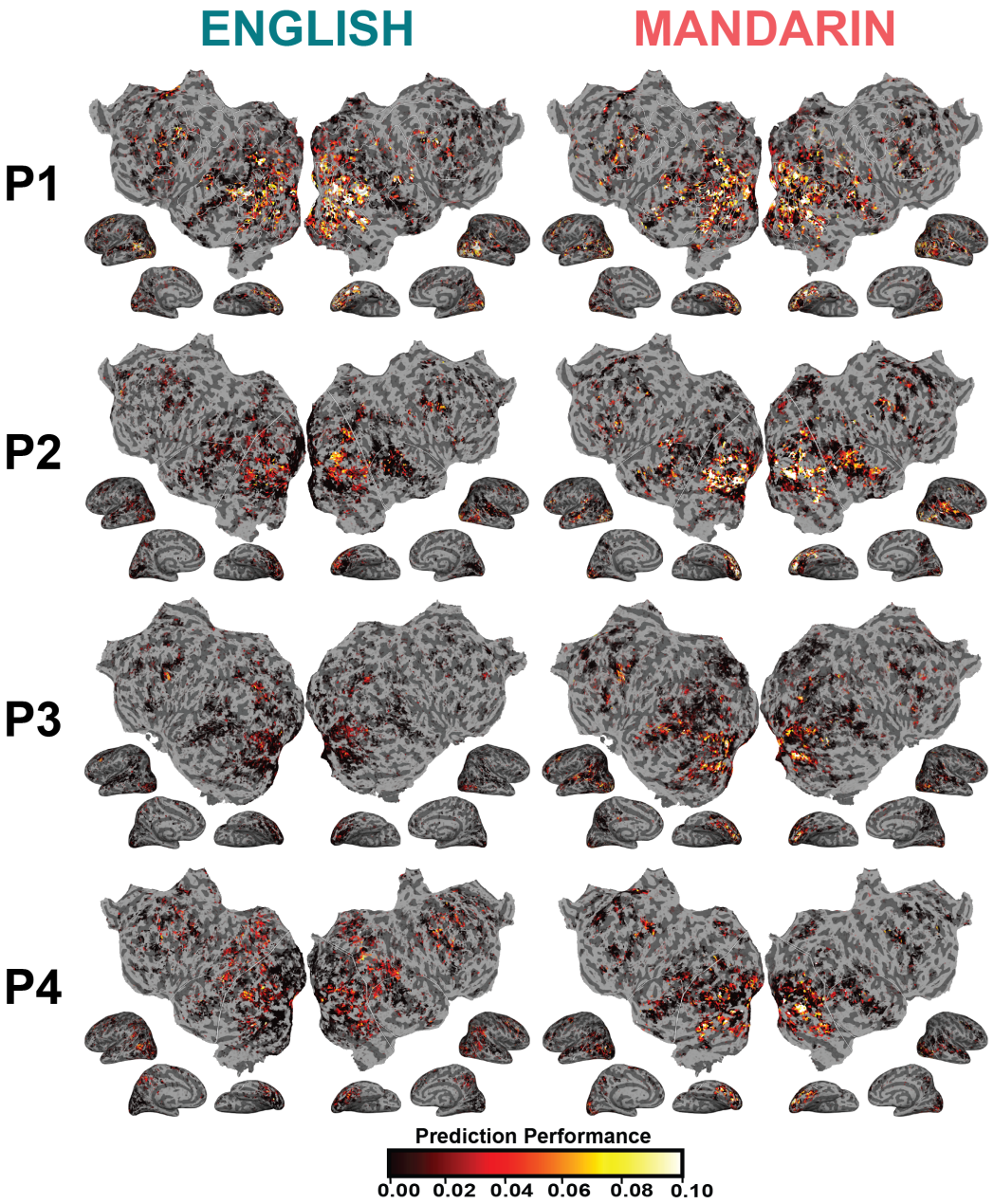


Figure C.4: **Split prediction performance of orthographic features for individual participants per language.**

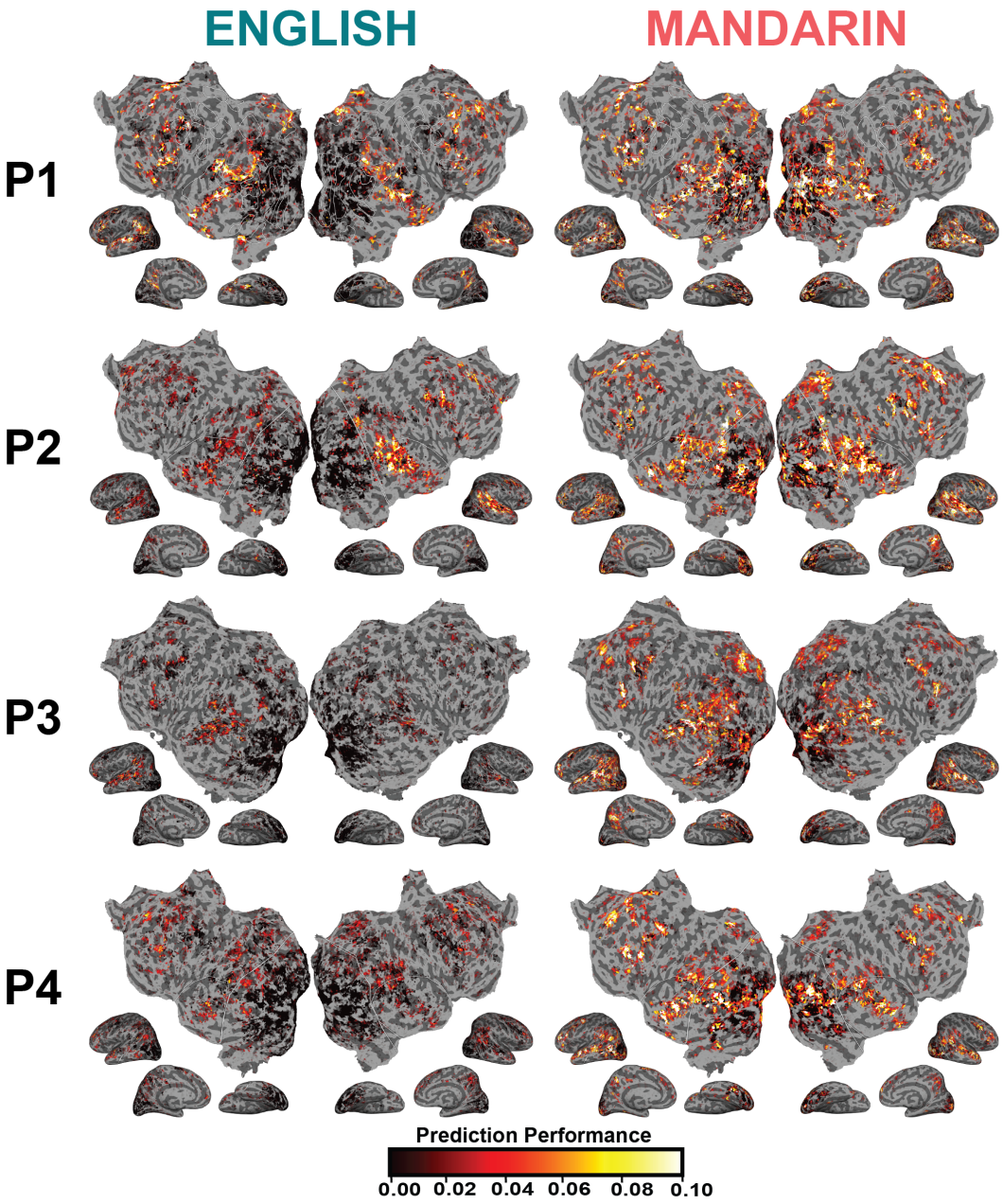


Figure C.5: **Split prediction performance of semantic features for individual participants per language.**

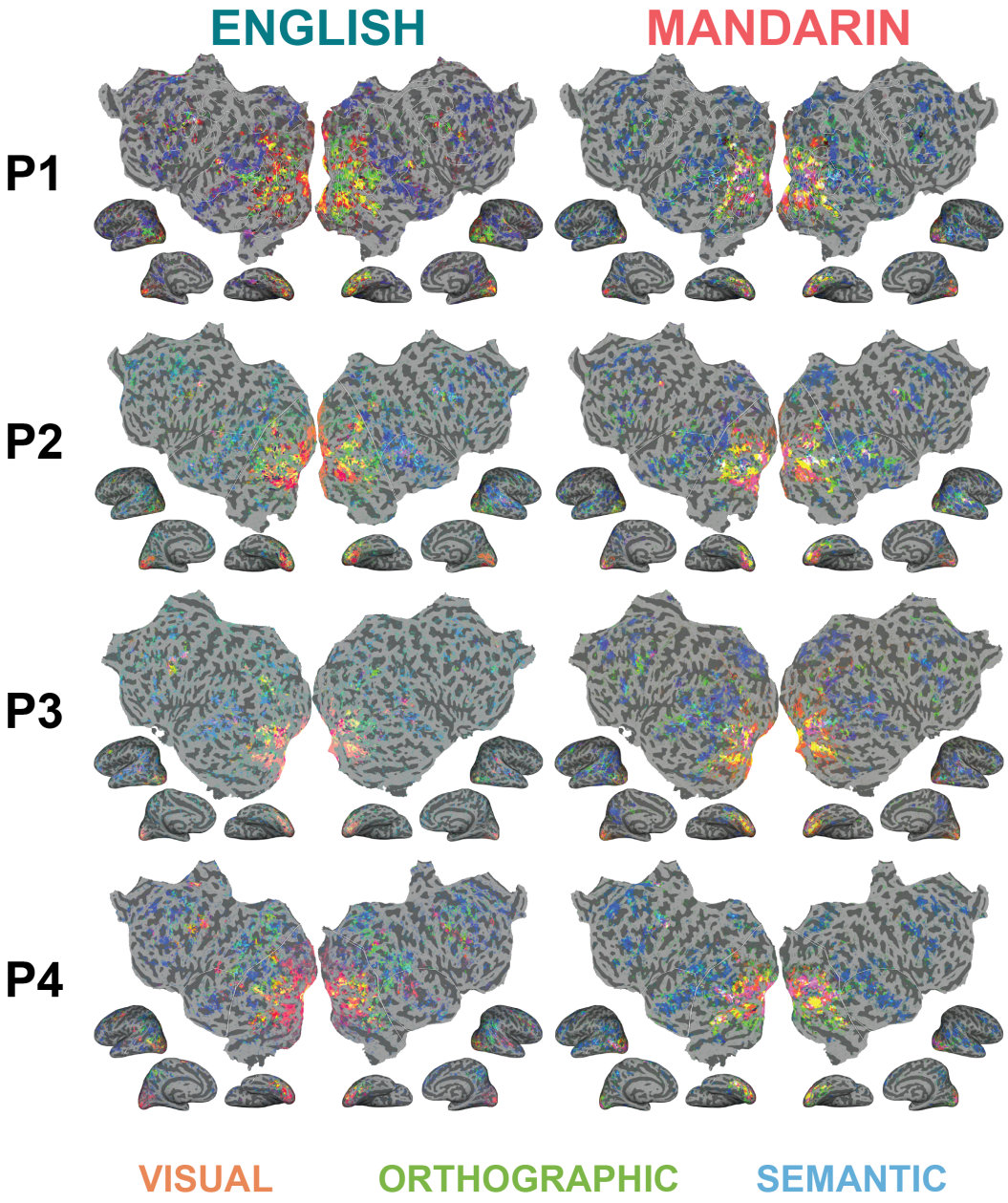


Figure C.6: **Prediction accuracy is shown for visual (red), orthographic (green) and semantic (blue) feature spaces on the flattened cortical surface of individual participants for English (left) and Mandarin (right).**

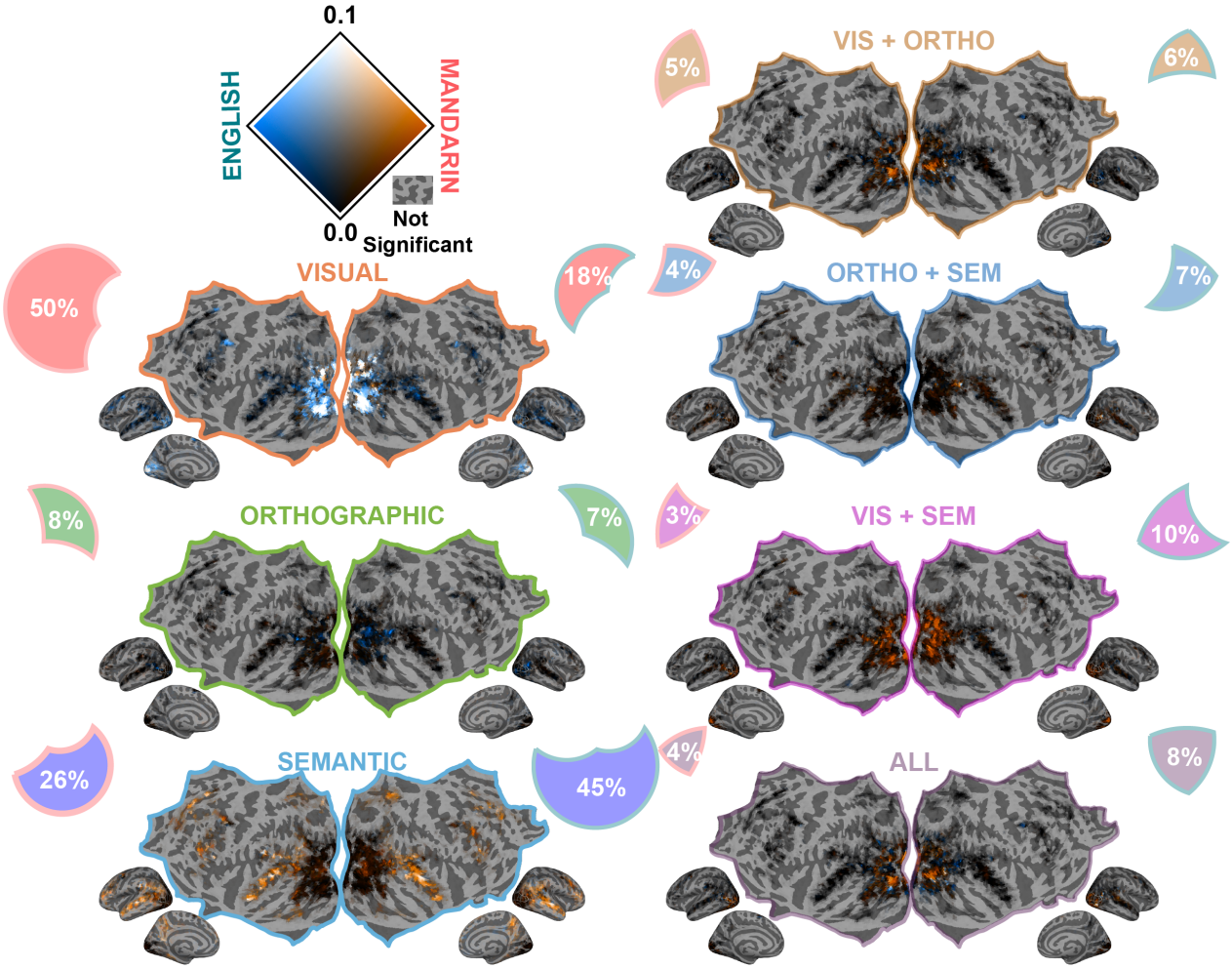


Figure C.7: **Grouped-averaged prediction accuracy for the all seven partitions in English and Mandarin are shown on the flattened cortical surface of the standard fsAverage brain with a 3D colormap.** Vertices better explained by English are shown in blue. Vertices better explained by Mandarin are shown in red. Vertices equally well explained by both English and Mandarin are shown in white.

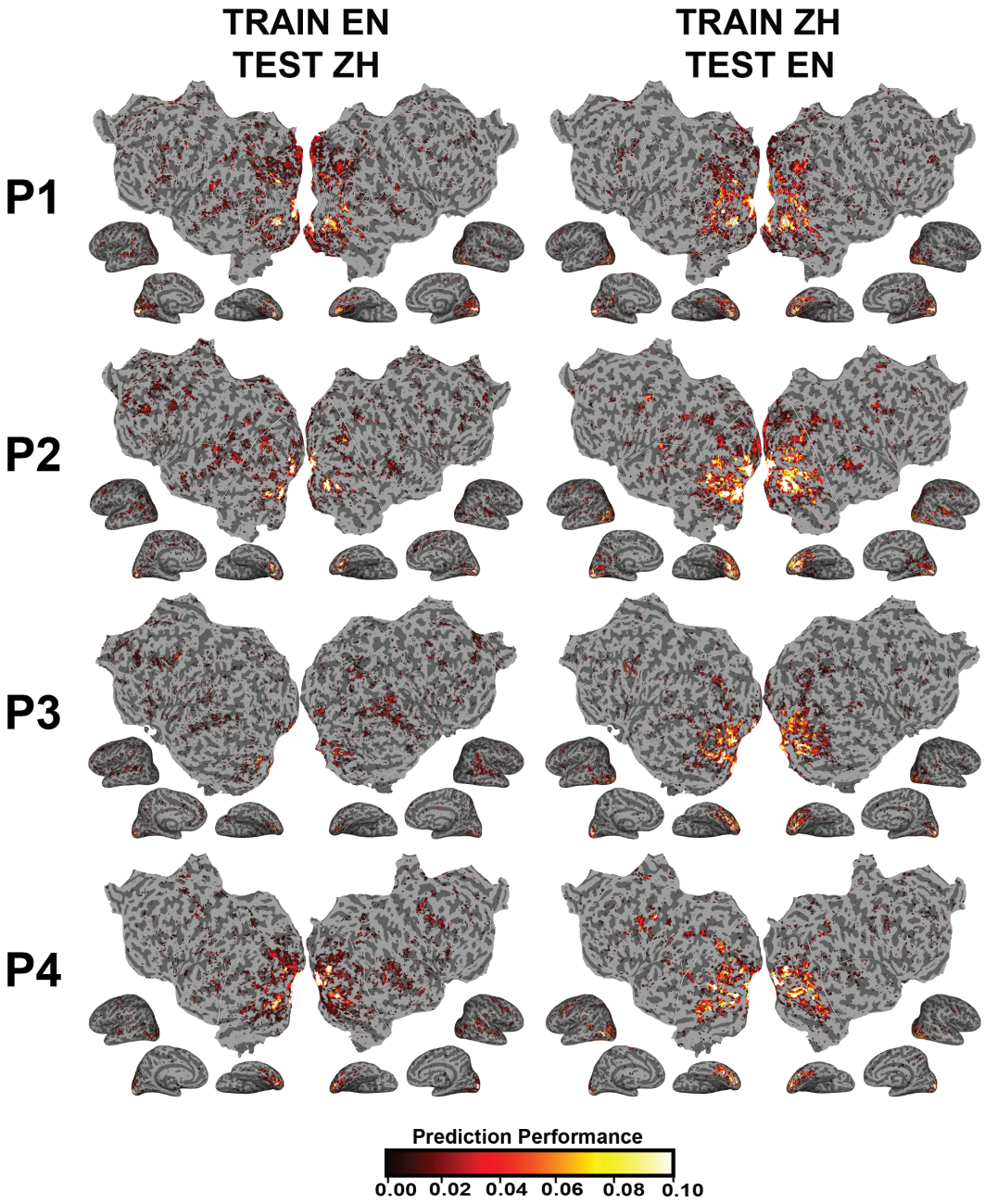


Figure C.8: **Across-language prediction accuracy of visual features for individual participants per language.**

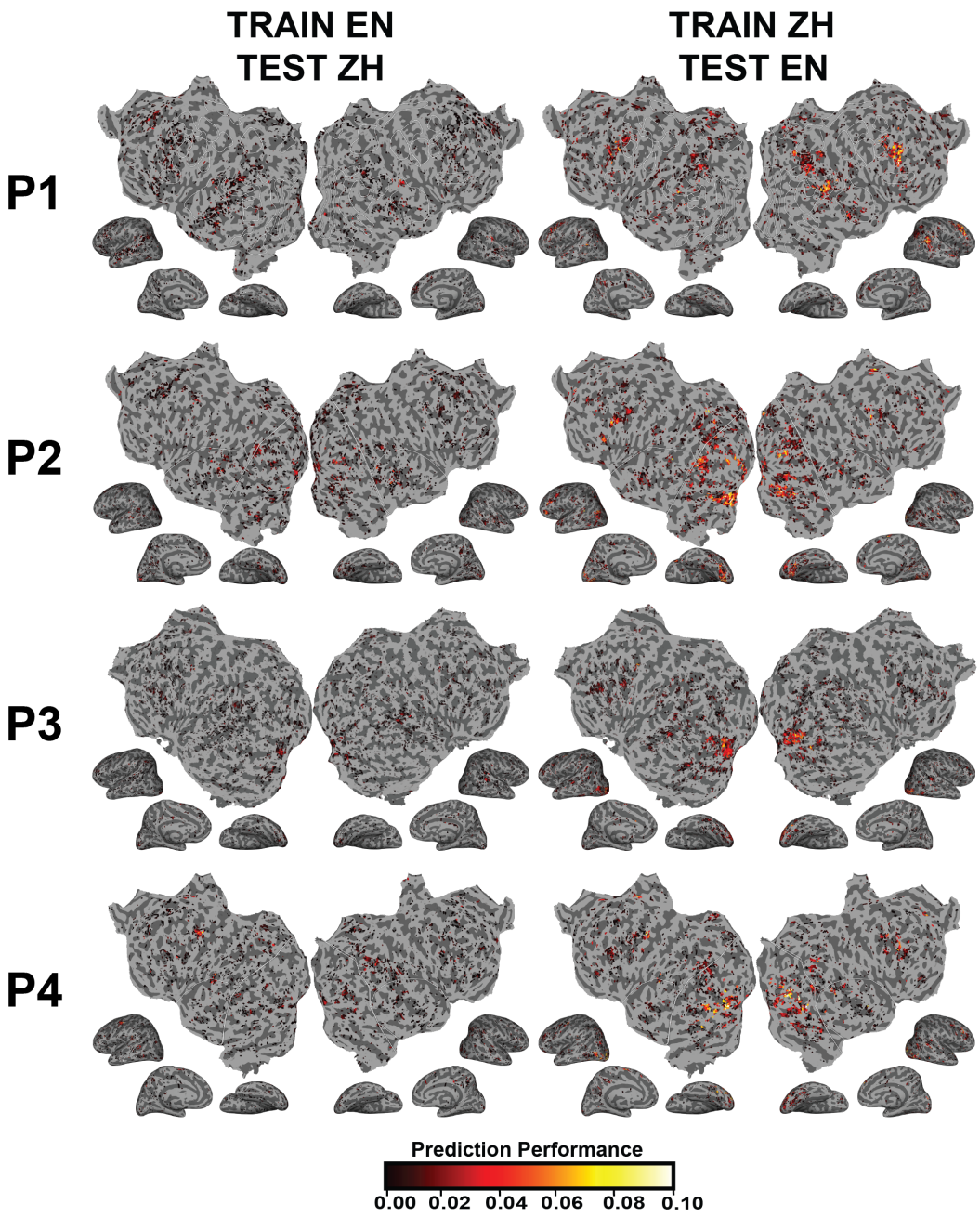


Figure C.9: **Across-language prediction accuracy of orthographic features for individual participants per language.**

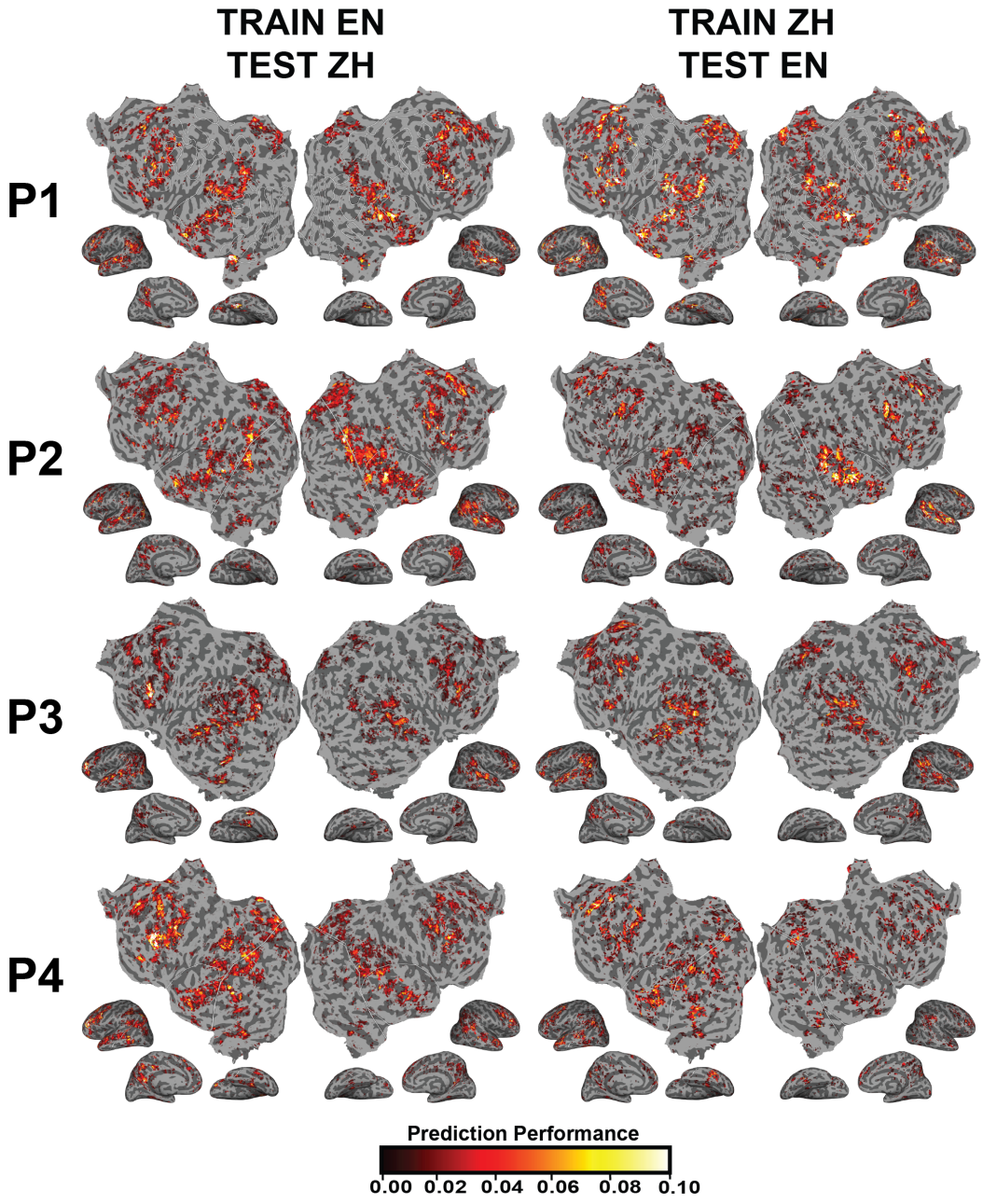


Figure C.10: **Across-language prediction accuracy of semantic features for individual participants per language.**

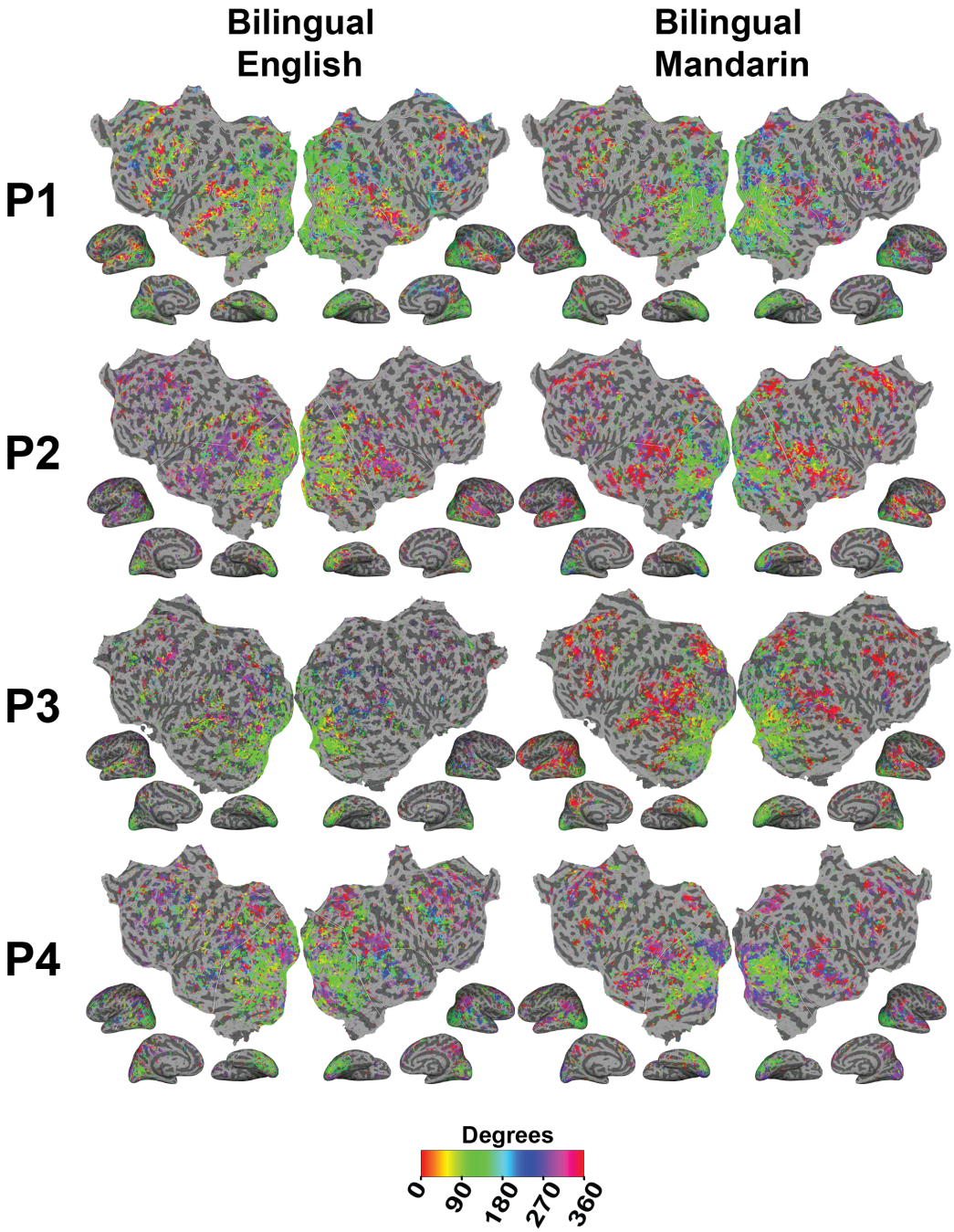


Figure C.11: **Visual features direction tuning for individual participants per language.**

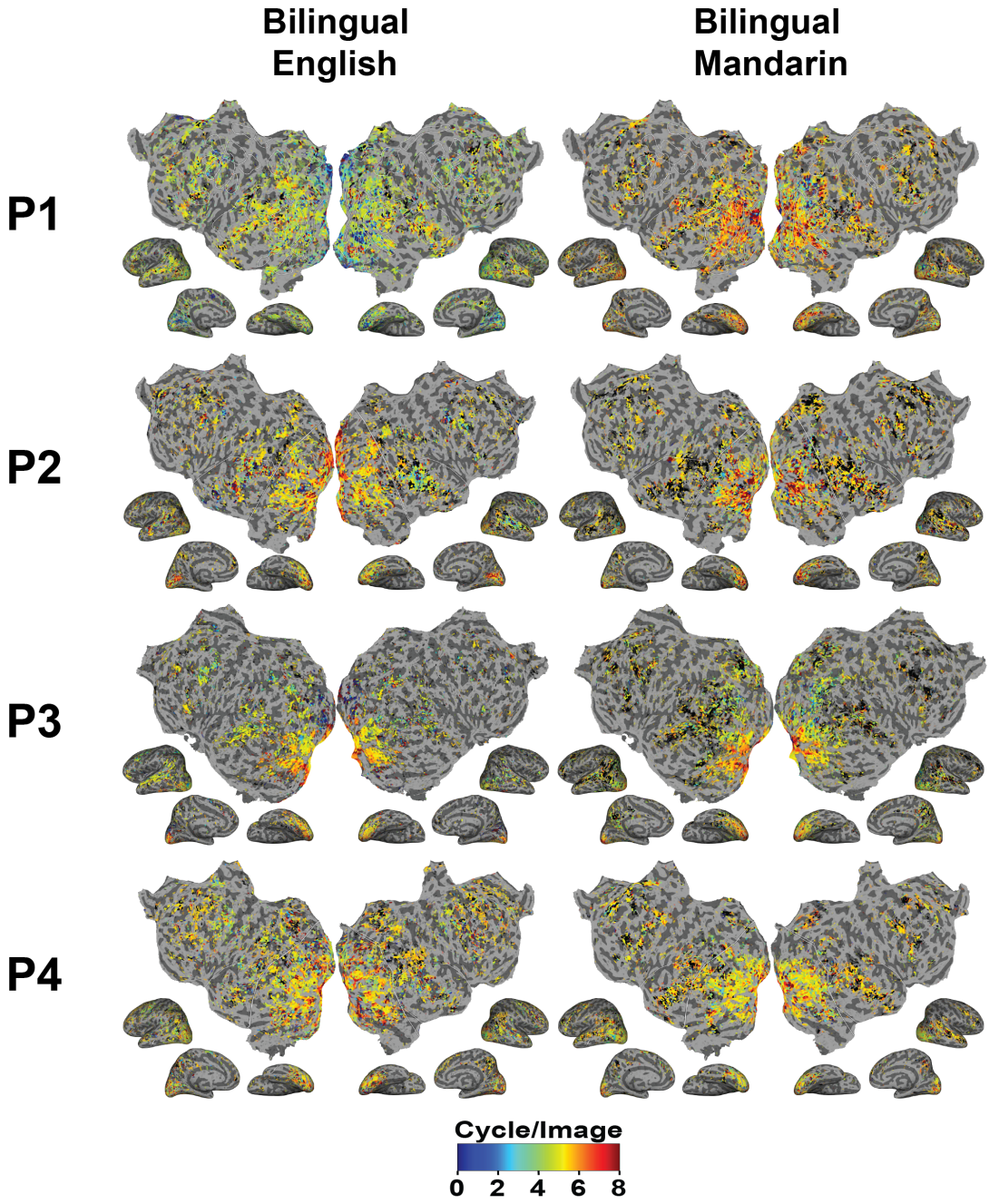


Figure C.12: **Visual features spatial frequency tuning for individual participants per language.**

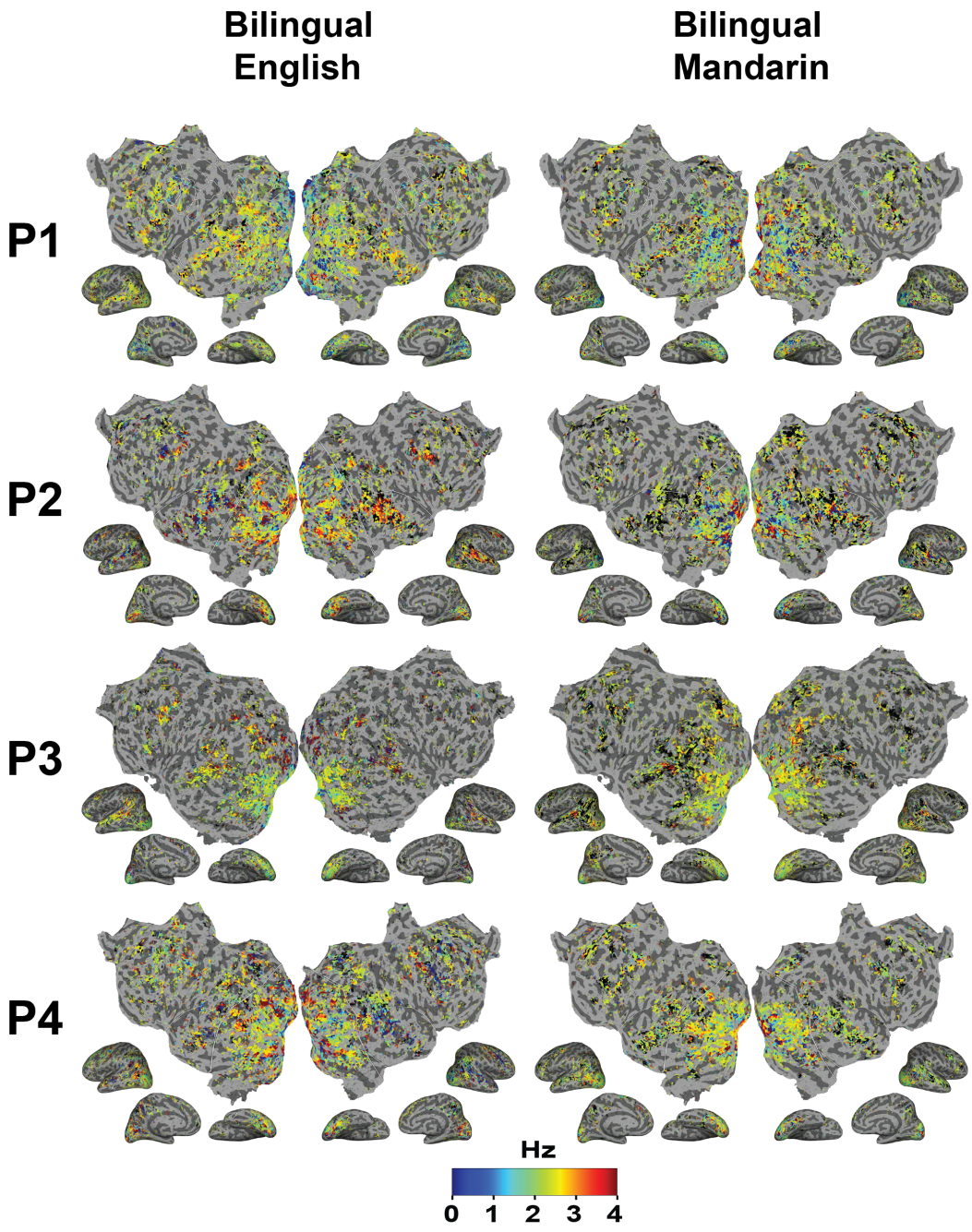


Figure C.13: **Visual features temporal frequency tuning for individual participants per language.**