Using Machine Learning for Time Series to Elucidate Sentence Processing in the Brain

Nicole S. Rafidi

AUGUST 2018
CMU-ML-18-107

Center for the Neural Basis of Cognition &
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Tom Mitchell, Chair
John Anderson
Geoffrey Gordon
Mark Richardson, UPMC Dept of Neurology
Stanislas Dehaene, Collège de France

Submitted in partial fulfillment of the requirements
for the degree of Doctorate of Philosophy in Machine Learning.

Nicole Rafidi is a PhD student in the joint Machine Learning and Neural Computation program at Carnegie Mellon University. The work in this thesis has been supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1252522, the National Institutes of Health grant number 1r01hd075328-0, IARPA afrl award number 31081.1 and the Carnegie Mellon University Richard K Mellon Fellowship

Keywords: Magnetoencephalography, machine learning, decoding, classification, representational similarity analysis, sentence processing, syntax, semantics

To my parents

Abstract

Language comprehension is a crucial human ability, but many questions remain unanswered about the processing of sentences. Specifically, how can sentences that are structured differently, e.g. "The woman helped the man." and "The man was helped by the woman." map to the same proposition? High temporal resolution neuroimaging, coupled with machine learning can potentially provide answers. Using magnetoencephalography (MEG) we can measure the activity of many neurons at a rate of 1kHz while humans read sentences. With machine learning, we can decode sentence attributes from the neural activity and gain insight into the inner computations of the brain during sentence comprehension.

We collected data from subjects reading active and passive voice sentences in two experiments: a pilot and a confirmation set The pilot set constituted a testbed for optimizing the application of machine learning to MEG data, and was used for exploratory analysis to generate data-driven hypotheses. The confirmation set allowed for confirmation of these hypotheses via replication.

Through exploration of the pilot data set, we are able to make several concrete recommendations on the optimal application of machine learning to MEG data. Specifically, we demonstrate that by combining data from multiple human subjects as additional features, classifier performance is significantly improved, even without additional data samples. Furthermore we show that while test set signal-to-noise ratio (SNR) is critical for classifier performance, training set SNR has limited impact on performance. We achieve near-perfect classification accuracy on a wide range of decoding tasks from neural activity. We also explored a non-machine learning technique, representational similarity analysis (RSA) that is quite popular for analyzing neuroimaging data, and show that by combining data across subjects we can again greatly improve performance.

We examine how sentence processing differs between active and passive sentences by showing the information flow over time during the reading of each type of sentence. We additionally explore post-sentence wrap-up activity that carries information about syntax and integrated semantics of the sentence being read. We compare the ability of models that separate syntax, semantics, and integration to explain neural activity during the post-sentence time period. Our results provide converging evidence that after a sentence is read, its syntactic structure is processed, followed by a semantic integration of sentence meaning. These results refine previous theories of sentence processing as a purely incremental process by revealing the existence of a post-sentence wrap-up period.

Acknowledgments

This thesis would not have been possible without the Pomodoro technique, invented by Francesco Cirillo. 25 minutes at a time, I put 6 years of work into writing, much of it for the first time. Also necessary for the completion of this thesis was the music of John Darnielle, which provided the background for the aforementioned writing.

Every person who ever dogsat for me has contributed to this thesis: thank you all. Thank you also to my wonderful friends who were always happy to commiserate. And thanks to everyone who told me “A good thesis is a finished thesis." I very much needed to hear it.

Thanks of course goes to my advisor, Tom, who taught me to stand on my own two feet. Last but not least, I thank my labmates, who contributed to this work both intellectually and through moral support. Best of luck to you all.

Contents

1 Introduction 1

2 Related Work 4
2.1 Terminology . . . 4
2.2 Sentence Comprehension . . . 5
2.2.1 The Neurobiology of Simple Sentence Comprehension . . . 6
2.2.2 Passive Sentences and the Need for Reanalysis . . . 8
2.3 Machine Learning for MEG Data . . . 10
2.3.1 Decoding and Encoding Models . . . 10
2.3.2 Representational Similarity Analysis . . . 11

3 Methodology 14
3.1 Decoding . . . 15
3.1.1 Basic Approach . . . 15
3.1.2 Combining Data from Multiple Subjects . . . 18
3.1.3 Choice of Classifier . . . 21
3.1.4 Treatment of Time . . . 22
3.1.5 Treatment of Trials . . . 23
3.1.6 Cross-validation Folds . . . 26
3.1.7 Optimal Hyperparameter Selection . . . 27
3.2 Representational Similarity Analysis (RSA) . . . 28
3.2.1 Basic Approach . . . 28
3.2.2 Noise Ceiling Computation . . . 30
3.2.3 Treatment of Time . . . 32
3.2.4 Partial Correlations . . . 32

4 Good Practices 36
4.1 Decoding . . . 37
4.1.1 Failure to decode indicates nothing . . . 37
4.1.2 Causal inference of classifier weights is not possible . . . 38
4.1.3 Correlation between labels affects inferences . . . 38
4.2 Language . . . 39
4.2.1 Tasks influence how participants engage with the stimuli . . . 39
4.2.2 Language features correlate with one another . . . 40

4.3 Exploratory vs Confirmatory Analysis 41

5 Pilot Data Analysis **43**

5.1 Questions of Interest 43

5.2 Methods 45

5.2.1 Data Collection 45

5.2.2 Basic Decoding Approach 45

5.3 Results 46

5.3.1 Information Flow during Sentence Reading 46

5.3.2 Post-sentence Wrap-up 47

6 Confirmatory Data Analysis **59**

6.1 Questions of Interest 60

6.1.1 Hypotheses to Confirm 60

6.1.2 Further Exploratory Questions 60

6.2 Methods 61

6.2.1 Data Collection 61

6.2.2 Decoding Experiments 62

6.2.3 RSA Experiments 63

6.3 Results 64

6.3.1 Confirmatory Decoding Results 64

6.3.2 Source Localized Decoding Results 66

6.3.3 Exploratory RSA Results 66

7 Conclusion and Future Work **83**

A Data Collection and Preprocessing Details **90**

B Pilot Stimuli **91**

C Confirmatory Stimuli **93**

Bibliography **95**

List of Figures

2.1 **Model of Auditory Language Comprehension** Taken from [1] **A. Auditory language comprehension model.** Illustration of the processing steps taken during the processing of acoustic language stimuli. Note the parallel paths for general linguistic processing (syntactic and semantic) and for prosody. **B. The brain basis of auditory language comprehension.** Brain regions activated in relation to the different processing steps outlined in panel A. 6

3.1 **Data as a function of trial averaging** MEG sensor data for the same stimulus as more trials are averaged together. Note that as more trials are averaged, the visual response to each word in the sentence (presented every 0.5 seconds) becomes more pronounced. 17

3.2 **Data difference between subjects** Data from subject B, subtracted from the data from subject I. Each subject's data is the average of 10 trials of sentence 0. Values close to 0 for a given sensor, time point entry indicate that the subjects had similar data values at that point. Large differences indicate a temporal misalignment across subjects. Vertical lines show the onset of each word and the offset of the sentence. Note that after each word, the visual response is misaligned between the two subjects. 19

3.3 **Single subject vs Multi-subject approach.** Rank accuracy TGMs computed for decoding verb identity post-sentence from the pool of all sentences (active and passive). The white vertical line indicates sentence offset. Chance performance is 0.5 and optimal performance is 1.0. Cross-validation was leave-one-sentence out. **A. Rank accuracy TGM averaged over subjects.** A separate TGM was computed for each subject, training and testing on that subject individually. The resulting TGMs were then averaged over subjects. **B. Rank accuracy TGM from multi-subject approach.** TGM generated from using the concatenation over subjects as the feature vector for classification. 20

3.4 **Algorithm Comparison** Maximum 0/1 accuracy and fraction of maximum runtime for several algorithms on the verb decoding task. Red error bars on blue bars show standard deviation across subjects in mean classifier performance. . . . 23

3.5 **Time Averaging Example** Toy example illustrating how averaging over time can make classification more difficult. The signal for the red class and the signal for the blue class have the same average value in the window (selected in green). However, if all the temporal information were used, they would be highly distinguishable. 24

3.6 **Time Treatment Comparison.** Maximum accuracy for several window sizes on the verb decoding task. Blue bars show accuracy with temporal averaging, and green bars show without. 25

3.7 **Trial Treatment Comparison.** Maximum accuracy for several ways of combining trials on the verb decoding task. Blue bars show accuracy with test set averaging, and green bars show without, i.e. when the test set has received the same treatment as the training set. 26

3.8 **Cross-validation Comparison.** Accuracy over time for three classification tasks: Voice (A), Verb (B), and Proposition (C). In each case performance was assessed with several cross-validation techniques, ranging from leave-one-out (LOO) to 2-fold cross-validation, with intermediate schemes where possible. The dashed horizontal line shows chance performance, and the vertical black line indicates sentence offset. 27

3.9 **Sample Models.** Sample model RDMs for analyzing post-sentence data. Each entry in the matrix corresponds to a pair of sentence stimuli. **A. Syntax model.** Model that assigns sentence pair a distance of 0 if both sentences are the same length and voice, 0.5 if they are the same voice but different lengths, and 1 if they are different voices. **B. Sentence length model.** Model that assigns sentence pair a distance of 0 if both sentences are the same length, otherwise a distance of 1. . 30

3.10 **RSA Approach Comparison.** Rank correlations between length and syntax models and neural activity over time. Gray shaded region gives the noise ceilling. **A. Traditional RSA.** Models are correlated to individual subject data RDMs and reported correlation is the average over those scores. Noise ceiling is computed in the tradition single-subject manner. **B. Repetition-based RSA** Proposed novel approach to RSA. Models are correlated with the mean data RDM over subjects. Noise ceiling lower bound is computed by correlating an RDM constructed from half the trials with an RDM constructed from the other half, for all possible splittings. Upper bound is computed by correlating the half-trial RDMs with the full mean RDM. 31

3.11 **Time Treatment Comparison.** Mean of noise ceiling lower and upper bound over post-sentence time period, for each window size, both averaging and not averaging over the timepoints in that window. 32

3.12 **Correlation Type Comparison.** Comparison of max model correlations over time with neural activity over the post-sentence time period. Noise ceiling is given in gray. On the left are the zero-order correlations between each model and the neural activity. On the right we have the conditional correlations, where each model is conditioned *on the other model*. 34

5.1 **Rank accuracy TGMs for each during-sentence classification task.** Each plot shows all pairs of training and testing timepoints over sentence presentation. The y axis indicates training time, and the x axis indicates test time, with sentence onset starting in the upper left corner. White vertical lines indicate word onsets and final sentence offset. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on either the set of active or passive sentences separately, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out. **A. TGM for decoding the first noun from active-voice sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on active voice sentences. **B. TGM for decoding the verb from active voice sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on active voice sentences. **C. TGM for decoding the second noun from active voice sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on active voice sentences. **D. TGM for decoding the first noun from passive voice sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on passive voice sentences. **E. TGM for decoding the verb from passive voice sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on passive voice sentences. **F. TGM for decoding the second noun from passive voice sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on passive voice sentences. 50

5.2 **Rank accuracy over time during sentence reading.** Each plot shows rank accuracy over time for each word in the sentence. Black vertical lines indicate word onsets and final sentence offset. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on either the set of active or passive sentences separately, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out. **A. Information flow during active voice sentence reading.** Accuracy for decoding each of the constituent words from active sentences. **B. Information flow during passive voice sentence reading.** Accuracy for decoding each of the constituent words from passive sentences. 51

5.3 **Helmet plots for each during-sentence classification task at 0.15s post word onset.** Each plot shows the resulting importance map from training a classifier 0.15s post onset of the word of interest. Importance maps were computed from the classifier weights on the concatenation of all subjects' mean sensor activity from 0.15 to 0.25 s post onset as described in [2]. Maps were then averaged over subjects. Importance values for a single gradiometer are shown. Titles give the rank accuracy at the time examined. **A. Importance map for decoding first noun from active sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on active voice sentences. **B. Importance map for decoding verb from active sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on active voice sentences. **C. Importance map for decoding second noun from active sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on active voice sentences. **D. Importance map for decoding first noun from passive sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on passive voice sentences. **E. Importance map for decoding verb from passive sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on passive voice sentences. **F. Importance map for decoding second noun from passive sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on passive voice sentences. 52

5.4 **Helmet plots for each during-sentence classification task at 0.3s post word onset.** Each plot shows the resulting importance map from training a classifier 0.3s post onset of the word of interest. Importance maps were computed from the classifier weights on the concatenation of all subjects' mean sensor activity from 0.3 to 0.4 s post onset as described in [2]. Maps were then averaged over subjects. Importance values for a single gradiometer are shown. Titles give the rank accuracy at the time examined. **A. Importance map for decoding first noun from active sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on active voice sentences. **B. Importance map for decoding verb from active sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on active voice sentences. **C. Importance map for decoding second noun from active sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on active voice sentences. **D. Importance map for decoding first noun from passive sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on passive voice sentences. **E. Importance map for decoding verb from passive sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on passive voice sentences. **F. Importance map for decoding second noun from passive sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on passive voice sentences. 53

5.5 **Decodability of determiners over time.** F1 score when decoding three tasks: ‘a’ vs ‘the’, ‘a’ vs ‘dog’ and ‘the’ vs ‘dog’ over word presentation time. F1 score was used instead of classification accuracy because there are many more instances of the determiners than of ‘dog’. Classifier was balanced during training by weighting the samples, so as to eliminate bias towards the more common class. 54

5.6 **Information flow post-sentence.** Each plot shows rank accuracy over time for each post-sentence classification task. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on either the set of active and passive sentences separately as well as the pool of all sentences, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out. The black vertical line indicates the offset of the second noun (the end of the sentence). **A. Decoding accuracy post active-sentences.** Classification accuracy of agent, verb and patient training and testing on active sentences only. **B. Decoding accuracy post passive-sentences.** Classification accuracy of agent, verb and patient training and testing on passive sentences only. **C. Decoding accuracy post all sentences.** Classification accuracy of agent, verb, patient, voice, first noun, and proposition, training and testing on the pool of all sentences. 55

5.7 **Rank accuracy TGMs for each post-sentence classification task.** Each plot shows all pairs of training and testing timepoints from second noun presentaiton onwards. The y axis indicates training time, and the x axis indicates test time, with word onset starting in the upper left corner. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on the pool of both active and passive sentences, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out.The white line indicates the offset of the second noun (the end of the sentence). **A. TGM for decoding the agent of the sentence.** Classification task was to detect the identity of the agent. In active sentences, this is the first noun, and in passive sentences it is the second noun. **B. TGM for decoding the patient of the sentence.** Classification task was to detect the identity of the patient. In active sentences, this is the second noun, and in passive sentences it is the first noun. **C. TGM for decoding the verb of the sentence.** Classification task was to detect the identity of the verb. **D. TGM for decoding the first noun of the sentence.** Classification task was to detect the identity of the first noun. In active sentences this is the agent, and in passive sentences it is the patient. **E. TGM for decoding the voice of the sentence.** Classification task was to detect the voice of the sentence (active or passive). **F. TGM for decoding the proposition of the sentence.** Classification task was to detect the proposition of the sentence (where the active and passive version of the proposition were given the same class label). 56

5.8 **Helmet plots for each during-sentence classification task at 0.61s post last word onset.** Each plot shows the resulting importance map from training a classifier 0.61s post onset of the last word in the sentence. Importance maps were computed from the classifier weights on the concatenation of all subjects' mean sensor activity from 0.61 to 0.71 s post onset as described in [2]. Maps were then averaged over subjects. Importance values for a single gradiometer are shown. Classification tasks were conducted on the pool of both active and passive sentences. Titles give the rank accuracy at the time examined. **A. Importance map for decoding the agent.** Classification task was to detect the identity of the agent of the sentence, which in passive sentences is the second noun and in active sentences is the first noun. **B. Importance map for decoding the patient.** Classification task was to detect the identity of the patient of the sentence, which in passive sentences is the first noun and in active sentences is the second noun. **C. Importance map for decoding verb.** Classification task was to detect the identity of the verb of the sentence. **D. Importance map for decoding first noun.** Classification task was to detect the identity of the first noun of the sentence. **E. Importance map for decoding sentence voice.** Classification task was to detect the voice (active or passive) of the sentence. **F. Importance map for decoding proposition.** Classification task was to detect the identity of the proposition of the sentence. 57

5.9 **Rank accuracy TGMs crossing noun presentations for animacy decoding.** Each plot shows a rank accuracy TGM for a pair of training and testing scenarios from which we attempted to decode noun animacy. The y axis indicates training time, and the x axis indicates test time, with word onset starting in the upper left corner. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on the pool of both active and passive sentences, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out. **A. Training on first noun, testing on first noun.** Animacy decoding results when training and testing on data from first noun presentation. **B. Training on first noun, testing on second noun.** Animacy decoding results when training on data from first noun presentation (agnostic to sentence length) and testing on data from second noun presentation. **C. Training on second noun, testing on first noun.** Animacy decoding results when training on data from second noun presentation and testing on data from first noun presentation. **D. Training on second noun, testing on second noun.** Animacy decoding results when training on data from second noun presentation and testing on data from second noun presentation. 58

6.1 **Model RDMs.** Model Representational Dissimilarity Matrices (RDMs) for exploratory RSA analysis. Each element of an RDM is the distance between a pair of stimuli according to a theoretical model. The rows are ordered active sentences first, followed by passive sentences. Sentence length alternates every four rows. **A. Syntax RDM.** Each entry is determined by the voice and the length of the given sentence pair. A different in voice is assigned a distance of 1, while a difference in length is assigned a distance of 0.5 if the voices are the same. **B. Bag of Words RDM.** Each sentence is represented by the average of the GloVe vectors for the constituent words. The entries of this RDM are the euclidean distances between these sentence representations. **C. Hierarchical RDM.** Each entry is the average of the pairwise distances between the GloVe vectors for the agent, patient, and verb. For short sentences, the missing noun (agent or patient) is represented by the average over the nouns in the experiment . 63

6.2 **Rank accuracy TGMs for each during-sentence classification task.** Each plot shows all pairs of training and testing timepoints over sentence presentation. The y axis indicates training time, and the x axis indicates test time, with sentence onset starting in the upper left corner. White vertical lines indicate word onsets and final sentence offset. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on either the set of active or passive sentences separately, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out. **A. TGM for decoding the first noun from active-voice sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on active-voice sentences. **B. TGM for decoding the verb from active-voice sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on active-voice sentences. **C. TGM for decoding the second noun from active-voice sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on long active-voice sentences. **D. TGM for decoding the first noun from passive-voice sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on passive-voice sentences. **E. TGM for decoding the verb from passive-voice sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on passive-voice sentences. **F. TGM for decoding the second noun from passive-voice sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on long passive-voice sentences. 69

6.3 **Rank accuracy over time during sentence reading.** Each plot shows rank accuracy over time for each word in the sentence. Black vertical lines indicate word onsets and final sentence offset. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on either the set of active or passive sentences separately, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out. **A. Information flow during active-voice sentence reading.** Accuracy for decoding each of the constituent words from active sentences. **B. Information flow during passive-voice sentence reading.** Accuracy for decoding each of the constituent words from passive sentences. 70

6.4 **Helmet plots for each during-sentence classification task at 0.15s post word onset.** Each plot shows the resulting importance map from training a classifier 0.15s post onset of the word of interest. Importance maps were computed from the classifier weights on the concatenation of all subjects' mean sensor activity from 0.15 to 0.25 s post onset as described in [2]. Maps were then averaged over subjects. Importance values for a single gradiometer are shown. Titles give the rank accuracy at the time examined. **A. Importance map for decoding first noun from active sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on active-voice sentences. **B. Importance map for decoding verb from active sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on active-voice sentences. **C. Importance map for decoding second noun from active sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on long active-voice sentences. **D. Importance map for decoding first noun from passive sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on passive-voice sentences. **E. Importance map for decoding verb from passive sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on passive-voice sentences. **F. Importance map for decoding second noun from passive sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on long passive-voice sentences. 71

6.5 **Helmet plots for each during-sentence classification task at 0.3s post word onset.** Each plot shows the resulting importance map from training a classifier 0.3s post onset of the word of interest. Importance maps were computed from the classifier weights on the concatenation of all subjects' mean sensor activity from 0.3 to 0.4 s post onset as described in [2]. Maps were then averaged over subjects. Importance values for a single gradiometer are shown. Titles give the rank accuracy at the time examined. **A. Importance map for decoding first noun from active sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on active-voice sentences. **B. Importance map for decoding verb from active sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on active-voice sentences. **C. Importance map for decoding second noun from active sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on long active-voice sentences. **D. Importance map for decoding first noun from passive sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on passive-voice sentences. **E. Importance map for decoding verb from passive sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on passive-voice sentences. **F. Importance map for decoding second noun from passive sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on long passive-voice sentences. 72

6.6 **Information flow post-sentence.** Each plot shows rank accuracy over time for each post-sentence classification task. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on either the set of active and passive sentences separately as well as the pool of all sentences, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out. The black vertical line indicates the offset of the last word (the end of the sentence). **A. Decoding accuracy post active-sentences.** Classification accuracy of agent, verb and patient training and testing on long active sentences only. **B. Decoding accuracy post passive-sentences.** Classification accuracy of agent, verb and patient training and testing on long passive sentences only. **C. Decoding accuracy post all sentences.** Classification accuracy of sentence length, agent, verb, patient, voice, first noun, and proposition, training and testing on the pool of all sentences. Agent, patient and proposition decoding tasks were conducted only on the set of long sentences. 73

6.7 **Rank accuracy TGMs for each post-sentence classification task.** Each plot shows all pairs of training and testing timepoints over sentence presentation. The y axis indicates training time, and the x axis indicates test time, with sentence onset starting in the upper left corner. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on the pool of both active and passive sentences, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out.The white line indicates the offset of the second noun (the end of the sentence). **A. TGM for decoding the agent of the sentence.** Classification task was to detect the identity of the agent from the pool of long sentences. In active sentences, this is the first noun, and in passive sentences it is the second noun. **B. TGM for decoding the patient of the sentence.** Classification task was to detect the identity of the patient from the pool of long sentences. In active sentences, this is the second noun, and in passive sentences it is the first noun. **C. TGM for decoding the verb of the sentence.** Classification task was to detect the identity of the verb from the pool of all sentences (short and long). **D. TGM for decoding sentence length** Classification task was to detect whether the sentence was long or short (two nouns or one noun). **E. TGM for decoding the first noun of the sentence.** Classification task was to detect the identity of the first noun from the pool of all sentences (short and long). In active sentences this is the agent, and in passive sentences it is the patient. **F. TGM for decoding the voice of the sentence.** Classification task was to detect the voice of the sentence (active or passive) from the pool of all sentences (short and long). **G. TGM for decoding the proposition of the sentence.** Classification task was to detect the proposition of the sentence (where the active and passive version of the proposition were given the same class label) from the pool of long sentences. 74

6.8 **Helmet plots for each during-sentence classification task at 1.44s post last word onset.** Each plot shows the resulting importance map from training a classifier 1.44s post onset of the last word in the sentence. Importance maps were computed from the classifier weights on the concatenation of all subjects' mean sensor activity from 1.44 to 1.54 s post onset as described in [2]. Maps were then averaged over subjects. Importance values for a single gradiometer are shown. Classification tasks were conducted on the pool of both active and passive sentences. Titles give the rank accuracy at the time examined. **A. Importance map for decoding the agent.** Classification task was to detect the identity of the agent of the sentence, which in passive sentences is the second noun and in active sentences is the first noun, from the pool of long sentences. **B. Importance map for decoding the patient.** Classification task was to detect the identity of the patient of the sentence, which in passive sentences is the first noun and in active sentences is the second noun, from the pool of long sentences. **C. Importance map for decoding verb.** Classification task was to detect the identity of the verb of the sentence, from the pool of long sentences. **D. Importance map for decoding sentence length** Classification task was to detect whether the sentence was long or short (two nouns or one noun). **E. Importance map for decoding first noun.** Classification task was to detect the identity of the first noun of the sentence. **F. Importance map for decoding sentence voice.** Classification task was to detect the voice (active or passive) of the sentence. **G. Importance map for decoding proposition.** Classification task was to detect the identity of the proposition of the sentence from the pool of long sentences. 75

6.9 **Voice Decoding Accuracy by Region and Time.** Rank Accuracy for decoding per-ROI over time for the voice decoding task. **A. Heatmap of accuracy by ROI and Time, Left Hemisphere** Each point in the grid is the rank accuracy for voice decoding at a given region and time point. The y-axis lists left hemisphere regions (from the Freesurfer atlas) and the x-axis shows time relative to last word onset. The vertical lines divide the post-sentence time period into 4 500ms periods. **B. Heatmap of accuracy by ROI and Time, Right Hemisphere** Same as in **A** but for right hemisphere regions. **C. Inflated Left Hemisphere Brain Maps of Accuracy for 500ms Periods.** For each 500ms window indicated in **A**, we took the max over time for each region and plotted it on the inflated brain map. **D. Inflated Right Hemisphere Brain Maps of Accuracy for 500ms Periods.** Same as in **C** but for right hemisphere regions. 76

6.10 **Verb Decoding Accuracy by Region and Time.** Rank Accuracy for decoding per-ROI over time for the verb decoding task. **A. Heatmap of accuracy by ROI and Time, Left Hemisphere** Each point in the grid is the rank accuracy for verb decoding at a given region and time point. The y-axis lists left hemisphere regions (from the Freesurfer atlas) and the x-axis shows time relative to last word onset. The vertical lines divide the post-sentence time period into 4 500ms periods. **B. Heatmap of accuracy by ROI and Time, Right Hemisphere** Same as in **A** but for right hemisphere regions. **C. Inflated Left Hemisphere Brain Maps of Accuracy for 500ms Periods.** For each 500ms window indicated in **A**, we took the max over time for each region and plotted it on the inflated brain map. **D. Inflated Right Hemisphere Brain Maps of Accuracy for 500ms Periods.** Same as in **C** but for right hemisphere regions. 77

6.11 **Agent Decoding Accuracy by Region and Time.** Rank Accuracy for decoding per-ROI over time for the agent decoding task. **A. Heatmap of accuracy by ROI and Time, Left Hemisphere** Each point in the grid is the rank accuracy for agent decoding at a given region and time point. The y-axis lists left hemisphere regions (from the Freesurfer atlas) and the x-axis shows time relative to last word onset. The vertical lines divide the post-sentence time period into 4 500ms periods. **B. Heatmap of accuracy by ROI and Time, Right Hemisphere** Same as in **A** but for right hemisphere regions. **C. Inflated Left Hemisphere Brain Maps of Accuracy for 500ms Periods.** For each 500ms window indicated in **A**, we took the max over time for each region and plotted it on the inflated brain map. **D. Inflated Right Hemisphere Brain Maps of Accuracy for 500ms Periods.** Same as in **C** but for right hemisphere regions. 78

6.12 **Patient Decoding Accuracy by Region and Time.** Rank Accuracy for decoding per-ROI over time for the patient decoding task. **A. Heatmap of accuracy by ROI and Time, Left Hemisphere** Each point in the grid is the rank accuracy for patient decoding at a given region and time point. The y-axis lists left hemisphere regions (from the Freesurfer atlas) and the x-axis shows time relative to last word onset. The vertical lines divide the post-sentence time period into 4 500ms periods. **B. Heatmap of accuracy by ROI and Time, Right Hemisphere** Same as in **A** but for right hemisphere regions. **C. Inflated Left Hemisphere Brain Maps of Accuracy for 500ms Periods.** For each 500ms window indicated in **A**, we took the max over time for each region and plotted it on the inflated brain map. **D. Inflated Right Hemisphere Brain Maps of Accuracy for 500ms Periods.** Same as in **C** but for right hemisphere regions. 79

6.13 **Proposition Decoding Accuracy by Region and Time.** Rank Accuracy for decoding per-ROI over time for the proposition decoding task. **A. Heatmap of accuracy by ROI and Time, Left Hemisphere** Each point in the grid is the rank accuracy for proposition decoding at a given region and time point. The y-axis lists left hemisphere regions (from the Freesurfer atlas) and the x-axis shows time relative to last word onset. The vertical lines divide the post-sentence time period into 4 500ms periods. **B. Heatmap of accuracy by ROI and Time, Right Hemisphere** Same as in **A** but for right hemisphere regions. **C. Inflated Left Hemisphere Brain Maps of Accuracy for 500ms Periods.** For each 500ms window indicated in **A**, we took the max over time for each region and plotted it on the inflated brain map. **D. Inflated Right Hemisphere Brain Maps of Accuracy for 500ms Periods.** Same as in **C** but for right hemisphere regions. . . 80

6.14 **Argument Binding Decoding Accuracy by Region and Time.** Rank Accuracy for decoding per-ROI over time for the argument binding decoding task. **A. Heatmap of accuracy by ROI and Time, Left Hemisphere** Each point in the grid is the rank accuracy for argument binding decoding at a given region and time point. The y-axis lists left hemisphere regions (from the Freesurfer atlas) and the x-axis shows time relative to last word onset. The vertical lines divide the post-sentence time period into 4 500ms periods. **B. Heatmap of accuracy by ROI and Time, Right Hemisphere** Same as in **A** but for right hemisphere regions. **C. Inflated Left Hemisphere Brain Maps of Accuracy for 500ms Periods.** For each 500ms window indicated in **A**, we took the max over time for each region and plotted it on the inflated brain map. **D. Inflated Right Hemisphere Brain Maps of Accuracy for 500ms Periods.** Same as in **C** but for right hemisphere regions. 81

6.15 **Whole brain RSA model comparison.** Spearman correlation between each model RDM in Fig. 6.1 and a MEG data RDM constructed from whole brain sensor-level data. Noise ceiling is shown in gray. Corrected significance is indicated via starred points. 82

7.1 **Computations underlying the reanalysis signal.** Diagrams demonstrating the expected decoding result under several hypotheses for reanalysis, as well as the true decoding result. Colored blocks indicate significantly above-chance decoding. **A. Hypothesis 1: No difference in post-sentence word decoding.** Under this hypothesis, there is no difference between active and passive sentences in terms of which of the constituent sentence words can be decoded. It is unclear what the reanalysis signal can be attributed to in this case aside from some purely syntactic function. **B. Hypothesis 2: Patient reactivation in passive sentences only.** Under this hypothesis, part of the reanalysis signal can be explained by the fact that the first noun in passive sentences must switch its role from agent to patient, leading to a reactivation. Since no reactivation is necessary in active sentences, we will not see any decodability of the first noun in that case. **C. Hypothesis 3: Verb reactivation in passive sentences only.** Under this hypothesis, it is not the noun's role that must be switched, but rather the argument bindings of the verb, leading to selective verb reactivation. **D. Results.** Results show that the patient is reactivated in passive sentences, while the agent is not reactivated in active sentences, conforming to Hypothesis 2 (panel B). 85

7.2 **Regions that maximally decode agent and patient identity.** During the post-sentence time period, we decoded agent and patient identity from the pool of all sentences (active and passive) for each ROI and each time point. In red are the regions for which agent decoding was maximal over the time period, and in blue are the regions where the patient was most decodable, with purple indicating overlap. Accuracy was thresholded at 0.8 for inclusion in this map. 86

7.3 **Post-sentence wrap-up.** Diagrams demonstrating the expected decoding result under several hypotheses of post-sentence wrap-up, as well as the true decoding result. Colored blocks indicate significantly above-chance decoding. **A. Hypothesis 1: Semantics-first sentence wrap-up.** Under this hypothesis, the a bag-of-words representation of the sentence is detectable (e.g. the component words), followed by the syntax and then the integration of syntax and semantics to get the proposition. **B. Hypothesis 2: Simultaneous sentence wrap-up.** Under this hypothesis, syntax and semantics are integrated simultaneously to generate the proposition. **C. Hypothesis 3: Syntax-first sentence wrap-up.** Under this hypothesis, syntax information is processed first, followed by semantic and then the final integration. **D. Results.** Results support a mixture of Hypothesis 2 and 3. Sentence voice and length (syntactic properties) are decodable first, along with the verb identity. Then, the proposition and agent become decodable as well. At the very end of the post-sentence time period, all sentence attributes are decodable. 87

7.4 **Regions that maximally decode verb, voice and proposition identity.** During the post-sentence time period, we decoded voice, verb, and proposition identity from the pool of all sentences (active and passive) for each ROI and each time point. In red are the regions for which voice decoding was maximal over the time period, and in blue are the regions where the verb was most decodable. Green represents maximal proposition decoding. Accuracy was thresholded at 0.8 for inclusion in this map. 88

List of Tables

3.1 **Basic decoding approach** 18
3.2 **Optimized decoding approach** 28
3.3 **Basic RSA approach** 29

4.1 **Sample Scientific Questions and Corresponding Prediction Tasks** 36

6.1 **Sample Stimuli from Confirmatory Experiment** 61

B.1 Active sentences used in the experiment. 91
B.2 Passive sentences used in the experiment. 91
B.3 Comprehension questions used in the Pilot experiment. 92

C.1 Active voiced Stimuli 93
C.2 Passive voiced stimuli 94

Chapter 1

Introduction

Overview

In this thesis, we sought to solve two problems: firstly, we wanted to understand sentence comprehension in the brain, specifically how the brain reconciles sentences with different syntactic structure. Secondly, we wanted to develop data analysis methods that enabled us to solve our first problem. To that end, we optimized two common analysis approaches, decoding and representational similarity analysis (RSA) and applied them to Magnetoencephalography (MEG) data. The high temporal resolution of MEG data allowed us to distinguish computations that occur in the brain during sentence reading from those that occur after the sentence has been read, henceforth referred to as the post-sentence time period.

By applying our improved decoding and RSA approaches to the MEG data, we were able to demonstrate that during sentence reading, each word is processed as it is read in an incremental fashion. Then, post-sentence, syntactic information such as sentence voice dominates the MEG data signal. When syntax becomes less decodable from the MEG data, the sentence proposition (i.e. "Who did what to whom?") is decodable. Our results from both decoding and RSA provide converging evidence of the existence of a post-sentence wrap-up period in which syntax is processed, followed by the integration of syntax and semantics to form the full sentence meaning.

The rest of the thesis is structured as follows: in this Chapter we give an overview of the scientific and methodological problems we solved. Chapter 2 is an overview of the relevant neuroscientific background on sentence processing as well as an introduction to the literature discussing the application of decoding and RSA to neural data. Chapter 3 details our methodological contributions, while Chapter 4 provides context on the limitations and relevant good practice measures needed for our methods. Chapters 5 and 6 present our results on a pilot and a confirmation data set, and Chapter 7 contains our discussion and ideas for future work.

Scientific Problem Statement

Language is an important part of the human experience, one that most humans use in their everyday lives. A loss of language ability can have a profound personal and psychological impact that has been likened to losing one's sense of self [3]. Understanding language processing has been a long-standing goal of linguistics, psychology, and neuroscience.

In order to communicate complex ideas and narratives, small linguistic units (such as words or phrases) are combined into sentences. Sentence comprehension is the first step on the linguistic pathway in which the meaning of single words are merged via structural rules into a linguistic whole. They are the smallest linguistic unit that requires a hierarchical structure for representation (as opposed to a simple list of elements), with potentially long-range structural relationships. Sentences constitute an essential building block of language, and thus make a natural entry point into the study of language comprehension.

However, many questions remain as to how humans understand even simple sentences. Sentences capture meaning in many ways; for example, "The dog found the peach," and "The peach was found by the dog," are the same proposition in different syntactic voices. In English, the former version, called active voice, is the most common. These two versions allow us to test theories of syntactic integration. For example, it is possible that sentences are read without any assumptions of syntax, and the human brain is flexible enough to understand active and passive sentences in the same manner. Alternatively, given the human tendency to make easy assumptions, reading the passive-voiced version may require a correction of the initial reading, which defaults to active voice, since it is more common. Evidence seems largely to point to the second theory of integration, and the correction has been referred to commonly as reanalysis.

The scientific goal of this thesis is to shed light on sentence processing, specifically on the comprehension of sentences with alternative syntactic structure. To that end, we contrasted active and passive sentences, dissociating syntax and semantics to better understand how the brain processes each type of information. We used both decoding and RSA to examine information flow in the brain during sentence reading and to probe the post-sentence time period for information. We found that words are processed incrementally during sentence reading, and observed that during the post-sentence time period syntactic processing is followed by sentence integration.

Methodological Problem Statement

While sentence comprehension in the brain can be studied in many ways, non-invasive neuroimaging is currently the easiest and most direct way to answer questions about the human brain. Of the techniques available, magnetoencephalography (MEG) holds the most promise for examining language. It measures the induced magnetic field produced by concerted neural activity at a rate of 1kHz [4]. This is ample resolution for our purposes, as humans typically process words on an order of hundreds of milliseconds [5]. Compared to functional magnetic resonance imaging (fMRI), which operates on the order of seconds, MEG is clearly the best choice, especially given that we want to distinguish processes that occur during sentence reading from those that occur post-sentence. MEG also has an advantage over electroencephalography (EEG) in that magnetic fields, unlike electric fields, are not distorted by the human skull. This improves the localization of MEG signals to actual brain regions (known as source-localization). Analyzing MEG data is difficult from a statistical standpoint: the signal-to-noise ratio of a single MEG trial is quite low, so a stimulus must be presented multiple times and the resulting trials are averaged together [6].

Given that we can only collect so many trials from a single participant in one scan session, we are severely data limited when analyzing MEG data. Furthermore, the time resolution greatly expands the feature space. This data deficit dictates which types of data analysis techniques (machine learning-based and otherwise) will perform well (with high sensitivity) on MEG data. Approaches that are standard for fMRI data cannot simply be transferred to MEG data applications without modifications.

The methodological goal of this thesis is to optimize two analysis approaches for MEG data: machine learning classification (referred to throughout as decoding), and representational similarity analysis (RSA) [7]. In decoding, we attempt to classify stimulus attributes from the neural activity. For example, we may want to decode whether the sentence was in the active voice or a passive voice from post-sentence neural activity. In RSA, we compare various models of sentence similarity in terms of their ability to correlate with neural activity. RSA can be thought of as asking the question: "Are stimuli that are similar according to this model close in neural activity space?" For example, we can compare a model that denotes two sentences as similar based only on their syntax to one that uses both syntactic and semantic information.

Both decoding and RSA, when applied in their standard forms, underperform on MEG data, and are insufficiently sensitive to detect the full range of effects we may be interested in [8]. In this thesis, we sought to improve the performance of decoding by combining data from multiple subjects (as opposed to the traditional single-subject approach), boosting test data signal-to-noise ratio (SNR), and using classifiers with regularization. Our RSA approach also combined data from multiple subjects, and we developed a novel method of noise ceiling computation to account for our multi-subject approach.

Thesis Statement

Magnetoencephalography (MEG) data can help us understand language comprehension in the brain, but only through careful tailoring of our analysis approaches.

By optimizing our methods, we were able to reveal the information contained in the MEG data to an unprecedented degree. Our approach revealed that sentence comprehension, previously theorized to be an entirely incremental process, involves a post-sentence wrap-up period in which syntactic and semantic information are integrated.

Chapter 2

Related Work

Overview

In order to accomplish the goals of this thesis, we need to understand the relevant work in two fields:

1. The neuroscience of sentence comprehension
2. Analysis methods for neuroimaging data

This chapter begins by defining key terminology related to sentences and sentence processing as used in this thesis. We then explore work that has been done on sentence comprehension in the brain. In addition to work on general sentence comprehension, we examine relevant conclusions in the literature pertaining to active and passive voice sentences and how the brain may respond differently to each type of sentence. In the latter part of the chapter we explore common analysis approaches for neuroimaging data, with special attention to decoding and to representational similarity analysis.

2.1 Terminology

For ease of understanding, below are the definitions of several key concepts in sentence comprehension, as they are used in this thesis.

A **proposition** is the basic information conveyed by a sentence. For a simple, noun-verb-noun sentence, it answers the question "Who did what to whom?" Multiple sentences can convey the same proposition. The proposition can be thought of as the **semantic** content of the sentence, whereas the form (the order of the words and overall structure) is the **syntax**. Properly parsing the syntax is necessary in order to determine the proposition.

The constituent words of a sentence can be described using the **thematic** roles that these words play in the proposition. The **agent** is the "who" in "Who did what to whom?" It is the noun that performs the verb. The **patient** is the object of the verb; it is the "whom." A simple proposition can be represented as a tuple: (agent, verb, patient).

In English, propositions are usually phrased in the **active voice**, which orders the words agent-verb-patient. An alternative is the **passive voice**, which orders them patient-verb-agent.

To illustrate, consider the following sentences:

1. The woman approached the man.
2. The woman was approached by the man.
3. The man approached the woman.
4. The man was approached by the woman.

In sentences 1 and 4, the agent is woman. In sentences 2 and 3, the agent is man. Sentences 1 and 3 are in the active voice, and sentences 2 and 4 are in the passive voice. Sentences 1 and 4 convey the same proposition, and sentences 2 and 3 convey the same proposition.

2.2 Sentence Comprehension

Sentence comprehension is essential to language understanding, and has been studied somewhat separately in two different fields: linguistics and psycholinguistics. Linguistic accounts for sentence comprehension tend to focus on the notion of a grammar, the set of rules by which a string of words are combined to get the full sentence meaning. Psycholinguistics and neuroscience characterize something slightly different: the actual *process* of combining the words. This distinction is important and accounts for several differences in understanding between fields, because contradictions can arise due to processing limitations in measured human subjects, as opposed to flaws in a given grammatical theory [9]. However, an ideal connection between fields would be for neuroscience to test in a rigorous manner grammatical theories put forth by linguists [10].

A central common point between pure linguistics and psycholinguistics is the concept of Merge [11]. Merge is the fundamental operation by which linguistic concepts are combined. In order to comprehend a sentence, the listener (or reader) applies Merge repeatedly, combining the constituent words into phrases and the phrases into the sentence. Merge can occur is theorized to occur at all levels of language processing. How humans know when to merge and which entities should be merged are key questions of the process of sentence comprehension.

In the neuroscientific and psycholinguistic literature, there are several prominent models of language comprehension [12, 13, 14, 15]. All have the following in common:

- Sentences are represented in the brain as *hierarchical*, tree-like structures.
- Online comprehension builds this structure by integrating each new linguistic unit into it *incrementally*.

This integration step is either referred to explicitly as Merge [14, 15] or as unification [12, 13].

The primary differences in models of sentence comprehension lie in the accounts of where in the brain the different kinds of merge (syntactic, semantic, phonetic) are performed and in what order. Many of these differences can be explained by the difficulty in recording from the healthy human brain: one must generally make a trade-off between spatial resolution (fMRI) and temporal resolution (M/EEG). Furthermore, differences in stimulus selection, task, and analysis approach can make certain effects appear larger or smaller. Lastly, while the works reviewed here have involved either auditory or visual presentation of linguistic stimuli, little work has been done to evaluate the correspondence between the neural response across modes. However,

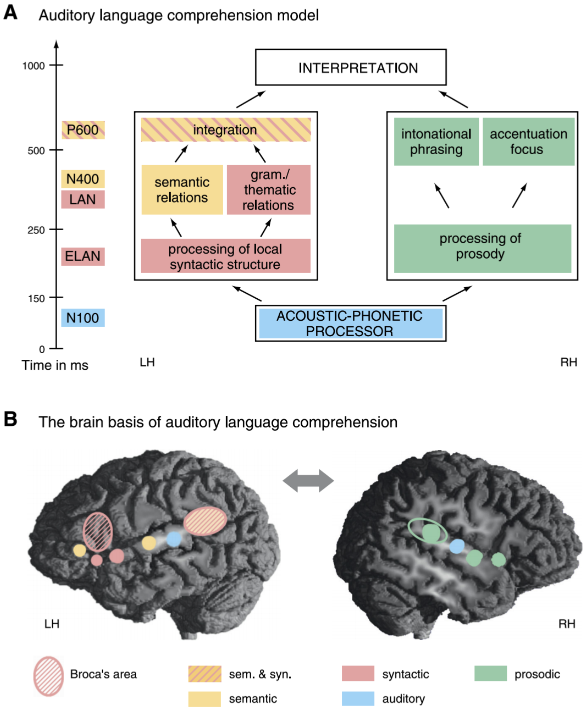


Figure 2.1: **Model of Auditory Language Comprehension** Taken from [1] **A. Auditory language comprehension model.** Illustration of the processing steps taken during the processing of acoustic language stimuli. Note the parallel paths for general linguistic processing (syntactic and semantic) and for prosody. **B. The brain basis of auditory language comprehension.** Brain regions activated in relation to the different processing steps outlined in panel A.

as will become apparent, there is a large overlap between the results gleaned from the separate types of stimuli.

2.2.1 The Neurobiology of Simple Sentence Comprehension

What happens in the brain as a person hears a word as part of a sentence? According to a large-scale review of the EEG literature [1], the first 150ms are spent recognizing the word acoustically. In the next 100ms, local syntactic structure is processed, e.g., the identification of the category of the word (verb, noun, etc.). In the next 250ms, the brain elucidates how the word semantically relates to the sentence, as well as how it relates syntactically. That is, the role the word plays in the sentence is understood. Finally, at the 600ms mark, the brain Merges the word into the rest of the sentence. See Fig. 2.1, A. for an illustration from [1]. Note that a similar perceptual-semantic gradient has been replicated for written nouns [5].

In most models of sentence processing, the left inferior frontal gyrus (lIFG), otherwise known as Broca's area, is thought to underlie syntactic Merge (see Fig. 2.1, B.) Additional important regions range from the left tempero-parietal junction anteriorly along the lateral sulcus to the superior and medial frontal gyri [1, 16, 17]. Similarly, the Memory, Unification, and Control model implicates primarily the inferior frontal gyrus, the lateral sulcus and superior temporal sulcus, with lIFG housing the unification operation [12].

These models and their neural substrates have arisen from decades of lesion studies as well as careful fMRI and EEG studies that contrast syntactically complex sentences with simpler sentences or word lists [13, 14, 18]. However, a somewhat contradictory account has arisen from studies that use more naturalistic language stimuli and more complex analysis techniques. An experiment consisting of natural story listening has implicated the left anterior temporal lobe (lATL) as essential for syntactic processing by correlating its activation with the number of open nodes in the sentence hierarchy [19], a result that has early historical precedent (see [20], and for a review, see [17]). An EEG study of sentences and phrases in Chinese potentially confirms the role of the lATL in sentence processing, although EEG lacks the spatial resolution necessary to be certain [21]. Chinese words corresponding to phrases and sentences were presented auditorally at a constant rate, with complete phrases occurring at a harmonic rate and sentences at another, slower harmonic. Different sensors revealed activity that synchronized with the word, phrase, and sentence presentation frequencies, with more anterior sensors along the left temporal lobe synchronizing to the presentation frequencies of greater syntactic complexity [21].

Additionally, the use of language localizers (contrasting linguistic and non-linguistic stimuli) has improved sensitivity in fMRI, thus allowing a larger number of distributed regions to be detected as involved with sentence comprehension and syntactic complexity [22], whereas previously it was thought that semantics was distributed while syntax was more localized (specifically, to the left IFG) [18].

These potentially contradictory explanations of the neurobiology of sentence processing can perhaps be accounted for via the general nature of the Merge/Unification operation, which can operate not only at the syntactic level, but at the phonological and the semantic levels [12, 23]. In fact, attempts to isolate merge in the context of syntactic hierarchy have localized the operation to a small sub-portion of lIFG [23], and there is evidence that phonological and semantic merge are similarly housed in distinct subregions [12]. Additionally, as sentence length and syntactic complexity increases, so too does working memory load. Attempts to distinguish working memory from core linguistic computation via fMRI have focused on whether or not additional linguistic material is hierarchically more complex, finding again that a small subportion of left IFG is selectively active for language-specific complexity [24].

Using intracranial, which has high spatial resolution while retaining the crucial temporal resolution needed to measure linguistic phenomena, work has been done explicitly measuring the building of syntactic hierarchies. In this experiment, the output of various parsers was correlated with neural activity, finally demonstrating that a bottom-up can strongly predict neural activity in lIFG. Furthermore, in a broad range of linguistic areas, gamma-band power was found to increase with each incoming word, and then decrease as soon as items could be successfully merged into a single unit (such as a phrase or clause), perhaps indicating a more global neural footprint of Merge [25].

From these many pieces of (sometimes contradictory) evidence, we can conclude the following about sentence processing:

- There is evidence of a neurobiological implementation of Merge likely localized to left IFG or Broca's Area [1, 12, 13, 16, 18, 23, 25].
- A set of left-lateralized brain regions activate in a manner correlated with incremental

sentence structure building [21, 25]

- A broad range of left-lateralized regions are implicated in sentence comprehension more generally [17, 21, 22, 25].

2.2.2 Passive Sentences and the Need for Reanalysis

The aforementioned accounts of sentence comprehension all characterize the brain as an online parser that integrates incoming words with the current tree structure as they are received. However, sentences with unconventional structure (e.g. passive voice sentences) pose a problem for such a theory. For many sentence structures, the true proposition cannot be known until all of the words have been received, and the original parse may not be correct. For example, in the sentence "The woman was approached by the man.", the reader or listener must wait until the end before it becomes clear that the man is the agent of the verb. If the brain has parsed the sentence greedily (as in a bottom-up parser [25]), the tree may incorrectly contain "woman" in the agent role. In the psycholinguistic literature, processes related to correcting the tree are broadly referred to as reanalysis. An alternative explanation to reanalysis is that all likely parses are maintained by the brain, with the brain finally committing to one once all information has been received. However, this hypothesis would predict minimal (if any) difference in the neurological signature elicited by active and passive sentences that cannot simply be explained by processing load. The work reviewed in this section will demonstrate that this is not the case, providing support for reanalysis as the solution to the voice parsing problem.

Early work using fMRI used active and passive voice sentences to try to dissociate semantic and syntactic processing in the brain [26]. Participants were assigned to one of two types of sentence similarity judgments: syntactic or semantic. In the case of the syntactic task, sentences with the same voice were regarded as the same, whereas in the semantic task, the meaning of the sentence was fully processed. Results implicated an anterior portion of Broca's area in syntactic processing, while a more inferior portion activated for semantic processing, a division that has since been replicated [12, 23].

Additional work using EEG recordings during the presentation of German sentences revealed evidence for a reanalysis effect: [27] found that sentences containing object-experiencer verbs, the German equivalent of passive sentences, elicited greater neural activity 250-500ms after final word (verb) presentation than active-voice sentences. This reveals a fundamental difference in how active and passive sentences are processed. Furthermore this difference is detectable only at the end of the sentence. However, it is impossible to say what the increase reflects, information-wise. Supposedly it corresponds to thematic reanalysis, but it can also be explained via an increase in processing load induced by distance between the agent noun and the verb.

Another study conducted in German sought to disentangle these two potential mechanisms by crossing passive and active sentences (thematic reanalysis, or argument reordering) with sentences with long- and short-range noun-verb dependencies (called argument retrieval) [28]. fMRI data revealed that reordering selectively activated Broca's Area (left IFG), as has been implicated in the general sentence processing literature, while retrieval elicits activity in the tempero-parietal (TP) junction. Using EEG, they could isolate the activation in the TP junction as occurring early (in the first 200ms), and the activity in Broca's area as occurring as late as 300-600ms post last word onset.

Studies conducted in English and Japanese using fMRI have also provided evidence consistent with the existence of a reanalysis effect, demonstrating that these results generalize to other languages. In both cases, passive-voice sentences elicit greater neural activation as measured by BOLD response in Broca's area as well as in temporal regions [29, 30]. The study conducted with Japanese stimuli went one step further and attempted to disentangle the semantic aspect of reanalysis (assigning the nouns to agent and patient) from the syntactic aspect (reordering the words), and found that while both Broca's area and the superior temporal gyrus are activated for thematic reanalysis, Broca's area shows greater sensitivity to syntax [30].

A parallel line of inquiry in sentence processing has looked at the effect of syntactic and semantic errors on neural activity. While errors elicit specific event-related potentials during the reading of the erroneous word, there is an additional change in neural activity during and after the reading of the last word in the sentence [31]. This implicates the post-sentence time period as crucial for sentence comprehension and can be likened to the reanalysis signal detected by other studies.

These studies all provide neural evidence for a reanalysis effect, housed in Broca's area and the superior temporal lobe, and coming into effect around 250-600ms post final word onset. In fact, they potentially point to two separate reanalysis effects: semantic and syntactic, both of which are required for understanding passive sentences. However, the question remains: what is the information content of the reanalysis signal? That is, what is the computation performed by the brain during reanalysis?

Multivariate pattern analysis (MVPA) can potentially answer this question [32]. For example, a study that presented video concepts of the form agent-verb-patient to subjects while recording fMRI demonstrated that the identity of the agent and the patient could be reliably decoded with a whole-brain analysis (excluding occipital lobe) [33]. There has been additional work along this line for both active and passive sentences, using a searchlight approach to classify the agent and the patient using the activity in many different regions in the brain separately [34]. Two proximal yet distinct subregions of left medial superior temporal cortex were found to selectively contain information about the agent and the patient of the sentence, respectively. The authors propose (speculatively) that the processing of sentences uses these two regions as registers to store the relevant roles of the nouns of the sentence. What follows naturally from such a theory is that if one of these registers is incorrectly filled, e.g. if a word that was supposedly the agent is found to truly be the patient, the brain must perform a transfer operation in order to correctly assign word role. This transfer operation could be a mechanistic explanation for the reanalysis effect. It should be noted that this work faces a challenge from subsequent work that reviews a wide array of regions that can semantically encode words in all grammatical roles in a sentence [35].

Taken together, this body of work is consistent with the idea that processing a passive-voice sentence involves reanalysis, in which the roles of agent and patient are swapped, in order for the proposition to be fully understood. There is some evidence that Broca's area underlies the syntactic aspect of this role-reversal, whereas the superior temporal lobe performs the semantic aspect of the operation. The work conducted in fMRI demonstrates that inferior frontal regions are involved, and that noun role assignment takes place in those regions ([30, 34], while M/EEG studies give us an indication that the process starts soon after the last word is presented [27, 28].

The presented work seeks to complete the picture by answering the following:

- What is the information content of the reanalysis signal as measured by MEG?
- Can we distinguish syntactic and semantic aspects of reanalysis using MEG?

2.3 Machine Learning for MEG Data

The use of machine learning in neuroscience has been steadily gaining popularity since the inception of multivariate pattern analysis (MVPA) [32, 36] and Representational Similarity Analysis (RSA) [7], although these techniques are still most commonly applied to fMRI data. The goal for each approach is to correlate neural data with behavioral measures (e.g. the stimulus being read) in order to elucidate the neural representation of those measures.

2.3.1 Decoding and Encoding Models

Multivariate pattern analysis (MVPA) is another name for the application of machine learning algorithms to brain imaging data (as apposed to traditional univariate analyses, such as ANOVAs). Typically this refers to a decoding approach, in which the stimulus (or some other variable) is predicted from the neural activity [32]. Machine learning can be applied in the opposite direction, predicting neural activity from stimuli, in what is referred to as an encoding approach [37]. While there are differences in how decoding and encoding results can be interpreted [38], the methodological approach remains consistent. Each case reduces to the standard machine learning problem of prediction.

When evaluating prediction performance, one typically trains on a subset of the data samples and tests the trained model on another subset of data samples. With M/EEG data, what constitutes a data sample can vary. For example, during the experiment subjects are typically shown the stimuli in trials. Multiple trials in which the same stimulus was shown can be averaged together to improve signal-to-noise ratio (SNR). For machine learning purposes, a data sample can be either a single trial or the average of several trials.

In recent years, machine learning has been applied to M/EEG data, not just fMRI (for a basic tutorial for MEG, see [8], for fMRI see [36]). The most straightforward way to apply machine learning to neural timeseries data is the sliding window approach. At each time point, train a model using some window of data, e.g. the data from time point t to time point $t + w$. This creates T different models for a timeseries of length T. We can then cross-validate over data samples to generate a classification accuracy for each timepoint. A simple extension to this approach is the temporal generalization method (TGM) [39], which takes the trained model at each time point and tests it on every other timepoint. This can help assess whether the neural representation of the class to be decoded at time t is similar to that at t': if the two representations are similar, then a model trained on t will be able to successfully decode at time t'.

Models are typically trained and tested on the data from each subject separately, with the resulting classification accuracies averaged over subjects. Because so many models are trained in this manner, testing for significance via a permutation test, the field standard [36], is extremely computationally expensive. For that reason, the Wilcoxon signed-rank test [40], which tests for consistency in results across subjects, is often used as an alternative [8]. Once significance is established, the results must be corrected for multiple comparisons over time (and potentially over regions in the brain), either via a cluster-based approach (see [41] for a primer) or standard False Discovery Rate (FDR) correction [42].

There has been some work combining data from multiple subjects for the purposes of decoding. One approach involves transforming the individual subject brains into a common space via Canonical Correlation Analysis [43]. Another competing approach, hyperalignment, computes pairwise alignments between subjects [44]. The former has been applied specifically to MEG data, whereas hyperalignment has traditionally been applied only to fMRI data. In both cases the goal is to mitigate the effect of brain morphology and timing differences across subjects for the purposes of training a classifier.

Early work decoding linguistic stimuli from MEG helped to reveal the timecourse of single-word processing by decoding semantic and visual attributes about the words [5]. This approach was later extended to Adjective-Noun phrases [45], and to tracking the neural correlate of context in stories [46]. For a review of this body of work, see [47].

The current work applies decoding to MEG data to better understand the information content of the neural signal. The prevailing assumption is that if we can reliably decode an attribute of the stimulus from the neural data, information about that stimulus is contained in the data (note that the inverse is not true [38]). To that end, in the next two chapters we will explore how to optimize a decoding approach (Chapter 3), as well as discuss potential pitfalls in interpretation (Chapter 4).

2.3.2 Representational Similarity Analysis

Another popular multivariate approach to neural data analysis is Representational Similarity Analysis (RSA). While not quite machine learning, *per se*, RSA faces many of the same challenges of decoding/encoding approaches, while also having its distinct advantages and challenges.

RSA answers the question: are stimuli that are similar according to some set of assumptions also similar in neural activity space? That is, if we think of the neural response to a stimulus as a point in a high-dimensional space, do the distances between points corresponding to different stimuli correspond to some hypothesis-driven notion of similarity between those stimuli? The goal is to capture the “representational geometry” of the neural space and see which hypotheses or models of behavior can best explain that geometry [7].

In a typical RSA pipeline, one has neural data collected in response to several different stimuli. One also has several candidate models of how the stimuli might be similar. For example, let us say that the stimuli are words. Words can be similar orthographically (they are spelled the same way), by part-of-speech (they are both nouns), or semantically (they mean similar things). For example, the words ‘dog’ and ‘boy’ are orthographically similar, and are the same part-of-speech. However, ‘boy’ is more semantically similar to ‘man’ than it is to ‘dog’. Each of these notions of similarity corresponds to a different hypothesis or model[1]. The critical computational unit of RSA is the Representational Dissimilarity Matrix (RDM), which consists of all the pairwise distances between stimuli. The neural data will produce one RDM, and a separate RDM can

[1] Note that model in the RSA sense refers to something different from model in the decoding sense. As opposed to being a model with learned parameters, a model used for RSA is usually a theory about how the stimuli are related.

be produced for each model of similarity. The neural data RDM can be computed via a variety of distance metrics: cosine, euclidean, or even classification accuracy (the idea being that neural codes that are more easily distinguished by a classifier are further apart). Then the rank correlation between each model RDM and the neural RDM is computed. The choice of rank correlation is somewhat critical so as to maximize sensitivity of the approach. To combine data across participants, a single RDM is computed per-person and the correlation between each of those neural RDMs and the model RDMs is averaged over participants. To determine significance, one can either use a label-permutation test (i.e. the Mantel test [48]), or the Wilcoxon signed-rank test over the individual participant correlations, comparing with 0 [40].

One can also ask: how well can a model reasonably perform, conditioned on the noise intrinsic to the neural RDM? This level of performance is known as the noise ceiling. Think of each neural RDM as a noisy draw from some true signal distribution. While the true noise ceiling is impossible to compute without knowing the true underlying signal, we can attempt to bound its value above and below using the data itself. The standard approach to computing the lower bound is to correlate each individual subject's data to the group average (excluding that subject). To generate the upper bound, correlate each individual's data with the total group average. The reasoning is that the lower bound underfits the data, while the upper bound overfits. For a full tutorial on RSA with examples of its application to fMRI data, see [49].

What if the models of interest are correlated with one-another because they capture related stimulus attributes? For example, if they are competing representations of sentence meaning, it is natural for some correlation to exist between models (unless some models capture nothing about the sentence meaning). Correlation across models can be detected by rank-correlating the model RDMs with each other. While no RSA tutorial makes mention of this possibility [49], there are examples of studies that use partial correlation, conditioning on alternative models, whenever this is the case [50, 51].

Like most neural data analysis techniques, RSA has mainly been used for fMRI data. However, there have been a few examples applying RSA to MEG data [8, 52]. Just as with MVPA, the simple extension of RSA to neural timeseries data such as MEG is to use a sliding window approach and compute separate neural RDMs at each timepoint, so as to trace the evolution of the representational geometry over time. Different models of similarity may correlate best with the MEG RDMs at different times.

Importantly, the correlations achieved via RSA for MEG are quite low, with the noise ceiling lower bound sometimes hitting a correlation of 0 [8, 52]. This is a strong indication that there is much room for optimizing RSA for MEG data, which we will discuss later on in this thesis.

Conclusions

In this chapter we have reviewed what is known about sentence processing in the brain. While linguistic, psycholinguistic, and neuroscientific research indicate that sentences are parsed incrementally, sentences with alternative syntactic structure (such as passive voice sentences) pose a problem for this account. Previous studies contrasting active and passive voice sentences reveal a difference in the neural activity post-sentence across the two conditions, but it is unclear what information and computations underlie that difference.

We further reviewed two prominent methods for analyzing neural activity data: decoding and RSA. In decoding, the stimulus or stimulus attributes are predicted from the neural data using machine learning classifiers. In RSA, the pairwise distances between stimuli in neural activity space are compared to the distances predicted by theoretical models. Both approaches have the potential to reveal unique insights about the brain.

Chapter 3

Methodology

Overview

Historically, neuroimaging experiments consisted of the comparison of two or more stimulus conditions. The mean neural activity from one condition in a particular brain region was subtracted from the mean activity from the other condition in that same region, and if that difference was significantly large, one concluded that that region treated the stimulus conditions differently. This approach is not sufficiently expressive to answer all types of scientific questions. In particular, it does not make use of the fact that multiple attributes of the neural signal could work in tandem to represent a given stimulus or produce a given behavior.

To improve the sensitivity of neuroimaging data analysis, the field has turned to multivariate approaches, which make use of the neural activity across multiple areas of the brain simultaneously to draw scientific conclusions. In addition to being more sensitive, multivariate approaches can be more expressive by allowing us to extract the *information content* of the neural activity, that is, what aspects of the stimuli are detectable in which parts of the brain and at which times.

Two of the most popular multivariate approaches for analyzing neural activity are decoding (otherwise known as multivariate pattern analysis (MVPA) [32]) and representational similarity analysis (RSA) [7]. The former uses classification or regression to predict stimulus (or behavior) information from the neural data. The latter captures the similarity structure across stimuli according to the neural activity, for comparison with theoretical similarity structures. In each method, neural activity for a given trial is considered a point in a high-dimensional space: decoding seeks to separate these points according to some class label, whereas RSA estimates the similarity structure of stimuli as distances in the space.

Each technique has the potential to make different contributions to our understanding of the data. When using decoding, we require representations of the stimuli to form labels for classification. While this is straightforward, it can also limit the kinds of decoding tasks we can perform. RSA, on the other hand, requires a well-defined notion of distance between the stimuli in the experiment. This can sometimes restrict the analysis, but it also has the potential to make RSA more expressive than decoding. For example, instead of having to develop a full-fledged representation of the stimulus (for example, of a sentence), we can instead develop a similarity metric between stimuli, which may be an easier task. RSA has a further advantage in that one can estimate a noise ceiling, i.e. we can determine how well the best model we have can perform given the noise in our data.

Both approaches extract information from the MEG signal and have the potential to be quite sensitive (indeed, they have been shown to be highly sensitive when applied to fMRI data [7, 32]). However, in order for these approaches to take full advantage of the high temporal resolution of MEG data, and in order for them to account for the low SNR of individual MEG trials, modifications to the standard procedures are required.

What follows are two sections, one for decoding and one for RSA. In each section we discuss the basic out-of-the-box application of the technique. Then, using a sample analysis task, we detail how changes to the basic methodological approach can improve the sensitivity of that approach. The knowledge gained here is then used to answer our scientific questions of interest in Chapters 5 and 6.

3.1 Decoding

3.1.1 Basic Approach

A typical neuroimaging experiment consists of recording neural activity while a person experiences a stimulus of some kind and/or while that person performs a behavior. In language studies, our goal is often to understand how the brain represents a linguistic stimulus. To that end, it is useful to see what linguistic information can be extracted from the neural data.

Decoding is a kind of signal detection technique. If the stimulus can be reliably decoded from the neural data, that indicates that relevant information is present in the neural data. The ideal decoding accuracy is 100%, although whether this is achievable given the noise level of the data is unclear. However, our ability to make inferences from decoding accuracy is contingent upon that accuracy being above chance performance (the performance of a classifier that was just randomly guessing), and the better it does ,the more confident we can feel that we have found a true effect in the data. Additionally, we can test whether decoding accuracy is significant with limited assumptions by using a permutation test. We permute the order of the labels with respect to the data samples and re-train and re-test our decoder many times, producing a histogram of accuracies that estimates what the decoding accuracy would look like if there were no relationship between the data and the labels.

In order to measure decoding accuracy, the classic approach is 0/1: if the classifier chooses the correct class for a given data sample, then that sample is marked as correct; otherwise, it is incorrect. While 0/1 is straightforward, it is not as sensitive as the alternative: rank accuracy. Rank accuracy requires a classifier that yields a distribution of confidence values over the potential classes (e.g. the log-likelihood that the data sample belongs in each class). We can then order the class labels by these confidence values and find the rank r of the correct class in that sorted list of C classes. We can take this rank and transform it into an accuracy, a:

$$a = 1 - \frac{r-1}{C-1} \tag{3.1}$$

Intuitively: if the correct class is ranked first, then the accuracy will be 1, as in 0/1 accuracy.

For each rank lower for the correct class, the accuracy decreases by $\frac{1}{C-1}$, the fraction of remaining classes that outrank the correct class. Regardless of how many classes there are, chance rank accuracy is 0.5, because in the event that there is no relationship between the classes and the data, the resulting ranks should be uniformly distributed. With the exception of our comparison of classification algorithms, all decoding accuracies reported in this thesis are rank accuracies.

Optimizing decoding accuracy is an important task for scientific inference. We can think of accuracy as an indicator of the sensitivity of our approach, in the statistical sense: that is, our likelihood of detecting a true effect that is present in the data is directly represented by accuracy. Ideally, we would like to be as sensitive as possible, which motivates tailoring our decoding approach to maximize accuracy.

To apply decoding to MEG data (or to other timeseries data in general), a straightforward approach is to use a sliding window over time. That is, if we think of the trial pertaining to a stimulus as a timeseries of length T, we train and test our classifier on a subwindow of size w, where $1 \leq w \leq T$. We can do this for all possible subwindows of the timeseries. This has been used to great effect to capture the temporal evolution of the information in MEG data [5, 8, 45, 46]. One can either use all the time points in a window or some function of the timepoints, such as the average, over that window.

A natural extension of the sliding window approach is the Temporal Generalization Method [39], which creates a Temporal Generalization Matrix (TGM) of decoding accuracies from all pairs of time points. That is, instead of training and testing on data from the same subwindow of the trial, accuracies are additionally computed for all possible pairs of subwindows. This TGM can help us understand whether the neural representation of the stimulus that we use for decoding at window i is the same as that at window j. If it is, then training on window i will lead to high accuracy when testing on window j, and vice-versa. This interpretation, however, is contingent on the signal to noise ratio (SNR) of the data being the same at these two times [39].

Because MEG is a highly sensitive recording technique, individual MEG trials are susceptible to many sources of noise [6]. Thus individual data trials are quite noisy and it is customary to average over trials for the same stimulus in order to improve SNR. This creates a trade-off for classification: we could have fewer, high-SNR samples, or a larger number of low-SNR samples. For example, if we have 10 trials of a given stimulus, we could use each trial separately (10 instances/sentence), average pairs of trials (5 instances/sentence), average half the trials separately (2 instances/sentence), or average all of the trials (1 instance/sentence). As the number of instances decreases, SNR increases. This is intuitive if one models the MEG data at trial i, x_i, as the sum of a true signal s related to the stimulus presented during that tiral and zero-mean Gaussian noise, $\mathcal{N}(0, \sigma)$:

$$x_i = s + \epsilon_i \tag{3.2}$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma) \tag{3.3}$$

Averaging multiple trials is akin to taking the expected value of Equation 3.2 as the trials i go to ∞:

$$\mathbb{E}_i[x_i] = \mathbb{E}_i[s + \epsilon_i] = s \tag{3.4}$$

For an empirical illustration of the profound effect of averaging on data quality, see Fig. 3.1.

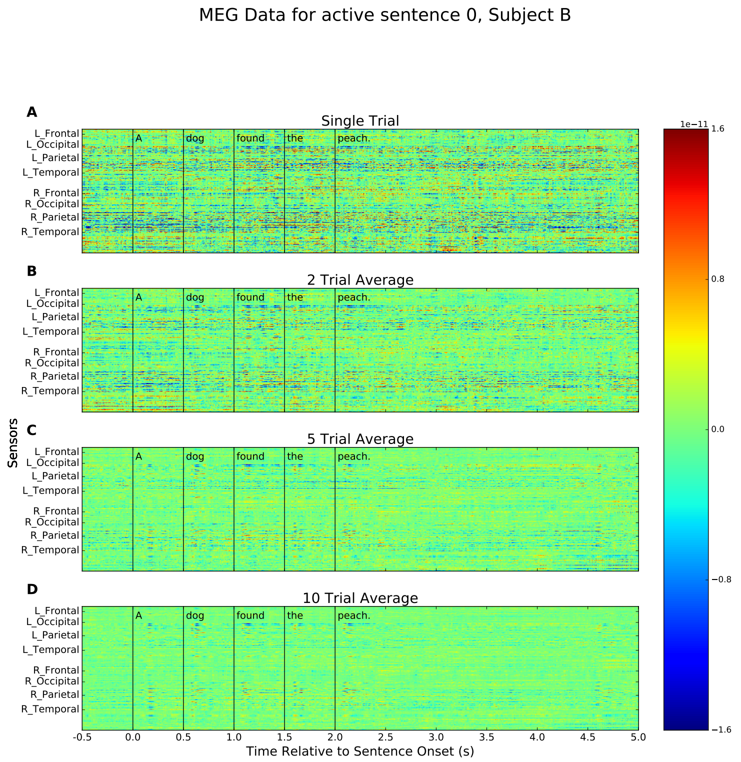


Figure 3.1: **Data as a function of trial averaging** MEG sensor data for the same stimulus as more trials are averaged together. Note that as more trials are averaged, the visual response to each word in the sentence (presented every 0.5 seconds) becomes more pronounced.

Related to the question of data samples is the cross-validation scheme used to evaluate performance. When using decoding as a signal detection technique, as we are here, one must demonstrate that the classifier can distinguish the classes of interest at a rate better than chance (i.e., better than random guessing) on a test data set that the classifier did not have access to during training. To make optimal use of the data available, cross-validation is used: the data samples are partitioned into train and test sets, and the classifier is trained and tested (respectively) on those sets. Then the data samples are partitioned again, with new samples in the test set, and the process is repeated. The reported accuracy is the average performance over all the test sets. One can perform leave-one-out cross-validation, in which each data sample is held out as the test sample once, or one could perform k-fold (where $2 \leq k \leq N$, if N is the number of samples) cross-validation, and divide the data into k partitions. While leave-one-out seems likely to yield the highest accuracy, because there is only one way to partition the data in that scheme, it tends to yield high variance results. On the other hand. 2-fold cross-validation uses less data per fold, but tends to yield more stable results [53].

We must additionally consider the question of how best to combine data across subjects. Typical decoding experiments consist of training and testing a separate classifier on each individual subject's data, then averaging the resulting decoding accuracies across subjects. If the data are well aligned in time across subjects, averaging the decoding accuracies will result in minimal loss of signal, but even a small delay between subjects can reduce accuracy. To gain some insight into the reliability of data across subjects, see Fig. 3.2. Note how even the visual response to the word presentations, which are generally quite stereotyped within a subject (see Fig. 3.1D), are misaligned.

Attempts to increase power by combining data from multiple subjects as additional training instances are generally not successful without some transformation on the feature space (e.g. PCA or Hyperalignment [44, 54]). However, an alternative is to combine data from multiple subjects as additional features, an approach that has been used successfully in fMRI [55].

The basic decoding approach has four attributes that we must consider: first, the way data from multiple subjects are combined, second, the choice of classifier, third, the treatment of time, and fourth, the treatment of trials. We will explore the effects of these attributes in the rest of this section. Unless indicated otherwise, the basic parameters used for decoding are those given in Table 3.1.

Attribute	Choice
Accuracy Measure	Rank accuracy
Subject Combination	Concatenate subject feature vectors
Classifier	ℓ_2-penalized Logistic Regression
Window Size	4ms (Averaged)
Stride between Windows	4ms
Instances	1 Instance/Sentence
Cross-validation	Leave one sentence out

Table 3.1: **Basic decoding approach**

As a running example, we will decode the sentence verb from the pool of all sentences (active and passive) from the pilot data set (full description in Chapter 5). In this experiment, 8 participants read active and passive voice sentences (e.g. “The dog found the peach.” and “The peach was found by the dog.”) while MEG data was recorded. After preprocessing, the data is sampled at 500Hz, and there are 10 single trials per sentence. We will decode sentence verb over time, aligning the sentences to the onset of the last word of the sentence, and decoding from that point until the next sentence is presented.. Throughout this section we will occasionally present full TGMs and full decoding timeseries accuracies for illustration, however, the metric of interest will largely be max classification accuracy over time for the sample task.

It should be noted that because so many classification tasks were run for the exploration in this chapter, it was computationally unfeasible to perform permutation testing and subject-level cross-validation (discussed in the following section) for all results. In Chapters 5 and 6 we provide that statistical analysis.

3.1.2 Combining Data from Multiple Subjects

In a typical neuroscientific language experiment, data is collected from multiple subjects. Decoding analysis is usually applied separately to each subject, with classifiers being trained and tested only on the data from a single subject at a time. This is due to the fact that slight differences in brain morphology across subjects can hurt classifier performance. There exists an alignment approach for classifying cross-subject using fMRI data, namely Hyperalignment; however it is quite computationally expensive [44]. Generally population-level decoding accuracy is estimated by averaging the separate decoding accuracies from each of the subjects. Inference on this population-level effect can be done by looking at consistency across subjects, potentially combined with a permutation test[1]. Averaging over subjects is straightforward, but it can result in a loss of signal if the subjects are not perfectly aligned in time.

How aligned are subjects in time? Unfortunately, time misalignment across subjects is quite common. See Fig. 3.2 for the difference between subject B and subject I for the 10-trial average of sentence 0. Note how even the visual response to the word presentations are not exactly aligned.

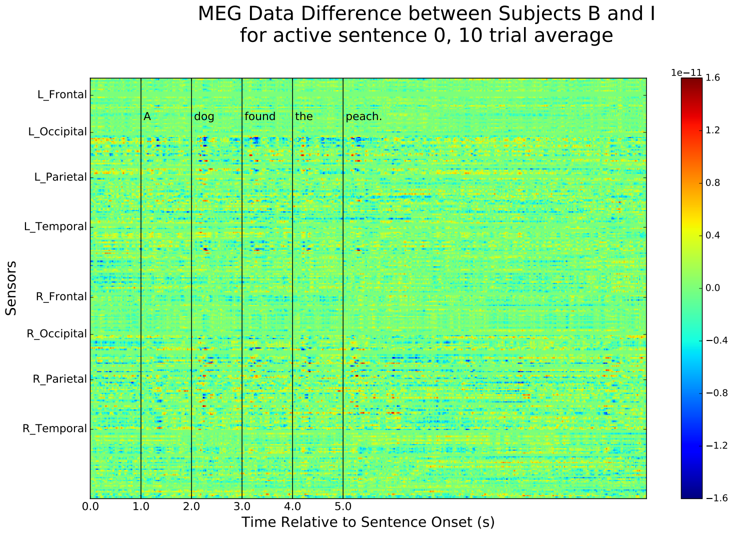


Figure 3.2: **Data difference between subjects** Data from subject B, subtracted from the data from subject I. Each subject's data is the average of 10 trials of sentence 0. Values close to 0 for a given sensor, time point entry indicate that the subjects had similar data values at that point. Large differences indicate a temporal misalignment across subjects. Vertical lines show the onset of each word and the offset of the sentence. Note that after each word, the visual response is misaligned between the two subjects.

While combining subjects as additional data samples may hurt as opposed to help classification, an alternative is to combine subjects along the feature dimension. That is, we can concatenate the f features of each of N subjects to create a feature vector of length $f \times N$. This has the benefit of boosting the amount of signal available, and for temporal decoding with MEG, is potentially more robust to small differences in timing across subjects. At any given time t only one subject needs to have the necessary signal in order for us to decode accurately. If multiple subjects contain signal, even weakly, then decoding accuracy will improve.

At first it may seem counter-intuitive; adding more features when there are few data samples could lead to overfitting. However, as is discussed in the following section, regularization of our classifier can help prevent overfitting even when the number of features is quite large.

[1] It should be noted that no consensus exists for combining permutation distributions across subjects

Empirically, we can see that the multi-subject approach strongly outperforms the single subject approach. Figure 3.3 illustrates the sample verb decoding task for the single subject (panel A) and multi-subject (panel B) approaches. While qualitatively the results are similar in that above chance accuracy is detected at roughly the same time points relative to sentence offset, the multi-subject approach has a much higher accuracy at those time points as compared with the single subject approach.

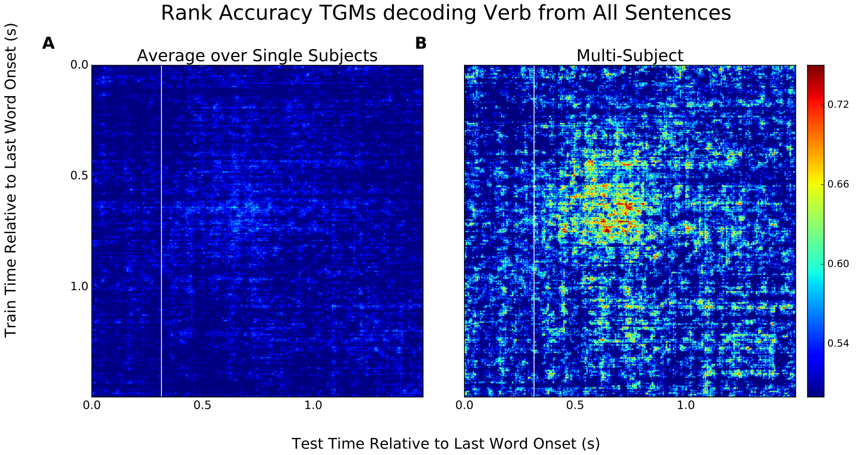


Figure 3.3: **Single subject vs Multi-subject approach.** Rank accuracy TGMs computed for decoding verb identity post-sentence from the pool of all sentences (active and passive). The white vertical line indicates sentence offset. Chance performance is 0.5 and optimal performance is 1.0. Cross-validation was leave-one-sentence out. **A. Rank accuracy TGM averaged over subjects.** A separate TGM was computed for each subject, training and testing on that subject individually. The resulting TGMs were then averaged over subjects. **B. Rank accuracy TGM from multi-subject approach.** TGM generated from using the concatenation over subjects as the feature vector for classification.

Estimating significance of the accuracy is straightforward via a permutation test. To estimate the null hypothesis that there is no decoding signal in the data, we randomly shuffle the order of the labels, breaking the relationship between them. We then retrain and retest the classifier as we normally would. This generates a distribution of null accuracies that estimates chance performance on the finite data set we have available. The p value against the null is simply the fraction of permutations that achieve a higher accuracy than the observed classification accuracy [56].

A potential drawback of the multi-subject approach is that population-level inference is no longer as straightforward as in the single-subject approach. The strength of the multi-subject approach from a decoding perspective is its weak point from an inference perspective. Because only one subject's data needs to contain the necessary signal, it is possible that the observed decoding accuracy at any given time can only be attributed to a small set of subjects, as opposed to the population of subjects

In order to estimate reliability over the population, we can cross-validate over subjects. For each subject, we leave it out, forming instead a $f \times N - 1$ feature vector, and classify as before.

We can quantify consistency in the subject population via the fraction folds that achieve above-chance accuracy. If only a few subjects are responsible for above-chance decoding at a given time point, this fraction will be less than 1, because only draws containing those subjects will be decodable. If, however, the decodability comes from the subject population together (e.g. via a weak contribution from each subject), then nearly all folds will achieve above-chance accuracy. Ideally the folds should cluster near the full N subject accuracy.

Concatenating subjects as features for multi-subject analysis substantially improves decoding accuracy. It also lends itself in a straightforward manner to significance testing and to evaluating population-level effects. If our goal is to draw conclusions from decoding results, multi-subject analysis is a clear choice over the conventional single subject approach.

3.1.3 Choice of Classifier

The field of machine learning has produced a wide variety of classification algorithms. These classifiers seek to separate the given data into the labeled classes, and do so by minimizing an objective function that estimates the error on a training data set. During training, the classifier "learns" the parameter values that best separate the data. We then evaluate classifier performance on a test set, asking the classifier to provide labels for data that it has never seen before. If there are more parameters to learn than training data samples, most classifiers will "overfit" and pick up patterns in the data that are actually noise. This is evidenced by a training error that is close to 0 but a high test error [57].

In the case of MEG data, there are 306 sensors and many time points per trial. The number of features used for classification can be quite large if we consider the neural activity at all sensors, at all time points, and concatenated across all subjects. However, scanner time is expensive and humans have limited energy for performing experimental tasks, thus the number of trials we can collect is generally quite small (in this case the experiment has a total of 320 trials). Due to the aforementioned noise in the MEG data, multiple trials are typically averaged together to make a single data point, which reduces the amount of training samples for a given classifier even further. This makes classifiers likely to overfit on this kind of data.

Classifiers are mainly distinguished by the assumptions they make about the nature of the data. For example, Gaussian Naive Bayes assumes that the features of the data are independent conditioned on the label, while Logistic Regression assumes that the boundary between labels is linear. Stronger assumptions generally indicate simpler classifiers. A simpler classifier (such as a linear classifier) is less likely to overfit. We can further prevent overfitting by regularizing our classifier - that is, instead of allowing it to fit the training data perfectly, we limit the complexity of the boundary between classes. There are several types of regularization; the two most commonly used with linear classifiers are ℓ_1 and ℓ_2. ℓ_1 results in sparse weights where some of the features are assigned a weight of 0 in the final classification decision. ℓ_2 penalizes large weights but does not result in any 0 weights. These indicate two different assumptions about the features of the data: in the first case, we assume that some features are useless and should be ignored. In the second case, we assume that while no one feature is especially useful, none is particularly useless, either [57].

Previous work has been done comparing the effects of different classifiers on MEG data; however, it was not viewed through the framework of regularization and preventing overfitting.

In this past comparison of classifiers by Grootswagers and colleagues, if principal component analysis had been applied or if the classifier were regularized, linear classifier performance was the same across several choices [8]. Here we take the analysis a step further by contrasting regularized and non-regularized classification algorithms, as well as contrasting two types of regularization. We compare performance on the sample task for the following classifiers:

- Unregularized Logistic Regression
- ℓ_1-penalized Logistic Regression
- ℓ_2-penalized Logistic Regression
- ℓ_1-penalized Support Vector Machine (SVM)
- ℓ_2-penalized SVM
- Unregularized Gaussian Naive Bayes
- Gaussian Naive Bayes with feature selection

The resulting comparison is shown in the blue bars in Fig. 3.4. Additionally, runtime was estimated for each algorithm[2] . To place the runtimes on the same scale as accuracy, the maximum runtime over all the algorithms was used to scale them. So a 1.0 indicates that that algorithm had the highest runtime, and smaller values should be interpreted as fractions of that maximum value. These runtime fractions are shown in green in Fig. 3.4.

The conclusion of this analysis is that so long as regularization or feature selection is a part of the decoding process, classifier choice is not important (this confirms previous work [8]).

3.1.4 Treatment of Time

Standard decoding of MEG data takes all the timepoints in a given window and averages over them [8, 39]. However, there are two key benefits to not averaging over time and instead using both the spatial and temporal information in a given subwindow:

1. The temporal information in a given window can help decoding accuracy.
2. Temporal averaging reduces the effective sampling rate, thereby reducing the maximum signal frequency detectable in the data.

Item 1 is illustrated in Fig. 3.5. Here note that while all trials have the same average value for the selected window, trials belonging to class A show an increase over time in that window, while trials belonging to class B show a decrease. Our classifier would be much more effective at distinguishing the classes if it could make use of that information.

Item 2 is simply a restating of Nyquist's theorem, that is, that in order to represent a frequency of f, one must sample at a rate of $2f$ [58]. One of the benefits of decoding neural activity from the time domain is that we can capture all frequencies present in the data. By averaging our temporal windows, we reduce the effective sampling rate and thereby reduce the maximum frequency our classifier can use. Given that relevant linguistic information has been shown to be present even in the high gamma frequency range (30-60Hz) of neural data [25], data fidelity is important.

[2] Algorithms were implemented with Sci-Kit Learn, with the exception of Gaussian Naive Bayes. Runtime was computed as the minimum runtime over 10 runs on the same machine

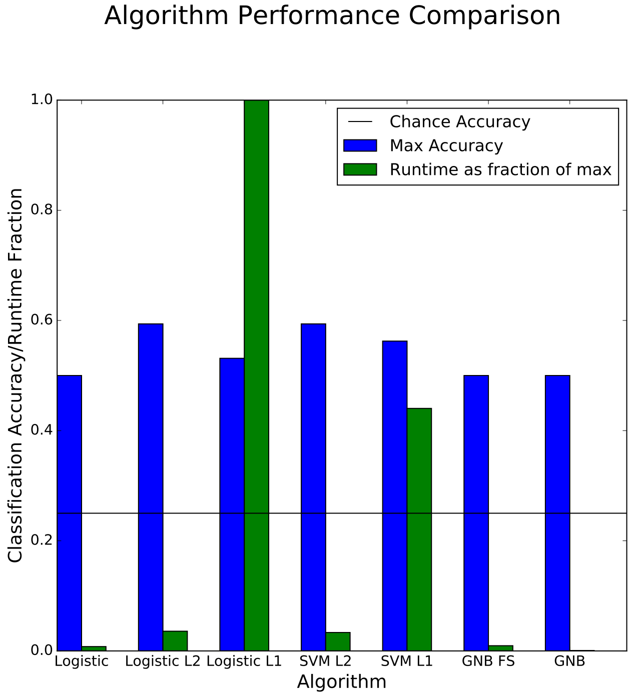


Figure 3.4: **Algorithm Comparison** Maximum 0/1 accuracy and fraction of maximum runtime for several algorithms on the verb decoding task. Red error bars on blue bars show standard deviation across subjects in mean classifier performance.

However, choosing to not average over time can create a problem in the interpretation of TGMs [39], namely that two windows may be time-shifted versions of one another, thus leading to a failure to generalize. Using averaged windows can suffer from the same problem, although it is less sensitive.

Previous work has examined the effect of using small windows of time for classification $w \leq 25$ms. For small time windows, window size and temporal averaging make little difference [8]. This is an indication that such timescales are smaller than the relevant effects. In this chapter we will examine a much wider range of time window sizes, $w \in [4, 25, 50, 100, 200]$ms. We will also compare the results with and without averaging over timepoints.

The resulting effect on accuracy is summarized in Fig. 3.6. Note that even for large window sizes, averaging over time results in similar if not better performance over using the full time window. In conclusion, the empirical evidence in favor temporal averaging is stronger than the theoretical evidence against. The optimal window size, for use in feature experiments, is 100ms, averaged over time.

3.1.5 Treatment of Trials

While linear classifiers are simple, they are still quite powerful and can separate even noisy data like MEG. Crucially, many low-SNR samples can be used to achieve similar performance as a

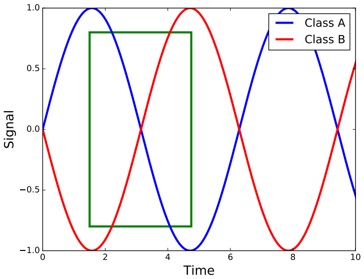


Figure 3.5: **Time Averaging Example** Toy example illustrating how averaging over time can make classification more difficult. The signal for the red class and the signal for the blue class have the same average value in the window (selected in green). However, if all the temporal information were used, they would be highly distinguishable.

small number of high-SNR samples. However, there may be an optimal point in the trade-off. For example, averaging over trials assumes that the timing of mental processes in each trial is consistent.

In a classical decoding setting, the classifier considers each test sample independently. Even if we provide many low-SNR test samples, they will still be more difficult to classify than a single high-SNR test sample. We can think of averaging over test data samples as providing additional information to the classifier: namely, that the test samples all belong to the same class. This may seem counter-intuitive at first, because the higher SNR test data is no longer drawn from the same distribution as the training data. However, at training time, the classifier is given the information that several low-SNR samples belong to the same class, information which is not available at test time since each test sample is considered separately. By averaging the test samples to create the best possible sample, we better present the signal that the classifier is looking for.

To illustrate the impact of training data SNR on classifier training, let us consider Gaussian Naive Bayes, and data for which we have M trials of N separate instances of C classes. For example, a class could be the word "dog", and each of the N instances would be the sentences that contain the word "dog." Each trial represents one presentation of a given sentence to the subject. Let the data for the j^{th} trial of the i^{th} instance and the c^{th} class be x_{ij}^c. Let the average over trials for the i^{th} instance be $\bar{x}_i^c = \frac{1}{M}\sum_j^M x_{ij}^c$. The single trial parameters for Naive Bayes are computed as follows:

$$\mu_c = \frac{1}{N}\sum_i^N \frac{1}{M}\sum_j^M x_{ij}^c = \frac{1}{NM}\sum_i^N\sum_j^M x_{ij}^c \tag{3.5}$$

$$\sigma_c^2 = \frac{1}{N}\sum_i^N \frac{1}{M}\sum_j^M (x_{ij}^c - \mu_c)^2 = \frac{1}{NM}\sum_i^N\sum_j^M (x_{ij}^c)^2 - 2x_{ij}^c\mu_c + \mu_c^2 \tag{3.6}$$

The parameters for Naive Bayes trained on the average over trials are:

$$\bar{\mu}_c = \frac{1}{N}\sum_i^N \bar{x}_i^c = \frac{1}{N}\sum_i^N \frac{1}{M}\sum_j^M x_{ij}^c = \frac{1}{NM}\sum_i^N\sum_j^M x_{ij}^c = \mu_c \tag{3.7}$$

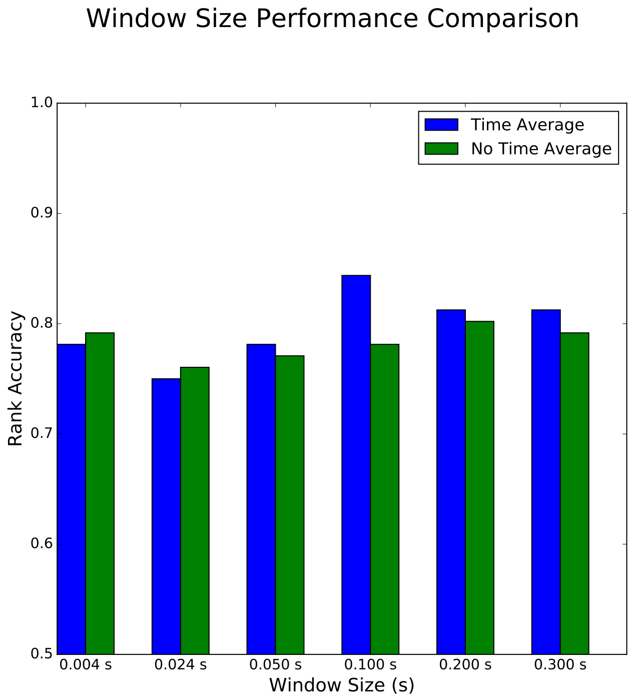


Figure 3.6: **Time Treatment Comparison.** Maximum accuracy for several window sizes on the verb decoding task. Blue bars show accuracy with temporal averaging, and green bars show without.

$$\bar{\sigma}_c^2 = \frac{1}{N}\sum_i^N(\bar{x}_i^c - \mu_c)^2 = \frac{1}{N}\sum_i^N(\bar{x}_i^c)^2 - \frac{2}{NM}\sum_i^N\sum_j^M x_{ij}^c\mu_c + \mu_c^2 \tag{3.8}$$

Note that Equations 3.5 and 3.7 are equivalent. The second two terms of Equations 3.6 and 3.8 are also equivalent. For comparing σ_c^2 and $\bar{\sigma}_c^2$ this leaves the the first term:

$$\frac{1}{N}\sum_i^N(\bar{x}_i^c)^2 = \frac{1}{N}\sum_i^N(\sum_j^M x_{ij}^c)^2 \leq \frac{1}{NM}\sum_i^N\sum_j^M(x_{ij}^c)^2 \tag{3.9}$$

By Jensen's Inequality we have demonstrated that $\bar{\sigma}_c^2 \leq \sigma_c^2$, since all the terms in the sum over i are positive [59].

Thus averaging over trials does not affect one of the parameters, μ_c, and the per-class variance σ_c for averaged trials is bounded above by the per-class variance for single trials. The potential theoretical impact of averaging over trials is small.

We can demonstrate the effect of trial averaging empirically on the sample task. Consider the following approaches to handling the training data:

- Average over all trials, creating 1 instance/sentence
- Average over half the trials, creating 2 instances/sentence

- Average over 2 trials, creating 5 instances/sentence
- Use single trials, 10 instances/sentence

We have additionally another choice to make: either the test set can receive the same averaging treatment as the training set, or it be averaged to create 1 instance/sentence.

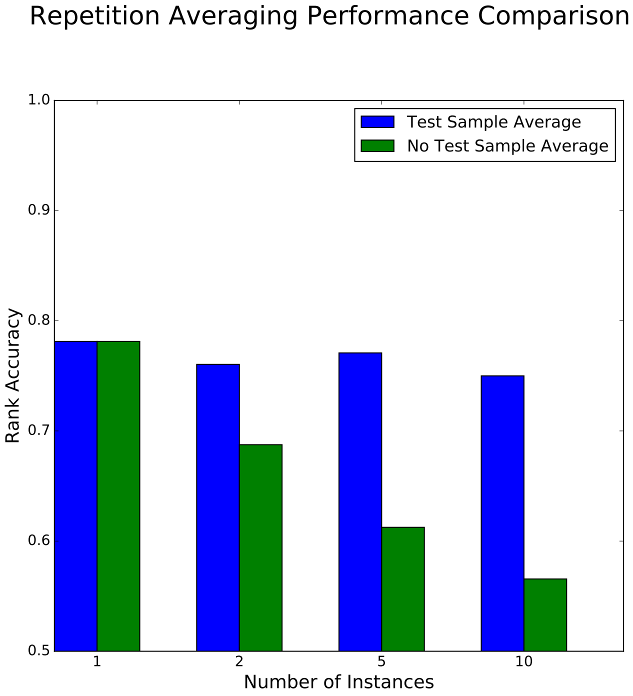


Figure 3.7: **Trial Treatment Comparison.** Maximum accuracy for several ways of combining trials on the verb decoding task. Blue bars show accuracy with test set averaging, and green bars show without, i.e. when the test set has received the same treatment as the training set.

Performance on our verb decoding task is shown in Fig. 3.7. Note that without test set averaging, accuracy drops off with SNR. But with test set averaging, all trial treatments perform roughly the same. This confirms the claim that the test set SNR is most critical for decoding performance.

3.1.6 Cross-validation Folds

Thus far we have mainly discussed the maximum accuracy over time, but also of interest is the minimum accuracy over time. It is the case that when using the leave-one-out scheme, as we do throughout this section, one can observe below chance accuracy. This presents somewhat of a puzzle - how is it possible for a classifier to do worse than guessing randomly?

The answer lies in the fact that we are data limited and in our cross-validation scheme. In order to assess true classifier performance, and particularly in order for a bad classifier to have performance that is perfectly at chance, one requires infinite data [60]. Any estimate we make of performance by cross-validating over a finite data set should be treated as a value bounded by confidence intervals. In the case of leave-one-out cross-validation, the confidence intervals tend to be much larger than for say, 2-fold cross validation [53].

To demonstrate this phenomenon, we looked at two additional decoding tasks: decoding the voice (active or passive) of the sentence, and decoding the proposition (the tuple of (agent, verb, patient)). For each decoding task, we computed results using the leave-one-out scheme, as well as the result from 2-fold cross-validation, using 100 different partitions of the data. For verb and voice decoding, since the number of classes was sufficiently high, we were also able to compute results from 4- and 8-fold cross-validation (also with 100 different data partitions).

The comparison of cross-validation schemes is shown in Fig. 3.8. Note how as the number of folds decreases, the accuracy over time becomes much smoother. This is due to the fact that there are many different partitions of the data possible under those fold schemes, so these estimates are more stable. Note also how, particularly in Fig. 3.8C the minimum accuracy over time increases as cross-validation fold number decreases.

What may seem puzzling at first is that maximum classification accuracy *also* increases. Why should this be the case? Despite the fact that 2-fold cross-validation uses less training data than leave-one-out, we are so data limited in either case that it is unlikely to cause much harm. In the case of the classification tasks explored here, the number of samples is either 32 or 62, which is the same order of magnitude, especially when compared with the number of features, which is 2448 at each time point.

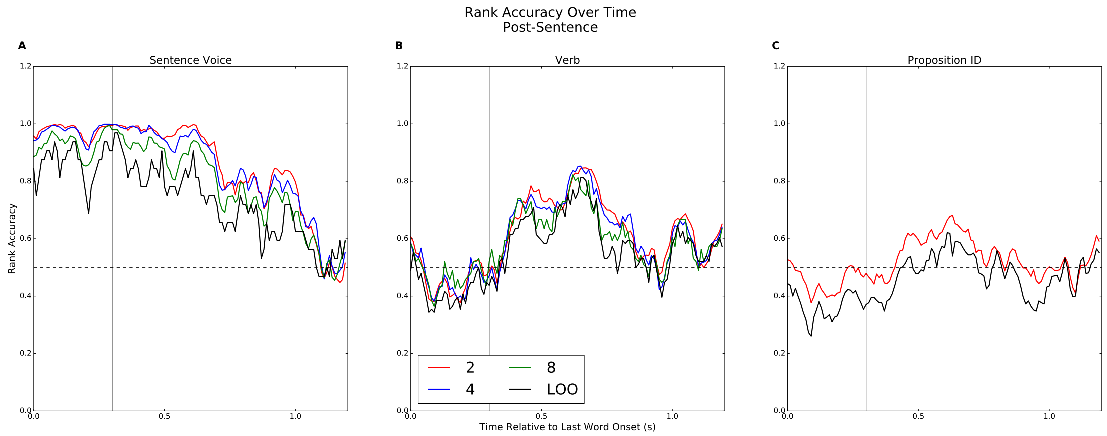


Figure 3.8: **Cross-validation Comparison.** Accuracy over time for three classification tasks: Voice (A), Verb (B), and Proposition (C). In each case performance was assessed with several cross-validation techniques, ranging from leave-one-out (LOO) to 2-fold cross-validation, with intermediate schemes where possible. The dashed horizontal line shows chance performance, and the vertical black line indicates sentence offset.

3.1.7 Optimal Hyperparameter Selection

In this section, we explored several key design choices for decoding from MEG data. From this analysis we can conclude that certain design choices are more important than others. Combining data across subjects by concatenating their feature vectors is much more powerful than building single-subject classifiers. Classification algorithm, which seems at first to be a crucial choice, is not as important as whether that classifier uses regularization. Using larger time windows improves decoding accuracy, but averaging over time as opposed to using all time points does not make a significant difference. Test sample SNR is crucial for performance, but training set SNR is less important.

Taken together, we can create an optimal decoding paradigm, summarized in Table 3.2. These parameters will be used in Chapters 5 and 6 to try and address scientific questions about MEG data. We choose to use a 10ms stride to save computation time, and we use 2 instances/sentence to facilitate nested cross-validation for choosing the ℓ_2 penalty weight.

Attribute	Choice
Accuracy Measure	Rank accuracy
Subject Combination	Concatenate subject feature vectors
Classifier	ℓ_2-penalized Logistic Regression
Window Size	100ms (Averaged)
Stride between Windows	10ms
Instances	2 Instances/Sentence

Table 3.2: **Optimized decoding approach**

3.2 Representational Similarity Analysis (RSA)

3.2.1 Basic Approach

Like decoding, RSA also seeks to reveal the information content of the neural activity; however, it does so by looking at the similarity between the stimuli. Typically in RSA we treat the neural activity for each stimulus as a point in a high-dimensional space, and we compute a Representational Dissimilarity Matrix (RDM), the pairwise distances between each stimulus. This RDM captures which stimuli are similar to one another according to the brain. We also typically have several models[3] that theorize what the similarity structure would look like under different hypotheses, also represented by RDMs. The rank correlation is computed between each model RDM and the brain RDM, which measures whether stimuli that are close in model space are also close in neural activity space. Separate RDMs are computed for each subject's data, and all correlations are averaged over subjects [7].

While the ideal rank correlation between the "true" model and the data should be 1, noise inherent in the data makes this impossible[4]. While it is not possible to compute the correlation of the true model with the neural activity without knowing this true model, we can bound that value above and below, creating what is referred to as the noise ceiling. In computing a noise ceiling the goal is to understand how well two different sets drawn from our data distribution could possibly correlate with one another.

[3] Here model is used to mean a theoretical model of how the brain might work, as opposed to a learned machine learning model. One could use a machine learning model as a candidate model for RSA, but it is not necessary.

[4] Note that this is also true of decoding.

In the traditional RSA approach, the lower bound is computed by taking each subject's neural activity RDM and correlating it with the average of *all other* subject's RDMs (not including that subject). The goal is to see how well the individual subject's data correlates with the data of all other subjects. The upper bound can be computed by taking the individual subject RDMs and correlating them with the *total* average RDM. The reasoning is that the lower bound underfits, because a single subject is a poor estimate for the data distribution. The upper bound, on the other hand, overfits, since the same data is used for both RDMs in the correlation. If the average correlation between a given model and the neural activity RDMs is within these two bounds, one can conclude that the model predicts the neural activity at least as well as the neural activity predicts itself. If two models are both within the estimated noise ceiling bound, they cannot be distinguished by the data due to excessive noise [7].

RSA has typically been performed on fMRI data, although some instances of its use on MEG data exist [8, 52], with somewhat lackluster results. As in decoding, extending RSA to timeseries data can be done by sliding a window over the timeseries, creating a different neural activity RDM at each timepoint. Crucially, the model RDM does not need to change over time. We can either use whole-brain sensor-level data to compute the neural activity RDMs or we can use source-localized ROIs, which is more similar to the typical RSA approach for fMRI data.

In this section we will use post-sentence neural activity from the confirmation data set, described in full in Chapter 6. In brief, 20 participants read sentences from four categories, long active sentences (e.g. "The man approached the woman."), short active sentences ("The man approached."), long passive sentences ("The woman was approached by the man."), and short passive sentences ("The woman was approached."). One of our candidate models will be a sentence-length model, in which two sentences are considered the same (distance of 0) if they are both short or are both long, and are considered different (distance of 1) if they differ in length. The other candidate model is a more comprehensive "syntax" model, in which two sentences are considered the same if they have the same voice and are the same length. They have a distance of 0.5 if they are the same voice but different lengths, and a distance of 1 if they are different voices. The model RDMs are given in Fig. 3.9.

Traditional RSA does not perform well on MEG data. In this section we will propose a novel approach to RSA, and demonstrate the superior sensitivity of that approach. We will also examine the effect of using multiple timepoints on RSA (as we did in the previous section with decoding). Lastly, we will discuss an extension to RSA for when two models are correlated with one another (as is the case for our sample models here).

Unless indicated otherwise, the basic parameters used for each RSA analysis are given in Table 3.3.

Attribute	Choice
Noise Ceiling	Computed by splitting data over trials.
Window Size	4ms (Not Averaged)
Stride between Windows	4ms
Correlation	Standard Kendall τ

Table 3.3: **Basic RSA approach**

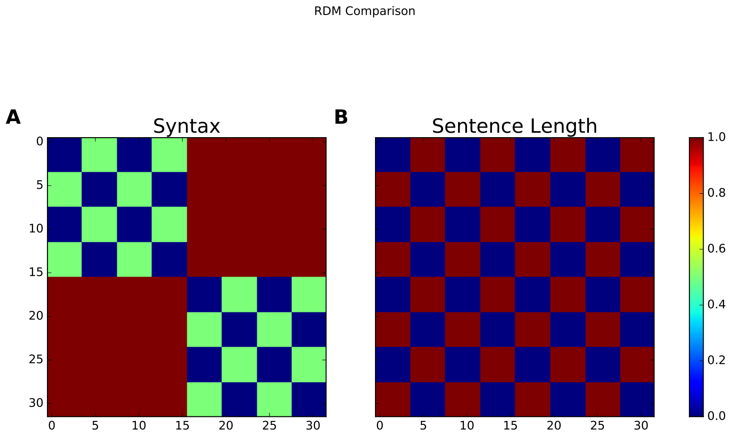


Figure 3.9: **Sample Models.** Sample model RDMs for analyzing post-sentence data. Each entry in the matrix corresponds to a pair of sentence stimuli. **A. Syntax model.** Model that assigns sentence pair a distance of 0 if both sentences are the same length and voice, 0.5 if they are the same voice but different lengths, and 1 if they are different voices. **B. Sentence length model.** Model that assigns sentence pair a distance of 0 if both sentences are the same length, otherwise a distance of 1.

3.2.2 Noise Ceiling Computation

As has been discussed earlier in this chapter, MEG data can be quite noisy. For that reason, using the RDM created from an individual subject's neural activity can often lead to lackluster results. See Fig. 3.10, panel A for the result yielded by traditional RSA on our sample task. Note that all correlations (including the noise ceiling) are very close to 0, and that the models are completely indistinguishable. While it is possible to potentially draw inferences from such a result, the effect size is so small as to be worrisome-the results seem lost in the noise.

Our proposed novel approach to RSA uses trials as an indicator of neural data reliability. Instead of correlating the model RDMs to individual subject RDMs and averaging the results, we instead correlate the model RDMs to the average RDM over all subjects, computed using only a subset of the trials from each subjects. This greatly boosts the SNR available.

To compute a noise ceiling in this approach, we again need an upper and lower bound. Recall that each stimulus is presented for multiple trials - these repetitions give us an intuitive way to measure data repeatability. One can ask: how correlated are these repetitions to each other? To get the lower bound for our noise ceiling, we compute the all-subject average RDM using half the trials for each stimulus, and compute the RDM again with the other half. This is repeated for all possible splittings of the trials. To bound from above, relate the half-trial RDMS to the full-trial RDM. This is overfit, as there is a replication of data, while the lower bound is underfit (the SNR of these data RDMs is lower than that used to test the models).

Our new approach yields the results shown in Fig. 3.10, panel B. All reported correlations have increased over the traditional approach, including the noise ceiling height. Note, however, that the two models are still indistinguishable given the noise level in the data. While there is a difference in how much each model correlates with the data, that difference could be attributable

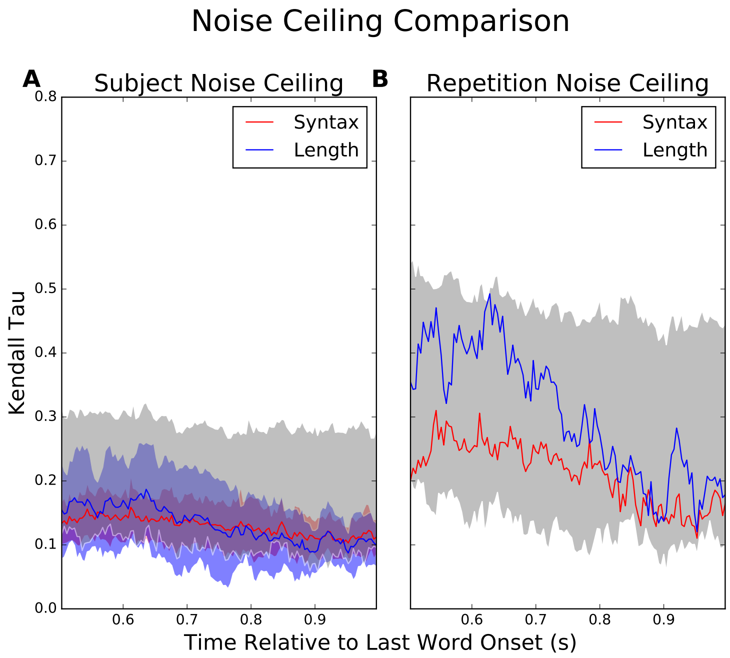


Figure 3.10: **RSA Approach Comparison.** Rank correlations between length and syntax models and neural activity over time. Gray shaded region gives the noise ceilling. **A. Traditional RSA.** Models are correlated to individual subject data RDMs and reported correlation is the average over those scores. Noise ceiling is computed in the tradition single-subject manner. **B. Repetition-based RSA** Proposed novel approach to RSA. Models are correlated with the mean data RDM over subjects. Noise ceiling lower bound is computed by correlating an RDM constructed from half the trials with an RDM constructed from the other half, for all possible splittings. Upper bound is computed by correlating the half-trial RDMs with the full mean RDM.

to noise in the data as opposed to a true difference in explanatory power between models.

The noise ceiling is a powerful aspect of RSA that is not shared by decoding approaches. Not only does it give us an estimate of how well any given model could possibly hope to correlate with the neural activity, but it also gives us a principled method for measure neural activity SNR. Without a noise ceiling, RSA would have the same problem that decoding has: it is clear whether a model fails to correlate with neural activity because of an issue with the model, or because of an issue with the MEG data itself (see Chapter 4 for a further explanation with this).

If we have multiple potential ways of creating a MEG RDM, say with different data processing choices or different distance metrics, we need a principled way to choose between them. Since using correlation with models creates circularity in that analysis, the clear answer is to look at how design choices affect our noise ceiling. The noise ceiling tells us whether the data RDMs become more reliable if we make a particular change.

3.2.3 Treatment of Time

In the case of decoding, it was clear that using multiple timepoints was superior to using a single timepoint, in terms of sensitivity. Is this also true for RSA? In this section we examine window sizes ranging from 4ms to 200ms, as in the decoding section. For each window size, we compute the noise ceiling as described in the previous section. We also compare averaging over time to not averaging over time.

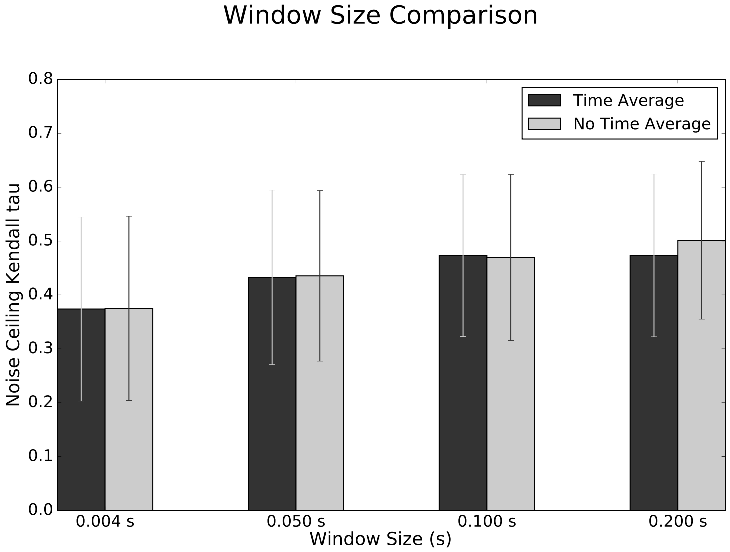


Figure 3.11: **Time Treatment Comparison.** Mean of noise ceiling lower and upper bound over post-sentence time period, for each window size, both averaging and not averaging over the timepoints in that window.

The results are shown in Fig. 3.11. As we saw with decoding, more timepoints generally improves performance. However, as in decoding, averaging over time seems to make little to no difference.

3.2.4 Partial Correlations

Recall Fig. 3.10, and note that neither model is distinguishable in terms of its performance, because both are as correlated with the MEG data as the MEG data is with itself (i.e. they are both within the estimated noise ceiling bounds). Furthermore, as is intuitive from the construction of these models, the model RDMs (shown in Fig. 3.9) are actually also correlated with one another (Kendall τ of 0.31). This creates a problem for interpretation, because any of the following scenarios could be true:

- Each model is correlated directly with the neural activity, and the correlation between models is incidental. That is, the models are conditionally independent given the neural activity.
- The syntax model is correlated directly with the neural activity and with the length model, and the correlation between the length model and the neural activity is incidental.

- The length model is correlated directly with the neural activity and with the syntax model, and the correlation between the syntax model and the neural activity is incidental.

If we want to be able to disentangle these scenarios, we must look not at the zero-order correlation across RDMs, but rather at the partial correlation. If the correlation between a given model and the brain data *conditioned on the other model* is still high, then we can feel confident that it has an independent relationship with the neural activity that cannot be explained by the other model. If, alternatively, all the explanatory power rests with one model, correlations conditioned on that model will be close to 0.

While the models in this section form a somewhat obvious example of this problem, it is by no means a rare issue. As discussed in more detail in Chapter 4, in language, many features of interest about a given stimulus are correlated (for example, word frequency is inversely correlated with word length). This makes it difficult if not impossible to create a stimulus set that resembles natural reading and spans an interesting linguistic space of stimuli. If we are comparing linguistic models, particularly a model that is an incremental change over another (say, successive layers in a neural network), there will also be correlations between model RDMs, independent of the stimulus choice. Thus, if we want to use RSA to study language, we must use partial correlations instead of zero-order correlations whenever necessary.

Zero-order Spearman's (rank) correlation ρ_s between variables X and Y is defined as the Pearson's correlation computed on the ranks of those variables, rg_X and rg_Y [61]:

$$\rho_s(X, Y) = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}} \tag{3.10}$$

where σ denotes the standard deviation. In order to condition this rank correlation on a third variable, Z, we first regress from Z to X and Y to produce $\hat{X} = B_{ZX}Z$ and $\hat{Y} = B_{ZY}Z$. Then we correlate the residuals from that regression to each other:

$$\rho_s(X, Y|Z) = \rho_s(\hat{X} - X, \hat{Y} - Y) \tag{3.11}$$

This easily generalizes to several conditioning variables via repeated regression and residual computation [62].

Figure 3.12 demonstrates the difference that partial correlations can make. The standard zero-order correlations are shown on the left. As we already know, both models perform very well and are in the noise ceiling. On the right, each model's partial correlation is plotted, conditioned on the other model. That is, the correlation for the syntax model is the partial correlation conditioned on the length model, and the correlation for the length model is the partial correlation conditioned on the syntax model.

As is potentially obvious from the construction of these RDMs, the syntax RDM contains all of the explanatory power of the two models. When the length model is conditioned on the syntax model, its correlation with the neural activity drops almost to 0.

Conclusions

Multivariate analysis approaches have great potential to provide new insights into MEG data. They can capture information content, moving beyond the simple contrasts of the past. However,

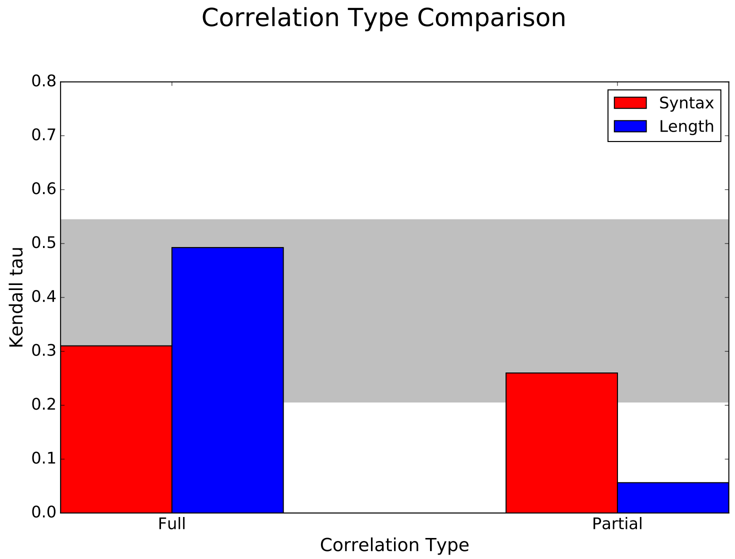


Figure 3.12: **Correlation Type Comparison.** Comparison of max model correlations over time with neural activity over the post-sentence time period. Noise ceiling is given in gray. On the left are the zero-order correlations between each model and the neural activity. On the right we have the conditional correlations, where each model is conditioned *on the other model*.

most of these approaches have been developed with fMRI data in mind. Some modifications are necessary so as to better tailor them to noisy, timeseries data such as MEG data. In this chapter, we examined two such methods, decoding and RSA, which each can provide unique insight. Both methods, when applied without care, are insufficiently sensitive. What follows is a set of guidelines distilled from the results in this chapter on how to most effectively employ each method.

Decoding provides a straightforward and sensitive way of determining what stimulus information is present in the neural signal. It draws its power from a wide array of classification algorithms. However, MEG data has few data samples and many features, restricting us to the domain of simple, linear classifiers. The most important attribute of a successful algorithm for decoding from MEG data is that it include regularization and/or feature selection. Furthermore, decoding can make use even of noisy single-trial MEG data during classifier training. However, the performance of that trained classifier hinges on the SNR of the test data. It is best to average all trials in which the test stimulus was presented to create the highest quality sample to test the decoder on. Combining data across subjects by concatenating their feature vectors provides a huge boost in performance over averaging single-subject results.

RSA allows us to compare theoretical models in terms of their ability to relate to neural activity. As with decoding, data SNR is crucial for the efficacy of this technique. Traditional RSA, which requires individual subject data to be highly reliable, is not effective when applied to MEG data. We propose the use of RDMs averaged over all possible subjects to boost RSA sensitivity. A strong advantage of RSA is the use of a noise ceiling to compute how well a model can possibly do given the noise inherent in the data. When using our proposed approach, the noise ceiling can be computed by comparing half the trials to either the other half of the trials (for the lower bound) or all of the trials (for the upper bound). This noise ceiling indicates much higher data reliability than the traditional single-subject noise ceiling. Furthermore, the RSA noise ceiling can be used to optimize the creation of the MEG data RDM in a principled and straightforward manner. We additionally recommend that when using RSA to compare models that are correlated to one another, partial correlations are used so as to distinguish the unique explanatory power of each model.

MEG data is spatio-temporal, and its high temporal resolution is a large part of its appeal as a neuroimaging technique. Both RSA and Decoding benefit from making use of the spatiotemporal nature of the data, namely by using multiple timepoints at once for analysis. In the case of decoding, it is crucial not to average over timepoints, but instead use them as they are. In RSA, this does not appear to make very much difference in performance.

Overall, both decoding and RSA have great potential in terms of MEG data analysis, but they must be tailored to suit the particularities of the MEG data in order to be effective. In Chapters Chapters 5 and 6 we will use both techniques to learn more about sentence processing in the brain.

Chapter 4

Good Practices

Overview

The goal of neuroscience is to answer scientific questions about the brain. Sometimes these scientific questions are broad and exploratory, such as "What happens in the brain of a person as they read a sentence?" Sometimes these questions are specific and pertain to a particular hypothesis, e.g., "Do humans post-process sentences after reading?" Neuroscientists record neural activity while subjects perform particular tasks and use the resulting data to answer the scientific question.

The choice of data analysis approach is crucial, because it determines what questions can and cannot be answered. Previously, neuroscientists collected data from two conditions and compared the data to see which parts of the brain behave differently under each condition, usually using statistical parametric mapping (SPM) [63]. While statistically straightforward, this cannot satisfactorily answer either of our two aforementioned example questions. Ideally we want to understand the information content of neural activity in a broad sense, not simply the difference between two conditions. A further ideal would be to record from the brain in a naturalistic setting, as opposed to using unnatural, carefully chosen stimuli.

To that end, we can employ machine learning. Machine learning can tell us what information a signal contains by testing the ability of that signal to distinguish between categories of interest. This requires us to transform our scientific questions into prediction tasks, as shown in Table 4.1.

While a more complex approach gives us more flexibility in the types of questions we can

Scientific Question	Prediction Task
What happens in the brain of a person as they read a sentence?	What information about the sentence can be decoded from the neural activity?
Do humans post-process sentences after reading?	Can sentence-relevant information be decoded from the post-sentence neural activity?

Table 4.1: **Sample Scientific Questions and Corresponding Prediction Tasks**

answer, it also requires more care. We should take note of the assumptions made by a given approach (for example, the assumption that a classifier has no access to information about the test data). We should also be concerned about the kinds of inferences we can make from our results. This relies both on the analysis approach and the way the data was collected. In this chapter, we detail several major concerns. While most of these concerns have been expressed by others, in the interest of painting a complete picture, we present them all here, and explain how these concerns informed the experimental and analysis choices made throughout the rest of the thesis.

4.1 Decoding

The most obvious concern with decoding is double dipping, in which information about the test set, on which we are validating performance, is leaked to the classifier during training [64]. This compromises all results by making them optimistic, i.e., the classification accuracy is higher than it would be on a true out-of-sample test [65]. However, this is not the only concern one should be aware of when employing decoding to answer a scientific question. The inferences one can make from decoding/classification results are not always straightforward.

4.1.1 Failure to decode indicates nothing

Given a data set contrasting the neural responses to two conditions, A and B, one can train a classifier to distinguish A and B from the data. Let us say that it achieves only 50% accuracy on an out-of-sample test, which is chance performance, equivalent to randomly guessing the class assignment. What are potential explanations for this?

1. The data contains no information distinguishing A and B because the brain does not distinguish these conditions.
2. The data contains no information distinguishing A and B because our neural recording approach cannot detect it.
3. The data contains information distinguishing A and B, but the difference between the two is within the noise margin of the data.
4. The data contains information distinguishing A and B, but the chosen classifier cannot detect it, e.g. because it is a nonlinear relationship and the chosen classifier is linear.

Scenario 1 is extremely unlikely, unless A and B are in fact the same in the brain. Scenario 2 is more likely, since each neural recording modality is limited in what it can detect. For example, MEG can only detect magnetic fields orthogonal to the skull, but the cortical folds of the brain are arranged such that neural activity propagates along a variety of directions. Scenario 3 is unfortunately also somewhat likely, depending on which regions in the brain underlie the difference in A and B. Scenario 4 is also likely, although our choice of classifier is often restricted by the number of data samples that we have. Because it is not possible to distinguish between these scenarios, one cannot make scientific claims from a failure to decode. Therefore, throughout the thesis we only make scientific inferences from positive decoding results.

4.1.2 Causal inference of classifier weights is not possible

It is tempting to believe that if a particular feature x (e.g. the neural activity at a particular sensor at a particular time)s is important for the success of a decoder (i.e. it is given a large weight), that indicates that x plays an important computational role in the brain for distinguishing the conditions of interest. Unfortunately this is an erroneous conclusion to draw.

It should be noted that the causal relationship between stimuli, neural activity, and behavior has strong implications for what can and cannot be inferred from decoding (and encoding) models. Stimuli "cause" neural activity and neural activity "causes" behavior. For example, in our experimental setup the subject reads a stimulus provided by the experimenter. Periodically the subject gives an answer to a question about the previous sentence. The causality relationship between the neural activity and the stimulus/behavior differs in these two settings. In the setting of passive reading, the stimulus is causing the neural activity. In the case of the question answering, the neural activity causes the behavior. Decoding the stimulus from the neural activity is actually anti-causal [38].

It is generally true that features useful for decoding are not necessarily the features that contain class-distinguishing information. Take as an example a two-dimensional regression task, with features x_1 and x_2 to predict an output variable y. Consider the following relationships:

$$\begin{aligned} x_1 &= y + \epsilon \\ x_2 &= \epsilon \end{aligned} \tag{4.1}$$

Here $\epsilon \sim \mathcal{N}(0, 1)$ is some noise that happens to be present in both features. Any reasonable regression approach will assign weight $w_1 = +1$ to x_1 and weight $w_2 = -1$ to x_2. If feature "importance" is indicated by $|w|$, both x_1 and x_2 will be considered signal-carrying. However, x_2 is obviously just playing a denoising role. Note that the signs could easily be reversed and the same would still be true. It is impossible to determine which is the signal-carrying feature just from decoding weights.

A typical neuroscientific goal is brain mapping, in which the neural regions recruited for the computation of interest are identified. Using decoding to form such a map is difficult, given the issues with interpreting classifier weights. However, recent methods have been developed to transform classifier weights into interpretable importance maps [2]. When examining our sensor-level decoding results, we will use this transform before interpreting our weight maps.

4.1.3 Correlation between labels affects inferences

In science we would ideally like to draw causal conclusions. However, due to ethical concerns, neuroscientists mainly use passive recording techniques such as MEG to study healthy subjects. These approaches can only yield correlational inferences.

Let us say that we can decode real-valued feature f (e.g. the length of a word stimulus) from neural activity. Does that necessarily mean that the brain uses feature f as part of its computational representation of the stimulus? Unfortunately, it does not. Let us suppose that the true neural feature is g (in our example, g could be the number of text-colored pixels on the screen). So long as the correlation between f and g, $\rho(f, g) > 0$, we will be able to decode f reliably.

This is the single most difficult problem in interpreting decoding results, although we can attempt to mitigate this issue through careful stimulus selection and choice of decoding task. For example, in the case of active and passive sentences, a clear confound with sentence voice is sentence length. In English, passive-voice sentences are longer than active voice sentences. So when we attempt to decode sentence voice, we may actually be decoding sentence length. However, by intermixing shortened active and passive sentences with longer ones in the stimulus set, we can decorrelate these two variables.

Correlation between models of stimulus processing can create issues for RSA as well. For example, let us say we are trying to correlate successive layers of a neural network with neural activity. These layers will likely be highly correlated with one another, and so too will the respective RDMs. In RSA, we can use partial correlation to counteract this problem.

4.2 Language

Studying language in the brain is particularly challenging. Because language is a uniquely human phenomenon, neuroscientists have been unable to use animal models to study it. Much of the causal inference in language neuroscience comes from lesion studies, which give Broca's and Wernicke's areas their prominence in nearly all language models. For most neuroscientists, language in healthy participants can only be studied passively.

Language processing is a naturally occurring behavior that is hard to replicate in an experimental setting. Natural reading involves eye movements that create noise, and speech production involves jaw movements that create noise. All methods currently available for recording neural activity will be adversely affected by these noise sources.

Furthermore, one has to ensure that the subjects are engaged: passively reading single words with no purpose is too boring, especially if the same words are presented over and over again (as they must be to boost signal-to-noise ratio (SNR)). However, the task that a subject is performing can directly impact how they think about the stimulus.

Lastly, as discussed in the previous section, correlation between stimulus attributes of interest can be problematic for scientific inference. Unfortunately, many attributes of language are correlated with one another, e.g. frequent helper words like "the" tend to be shorter than content words.

In this section we discuss language-specific concerns in detail and provide recommendations.

4.2.1 Tasks influence how participants engage with the stimuli

In general, most participants in neuroscientific experiments are college students from the university at which the research is conducted. Aside from the obvious issue with sample bias [66], this type of participant may lack motivation to perform the requested task.

Lack of motivation is a clear problem for SNR, especially if the task involves the repetitions of the same stimuli, which can be boring and tiresome. Furthermore, it has long been established that the task context can affect even low-level sensory perception of the stimuli [67, 68, 69].

To combat this problem, most experiments using linguistic stimuli include an engagement task. The purpose of this task is two-fold:

1. To incentivize the correct kind of engagement with the stimuli (e.g. motivate the subject to semantically engage with the material instead of passively look at it)
2. To indicate to experimenters whether a subject failed to engage (e.g. if they answered many more questions incorrectly than their fellow participants, they were likely not engaging) so that they can be excluded from further analysis.

However, humans are very good at subverting the desired "correct" kind of engagement if another option is easier, as has been reliably observed in Mechanical Turk workers [70]. Let us take the example of the semantic engagement questions used in this thesis. Participants could use either one of the following strategies:

1. Read the sentence and integrate it semantically as soon as possible, then rehearse/store that integrated meaning until a question is asked.
2. Read the sentence without integrating it, and just rehearse the component words, integrating only if there is a question.

If our goal is to detect sentence integration, we ideally want participants to integrate every single sentence. We also want sentence integration to occur at roughly the same time. By increasing the frequency of question trials, we can hopefully motivate participants to use Strategy 1.

When interpreting results, it is important to account for task effects by asking ourselves:

- Could these results be explained by the task the participants were required to perform?
- Do these results distinguish which strategy the participants used to complete the task?
- What implications do the answers to these questions have for my scientific conclusions?

4.2.2 Language features correlate with one another

The ideal scenario in hypothesis-driven scientific experiments is that we hold all stimulus attributes constant except the variables of interest. Unfortunately, language does not lend itself easily to such a scenario. Here are some examples of how language can complicate the scientific endeavor:

1. If we want to study syntactic complexity of sentences, we are forced to admit that complexity adds length to the sentences. For example, passive-voice sentences are always longer than active-voice sentences
2. In natural language, it is unusual for inanimate objects to be the agents of verbs
3. The usage frequencies of words are correlated with their parts-of-speech and their length

It is simply not possible to account for all of these minute correlations in designing stimuli while maintaining a naturalistic language experience for the subject.

Therefore the best we can do in analyzing language data is to try to give ourselves the opportunity to correct for these correlations when making scientific inferences. Looking at the examples given:

1. Include both short active and short passive sentences. While the passive sentences are longer than their active counterparts, a short passive sentence is shorter than a normal active sentence.

2. Eliminate animacy as a confound by only including animate nouns.
3. Explicitly model the unique contributions of different properties using partial correlations or using additional regressors.

4.3 Exploratory vs Confirmatory Analysis

Many neuroscientists would agree that the field of neuroscience suffers from a reproducibility crisis [71]. There has been discussion both of the theoretical explanation of the existence of the crisis (namely, small sample sizes and marginally significant results) [72], as well as attempts to quantify the lack of reproducibility empirically [73].

There are many factors that can make reproducing the results of others difficult:

1. Low statistical power and small effect sizes
2. Failure to communicate methods and data in a transparent manner
3. Conflating exploratory and confirmatory analysis (otherwise known as HARKing - "Hypothesizing After Results are Known") [74]

All three explanations have received attention in the community, culminating in a set of concrete recommendations to improve reproducibility [75]. I want to direct special attention to the third factor, HARKing, which is likely the easiest mistake to make when conducting neuroscientific research. HARKing can lead to optimistic results that will later be difficult to reproduce, since they are the result of overfitting to the data set. Recall our neuroscientific questions:

- What happens in the brain of a person as they read a sentence?
- Do humans post-process sentences after reading?

Our first question has no hypothesis, and everything we do, from data collection to analysis, is done without a hypothesis in mind. This has implications for our results. Because we want to be able to build appropriately on prior work it is important to draw a distinction between exploratory and confirmatory work. In exploratory analysis we are given data and ask the question: what is here? Conclusions drawn from exploratory work are not affirmative scientific statements, but rather hypotheses that are potentially true and require explicit confirmation. In a confirmatory analysis, a hypothesis and analysis plan are decided in advance, and the conclusion can thus be interpreted as the truth or the falsehood of the hypothesis.

While we would never dream of directly double-dipping within a given analysis, e.g. by reporting training accuracy instead of test accuracy, when we apply many analyses with many different hyperparameters to a data set, we are effectively overfitting to that data set. This can also be thought of as creating an unintended multiple comparisons problem [76].

What can we do about this? The ideal is to perform replications of our experiments to test any discovered effects explicitly in a confirmatory manner. Another solution that may be less expensive is to reserve a held-out test set from all explorations, although in order to maintain statistical power that test set will need to be large [72]. Alternatively, the field could shift to value the honest reporting of exploratory results.

Conclusions

Answering scientific questions with certainty is difficult, and while more complex analysis tools such as machine learning can improve our ability to uncover signal in the data, they also come with caveats to bear in mind during scientific inference. Language is a difficult phenomenon to study, as it puts in direct conflict experimental controls and naturalistic experience. While exploration of a data set for clues is a critical part of the research life-cycle, it is important to bear in mind that results of exploration are optimistic and additional data is needed to confirm any discovered hypotheses.

There is hope for neuroscience, as the field moves to a more open structure (sharing data, code, publishing preprints), we will naturally produce more reliable results. More and more neuroscientists are dedicated to tackling these issues as they arise [75].

In this thesis we make a strong effort to bear all of these concerns in mind. There are two components to this work: the first is optimizing our analysis approaches to MEG data, as presented in the previous chapter. All of this tuning is performed on a pilot data set. The second component is to uncover the information about syntax and semantics available in the MEG data, so as to better understand how humans process sentences. In the next chapter, we will present exploratory results on the pilot data set that yield several key hypotheses. In the following chapter, these hypotheses are directly tested in a confirmatory manner on an untouched data set.

Chapter 5

Pilot Data Analysis

Overview

Evidence suggests that when humans read a sentence, they greedily “merge” each word they encounter into a hierarchical tree-like representation of the sentence [25]. However, sentences with atypical syntax, such as passive voice sentences, present a problem for online sentence comprehension; assuming that the sentence is active voice when parsing is incorrect and will require correction. Only when one has read the full sentence can one hope to make the correct parse. Previous work examining this problem points to the existence of a post-sentence “reanalysis” effect, but no study has revealed what information is present in the neural signal during this reanalysis time period [27, 28].

To better understand the interplay of syntax and semantics in sentence reading, and to hopefully uncover the information content of post-sentence neural activity, we designed an experiment contrasting active and passive voice sentences.

The purpose of this pilot data set was two-fold: first, to serve as a testbed for optimizing the machine learning approach to analyzing MEG data, described in Chapter 3. The second purpose of this pilot data set was to conduct an exploratory analysis of neural activity during sentence reading and post-sentence. Previous work has shown that there is a difference in neural activity between active and passive sentences, and that this difference is most visible post-sentence [27, 28]. What is unknown is what kind of information that neural signal may contain.

In this chapter we first detail the initial scientific questions we hoped to answer through exploration of this data set. We then describe the experimental design and data collection, informed by these questions. Next we outline the basic decoding approach as developed in Chapter 3.

The results presented in this chapter show the information flow decodable from the MEG signal both during sentence reading and post-sentence. From these results we distill a list of data-driven hypotheses that we then confirm in Chapter 6.

5.1 Questions of Interest

In analyzing the pilot data set we were interested in the following broad questions:

1. What is the information flow over time in the MEG signal as a sentence is being read?

2. Is there a post-sentence ‘wrap-up’ period, and what information does it contain?

The questions of information content in the MEG signal can be distilled into a series of classification tasks in a straightforward manner. Each sentence contains a noun in the role of agent (e.g. ‘dog’ in ‘A dog found the peach’ and ‘The peach was found by a dog’), and a noun in the role of patient (e.g. ‘peach’ in the previous example), along with an action verb (e.g. ‘found’). If we can classify above chance (successfully *decode*) the identity of one of these words, e.g. the verb, we can claim that information pertaining to that word is present in the MEG signal. Note that, as discussed in Chapter 4, the inverse is not true.: when a classifier fails to decode above chance, there are many potential causes of that failure (insufficient data, poor SNR, weak classifier) between which we cannot distinguish.

Using the full timecourse from sentence onset until the presentation of the following stimulus, we attempt to decode the following, training and testing on active and passive sentences separately:

1. First noun identity
2. Verb identity
3. Second noun identity

As an additional task, we attempted to decode the identity of the first determiner (‘a’ or ‘the’) from the first 500ms of the sentence, and contrasted the neural activity elicited by these determiners to that corresponding to ‘dog’, a short noun. Given that determiners are very unique linguistically, we were curious to see if this difference is detectable in neural activity.

Using the timecourse from the last word (the second noun) onset onward, we attempt to decode the following sentence attributes, using active and passive sentences separately, but also the pool of active and passive sentences:

1. Proposition identity (e.g. ‘A dog found the peach’ and ‘The peach was found by a dog’ are considered the same class) (pooled sentences only)
2. Sentence voice (pooled sentences only)
3. Agent identity
4. Patient identity
5. Verb identity
6. First Noun identity (pooled sentences only)

There were minimal *a priori* hypotheses when collecting this data set, as is evident from the exploratory nature of our questions of interest. The data were subjected to many different classification approaches in Chapter 3, and while the results are consistent independent of approach, the exact decoding accuracies shown here are likely optimistic since analysis decisions were made based on test accuracy reported in Chapter 3, leading to overfitting (see Chapter 4 for a more detailed discussion of this). However, the work presented here presents an important first step in understanding sentence processing: it reveals that a wealth of information can be detected during a post-sentence ‘wrap-up’ period, when using an optimized decoding approach.

5.2 Methods

5.2.1 Data Collection

8 neurologically healthy, right-handed native English speakers from Carnegie Mellon University's Machine Learning Department read 32 noun-verb-noun sentences that varied in voicing. The sentences are composed of 8 nouns and 4 verbs. Each of 16 propositions was presented in both active and passive voice 15 times over the course of 5 blocks (e.g. "The dog found the peach" and "The peach was found by the dog"). Each word was presented for 300ms with 200ms rest in between, and there was a 2s rest period between sentences. Comprehension questions (e.g. "Was there a vegetable?") followed 10% of sentences, to ensure semantic engagement. All 8 subjects answered close to all of the questions correctly and so all were used in the final analysis. Of the 15 trials per sentence, the first 10 were used for analysis, as we were concerned about memorization of the stimuli. See Appendix A for data recording and preprocessing details and Appendix B for the full set of stimuli.

5.2.2 Basic Decoding Approach

Throughout this chapter we follow the optimized decoding approach described in Chapter 3, in which an ℓ_2-penalized, one-vs-all logistic regression classifier is trained and tested in a sliding window fashion over the MEG data timeseries. Each window is the mean activity at 306 sensors over 100ms, concatenated over subjects to form a $306 \times 8 = 2448$ dimensional feature vector. In addition to a simple sliding window approach, we also apply the temporal generalization method [39], in which a Temporal Generalization Matrix (TGM) is constructed from the accuracies of training and testing on all possible pairs of time points. The stride between windows is 10ms.

The cross-validation scheme throughout is leave-one-sentence-out, in which all trials of the test sentence are held out and averaged together to create one high-SNR test instance. To facilitate hyperparameter selection for the ℓ_2 penalty via nested cross-validation, we train on 2 instances per sentence, where each instance is the average of 5 trials.

Significance is established via a per-timepoint permutation test with 100 permutations (note that due to computational complexity, permutation results are only available for the post-sentence decoding experiments). Multiple comparisons are corrected via controlling the false discovery rate [42]. To determine whether the effect at a given timepoint (as measured by decoding accuracy) is representative of the subject population or due to the data from only a small subset of subjects, we re-run the analysis in a leave-one-subject-out fashion, using the data from all but one subjects (thus generating a $306 \times 7 = 2142$ length feature vector). We then report the fraction of folds (out of 8) for which the accuracy was above chance. If that fraction is high, then the decoding accuracy comes from a signal that is consistently present in the subject population. If not, then only a few subjects carry the effect. For a given timepoint, it is counted as "significant" if 100% of subject folds are above chance and if it has a corrected $p < 0.05$.

5.3 Results

5.3.1 Information Flow during Sentence Reading

The TGM for each during-sentence classification task is shown in Fig. 5.1. For each sentence type, each word in the sentence is decodable as it is being read, and the verb and second noun are again decodable post-sentence. However, the strength and duration of post-sentence reactivation seems to differ: active sentences have a strong, sustained verb reactivation, while the second noun is more strongly reactivated in passive sentences.

Two issues are apparent from these results: first, that there is some problem with the stimuli, since the second noun in active sentences can be decoded during first determiner presentation. Upon further inspection, we realized that the first determiner is in fact predictive of second noun identity due to the construction of the stimulus sentences. Secondly, there is insufficient post-sentence time for passive sentences, which may lead us to miss effects. These two issues are addressed in the confirmation experiment in Chapter 6.

Focusing on the diagonal of these matrices (when the classifier was trained and tested on the same time point) can give a potentially clearer sense of information flow. The overlay of when words are decodable in active and passive sentences is shown in Fig. 5.2. These are the composites of the diagonal entries of the TGMs in Fig. 5.1.

Figure 5.2 makes certain aspects of the results more apparent. For example, while decoding accuracy can be quite high above chance, it can also be quite far below chance. The verb is actually more decodable and has more sustained decoding in passive sentences during presentation, but, as we have already noted, post-sentence verb decoding is stronger in active sentences. It is even more clear in these plots that there is insufficient time post-sentence to see all potential effects in passive sentences.

From which brain regions are we decoding the identities of these words and at which times? To examine this question, we generated importance maps from the learned classifier weights as described in [2] at 0.15s post word onset (for the word of interest) and again at 0.3s post word onset. The feature vector used to train the classifier consisted of the concatenation across subjects of mean sensor activity over 100ms. The resulting transformed importance map was averaged over subjects, giving an importance weight to each of 306 sensors. Plots of the importance of the lateral gradiometer are shown in Figs. 5.3 for the 0.15-0.25s time window and 5.4 for the 0.3-0.4s time window.

In the early time period, occipital sensors dominate; however, for the verb, right temporo-frontal sensors are also important. In second noun decoding, frontal sensors are even more pronounced. In the late time period, left and right temporal and frontal sensors are the most important for decoding, but occipital sensors are still relevant.

In this experiment, we additionally contrasted the determiners used for the nouns, using ‘a’ and ‘the.’ These two determiners turned out to be highly distinguishable from one another, with F1 score 1.0 for many timepoints. It is reasonable to ask whether the two determiners are solely distinguishable due to the difference in string length (1 vs 3 letters). As control tasks, we contrasted both determiners to the noun ‘dog’ and observed the resulting accuracy. The results are summarized in Fig. 5.5. If the sole driver of determiner distinguishability were word length, then that ‘a’ vs ‘dog’ task would achieve a similarly high accuracy to the ‘a’ vs ‘the’ task. This is evidently not the case. While the ‘a’ vs ‘dog’ task achieves a higher accuracy than the ‘the’ vs ‘dog’ task, it does not match the performance of the ‘a’ vs ‘the’ task. Therefore we must conclude that there is something beyond word length that explains the decodability of determiners.

5.3.2 Post-sentence Wrap-up

While the difference between active and passive sentences was the original primary investigative goal, it is possible that some questions about the post-sentence period can be best answered by decoding from the pool of active and passive sentences. That is, we can examine what occurs post-sentence for all sentences, independent of syntax.

The performance at each decoding task over time is shown in Fig. 5.6. In active sentences (panel A), the patient (second noun) is decodable, but the agent (first noun) is not really decodable. As obseved previously, the verb is highly decodable. In passive sentences (panel B), both the agent and the patient are decodable (the agent is less strongly decodable, but is still above chance). The verb is decodable from passive sentences as well but in two distinct windows, concurrent with the agent. On the pool of all sentences (panel C), sentence voice is highly decodable. When voice declines in decodability, the proposition becomes decodable. All sentence components (agent, patient, verb) are weakly decodable as well.

How stable are the post-sentence representations over time? We can answer this question with temporal generalization, shown in Fig. 5.7. Voice decoding (panel E) seems to evolve somewhat rapidly and then dissipate, while verb decoding highly stable.

Again, we can examine the importance map over the brain underlying the decoding accuracies that we see, using the method described in [2] at 0.61s post last word onset (the time at which proposition decoding accuracy peaks). The feature vector used to train the classifier consisted of the concatenation across subjects of mean sensor activity over 100ms. The resulting transformed importance map was averaged over subjects, giving an importance weight to each of 306 sensors. Helmet plots of the importance of the lateral gradiometer for each classification task are shown in Fig. 5.8.

The sensors that contribute most to proposition decoding are left temporal and frontal, and verb decoding is similarly supported. Sentence voice, on the other hand, draws on activation from bilateral temporal sensors, as well as left parietal sensors in the neighborhood of the inferior frontal gyrus.

Conclusions

Hypotheses for Sentence Comprehension

Our results clearly support the existence of a post-sentence ‘wrap-up’ period in which the following information is present in the MEG signal:

- The voice of the sentence
- The identity of the verb
- The identity of the agent

- The identity of the patient
- The identity of the first noun
- The identity of the sentence proposition

This refines the view that sentence processing is fully incremental by demonstrating that voice reconciliation and further processing occur once the sentence has been completely read. Another important observation is that the post-sentence information content differs between active and passive sentences. We can decode both agent and patient identity from passive sentences, but not from active sentences. These results confirm two key findings in the literature: that there are differences in the post-sentence neural representations of active and passive sentences [27, 28], and that there are separate neural signatures for nouns in the roles of agent and patient [30, 34].

We expand on what is known by showing a temporal evolution of this post-sentence activity, where sentence voice, agent, and patient, are all present in the neural signal, followed by the verb and the proposition identity. While it is impossible to prove that integration is taking place when these attributes are decodable, we clearly see evidence of both the necessary pieces of integration and the result of that integration in rapid succession post-sentence.

Through the exploration of this pilot data set, we have the following hypotheses to confirm on the confirmation data set:

1. A post-sentence wrap-up period exists and contains relevant sentence integration information, and the result of that integration: the final sentence proposition, independent of syntax.
2. This wrap-up starts at the presentation of the last word and persists for 500ms beyond its offset.

Additional results of interest from this exploration are that we can decode constituent words of a sentence as they are being read, independent of the context in which the words are presented. Furthermore, determiners seems to have highly separable neural representations, a fact that cannot solely be attributed to their visual differences.

Alternative Explanations for Results

Unfortunately the stimuli for the pilot data set are not completely balanced. This leads to two problems for interpretation:

1. The patient is always inanimate and the agent is always animate. This confounds the difference between active and passive sentences at last word presentation and possibly beyond.
2. All active sentences are shorter than all passive sentences. Therefore the difference between active and passive sentences can be confounded by sentence length.
3. The first determiner (‘the’ or ‘a’) is predictive of the second noun in active sentences, and the second determiner is predictive of the first noun in passive sentences. This causes issues for interpreting the decoding accuracy timeseries observed during sentence presentation and explains why some words can be decoded before they are presented.

We can test explanation 1 directly by running a follow-up experiment. We can train an animacy decoder on data from the presentation of the first noun. We can then apply this animacy decoder to the data from the presentation of the second noun. If the difference in noun animacy is the primary driver of our sentence voice classification results, then the animacy decoder will perform nearly as well as the voice decoder does. The result is shown in Fig. 5.9. While animacy is decodable from the neural data, it only explains the early voice decoding peak and a secondary peak at 0.4s post second noun onset.

Additionally, as previously mentioned, when one performs so many classification tasks on the same data as we did in Chapter 3, even with cross-validation, the results will start to overfit the data set. Thus the effect sizes presented here are likely optimistic.

Lastly, it is natural to ask whether the post-sentence wrap-up period observed here is an artifact of requiring subjects to answer comprehension questions after a subset of the sentences. The comprehension questions are necessary in order to ensure the engagement of subjects in the experiment; however, they do make the reading setting less natural. It is possible that the only reason we observe a post-sentence 'wrap-up' is because participants are holding the sentence in working memory in preparation for the question that may come. In light of previous work on post-sentence activity (as discussed in Chapter 2), it seems unlikely that this is the major driver of this effect. While this is the first work to our knowledge on the precise information content of the post-sentence signal, post-sentence activity has been shown to distinguish between linguistic conditions in a variety of task settings [25, 27, 28].

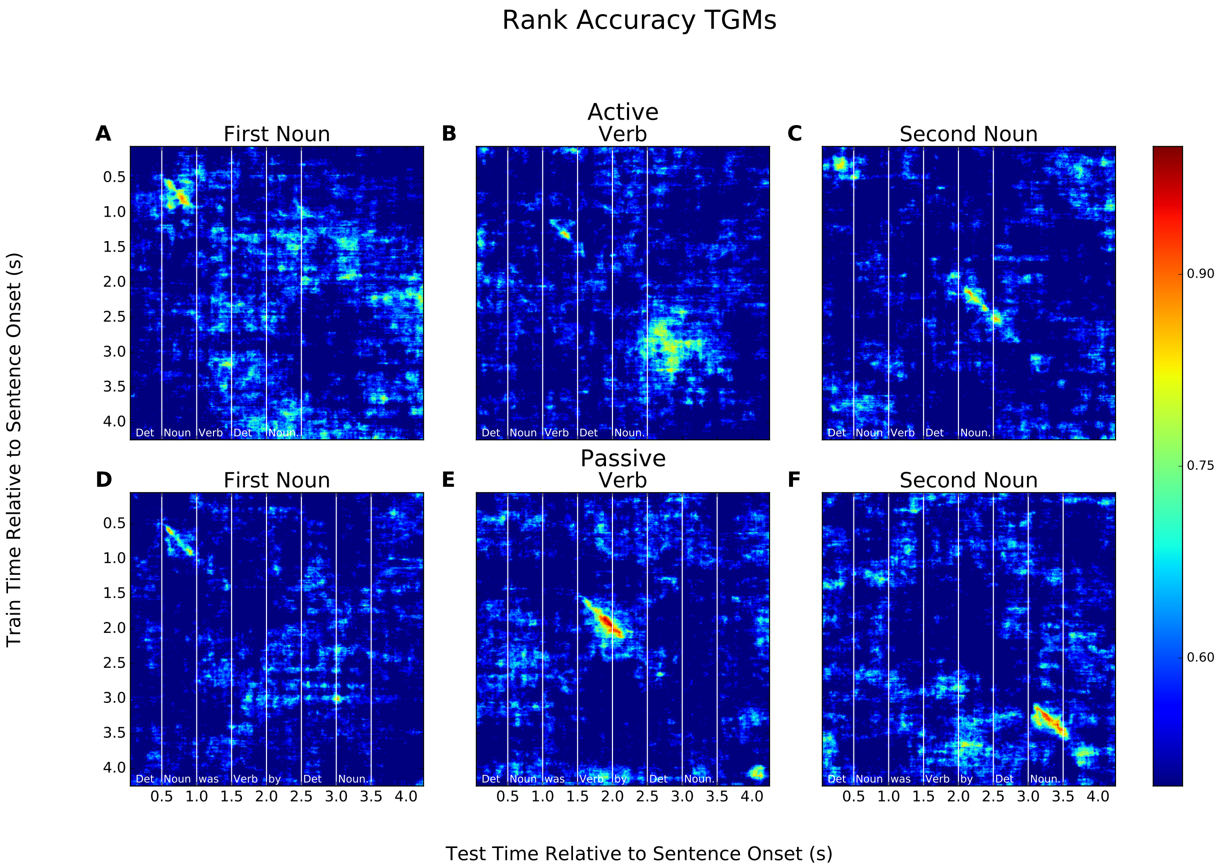


Figure 5.1: **Rank accuracy TGMs for each during-sentence classification task.** Each plot shows all pairs of training and testing timepoints over sentence presentation. The y axis indicates training time, and the x axis indicates test time, with sentence onset starting in the upper left corner. White vertical lines indicate word onsets and final sentence offset. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on either the set of active or passive sentences separately, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out. **A. TGM for decoding the first noun from active-voice sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on active voice sentences. **B. TGM for decoding the verb from active voice sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on active voice sentences. **C. TGM for decoding the second noun from active voice sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on active voice sentences. **D. TGM for decoding the first noun from passive voice sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on passive voice sentences. **E. TGM for decoding the verb from passive voice sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on passive voice sentences. **F. TGM for decoding the second noun from passive voice sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on passive voice sentences.

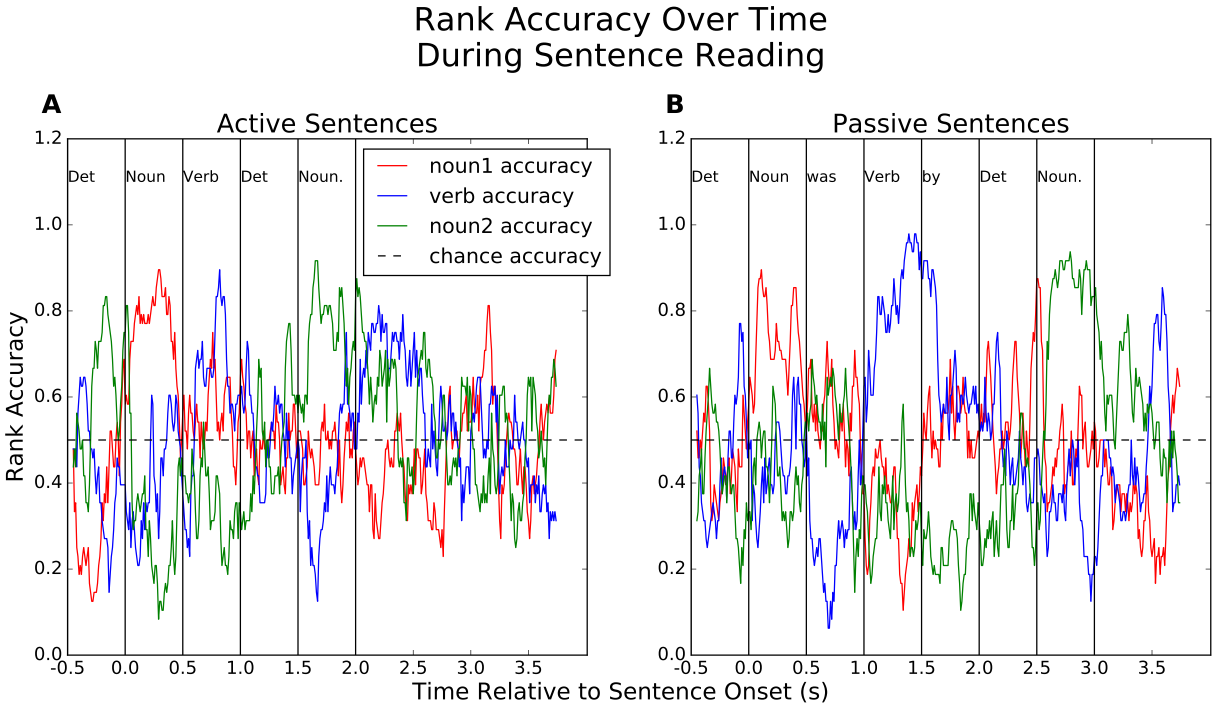


Figure 5.2: **Rank accuracy over time during sentence reading.** Each plot shows rank accuracy over time for each word in the sentence. Black vertical lines indicate word onsets and final sentence offset. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on either the set of active or passive sentences separately, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out. **A. Information flow during active voice sentence reading.** Accuracy for decoding each of the constituent words from active sentences. **B. Information flow during passive voice sentence reading.** Accuracy for decoding each of the constituent words from passive sentences.

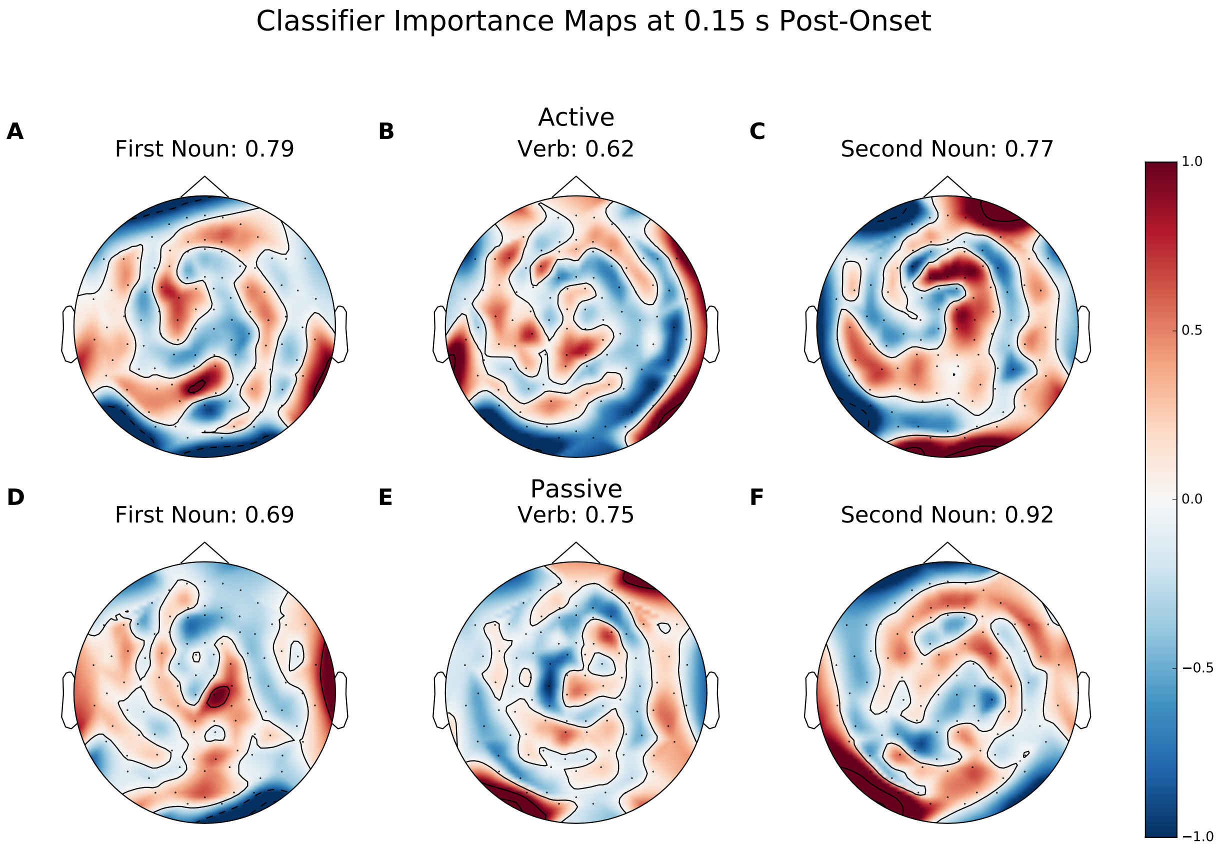


Figure 5.3: **Helmet plots for each during-sentence classification task at 0.15s post word onset.** Each plot shows the resulting importance map from training a classifier 0.15s post onset of the word of interest. Importance maps were computed from the classifier weights on the concatenation of all subjects' mean sensor activity from 0.15 to 0.25 s post onset as described in [2]. Maps were then averaged over subjects. Importance values for a single gradiometer are shown. Titles give the rank accuracy at the time examined. **A. Importance map for decoding first noun from active sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on active voice sentences. **B. Importance map for decoding verb from active sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on active voice sentences. **C. Importance map for decoding second noun from active sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on active voice sentences. **D. Importance map for decoding first noun from passive sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on passive voice sentences. **E. Importance map for decoding verb from passive sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on passive voice sentences. **F. Importance map for decoding second noun from passive sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on passive voice sentences.

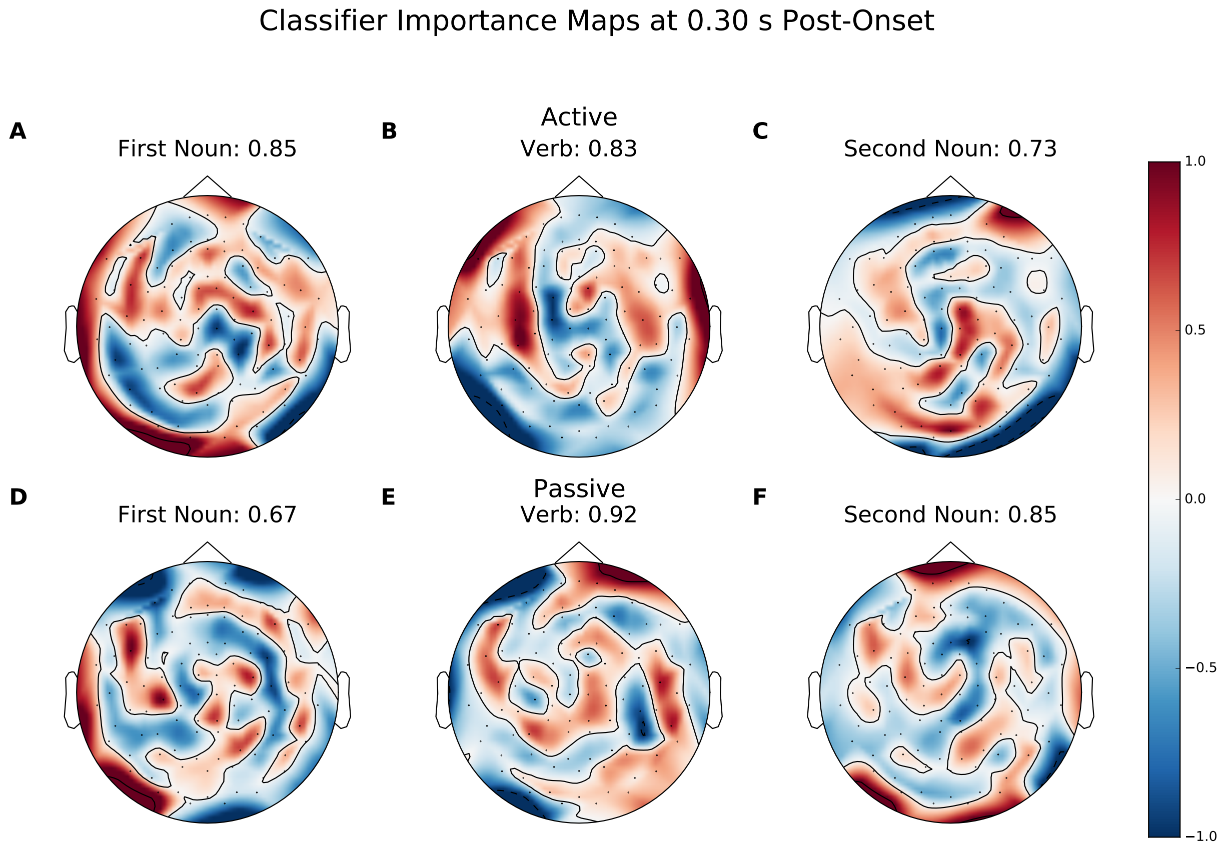


Figure 5.4: **Helmet plots for each during-sentence classification task at 0.3s post word onset.** Each plot shows the resulting importance map from training a classifier 0.3s post onset of the word of interest. Importance maps were computed from the classifier weights on the concatenation of all subjects' mean sensor activity from 0.3 to 0.4 s post onset as described in [2]. Maps were then averaged over subjects. Importance values for a single gradiometer are shown. Titles give the rank accuracy at the time examined. **A. Importance map for decoding first noun from active sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on active voice sentences. **B. Importance map for decoding verb from active sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on active voice sentences. **C. Importance map for decoding second noun from active sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on active voice sentences. **D. Importance map for decoding first noun from passive sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on passive voice sentences. **E. Importance map for decoding verb from passive sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on passive voice sentences. **F. Importance map for decoding second noun from passive sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on passive voice sentences.

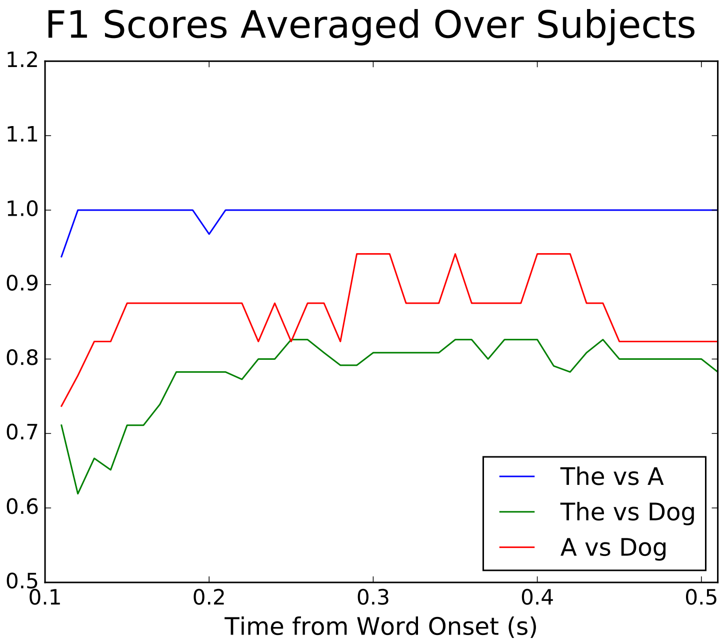


Figure 5.5: **Decodability of determiners over time.** F1 score when decoding three tasks: ‘a’ vs ‘the’, ‘a’ vs ‘dog’ and ‘the’ vs ‘dog’ over word presentation time. F1 score was used instead of classification accuracy because there are many more instances of the determiners than of ‘dog’. Classifier was balanced during training by weighting the samples, so as to eliminate bias towards the more common class.

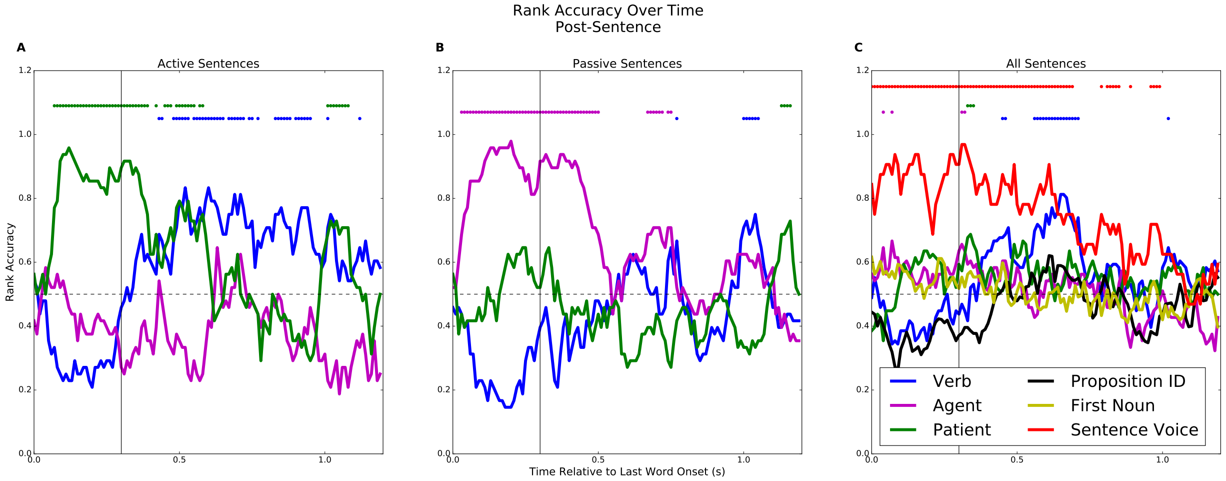


Figure 5.6: **Information flow post-sentence.** Each plot shows rank accuracy over time for each post-sentence classification task. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on either the set of active and passive sentences separately as well as the pool of all sentences, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out. The black vertical line indicates the offset of the second noun (the end of the sentence). **A. Decoding accuracy post active-sentences.** Classification accuracy of agent, verb and patient training and testing on active sentences only. **B. Decoding accuracy post passive-sentences.** Classification accuracy of agent, verb and patient training and testing on passive sentences only. **C. Decoding accuracy post all sentences.** Classification accuracy of agent, verb, patient, voice, first noun, and proposition, training and testing on the pool of all sentences.

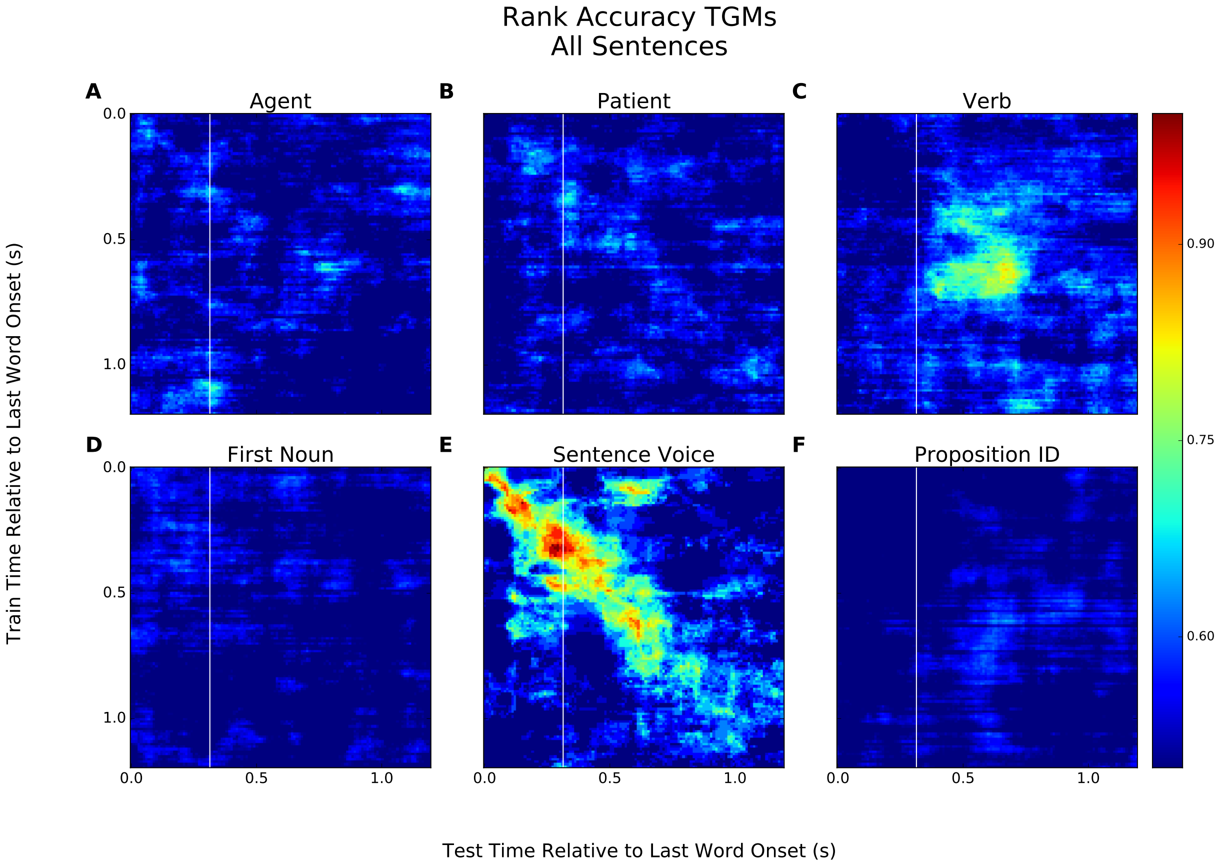


Figure 5.7: **Rank accuracy TGMs for each post-sentence classification task.** Each plot shows all pairs of training and testing timepoints from second noun presentaiton onwards. The y axis indicates training time, and the x axis indicates test time, with word onset starting in the upper left corner. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on the pool of both active and passive sentences, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out.The white line indicates the offset of the second noun (the end of the sentence). **A. TGM for decoding the agent of the sentence.** Classification task was to detect the identity of the agent. In active sentences, this is the first noun, and in passive sentences it is the second noun. **B. TGM for decoding the patient of the sentence.** Classification task was to detect the identity of the patient. In active sentences, this is the second noun, and in passive sentences it is the first noun. **C. TGM for decoding the verb of the sentence.** Classification task was to detect the identity of the verb. **D. TGM for decoding the first noun of the sentence.** Classification task was to detect the identity of the first noun. In active sentences this is the agent, and in passive sentences it is the patient. **E. TGM for decoding the voice of the sentence.** Classification task was to detect the voice of the sentence (active or passive). **F. TGM for decoding the proposition of the sentence.** Classification task was to detect the proposition of the sentence (where the active and passive version of the proposition were given the same class label).

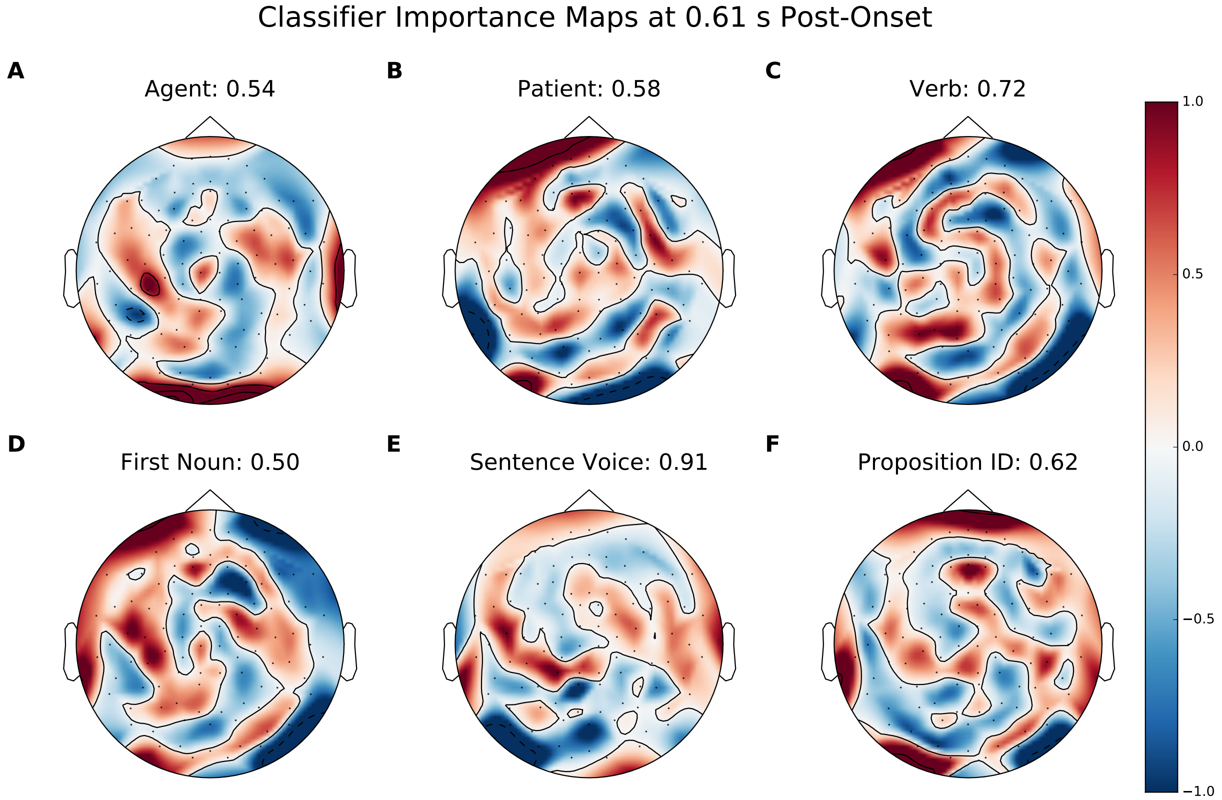


Figure 5.8: **Helmet plots for each during-sentence classification task at 0.61s post last word onset.** Each plot shows the resulting importance map from training a classifier 0.61s post onset of the last word in the sentence. Importance maps were computed from the classifier weights on the concatenation of all subjects' mean sensor activity from 0.61 to 0.71 s post onset as described in [2]. Maps were then averaged over subjects. Importance values for a single gradiometer are shown. Classification tasks were conducted on the pool of both active and passive sentences. Titles give the rank accuracy at the time examined. **A. Importance map for decoding the agent.** Classification task was to detect the identity of the agent of the sentence, which in passive sentences is the second noun and in active sentences is the first noun. **B. Importance map for decoding the patient.** Classification task was to detect the identity of the patient of the sentence, which in passive sentences is the first noun and in active sentences is the second noun. **C. Importance map for decoding verb.** Classification task was to detect the identity of the verb of the sentence. **D. Importance map for decoding first noun.** Classification task was to detect the identity of the first noun of the sentence. **E. Importance map for decoding sentence voice.** Classification task was to detect the voice (active or passive) of the sentence. **F. Importance map for decoding proposition.** Classification task was to detect the identity of the proposition of the sentence.

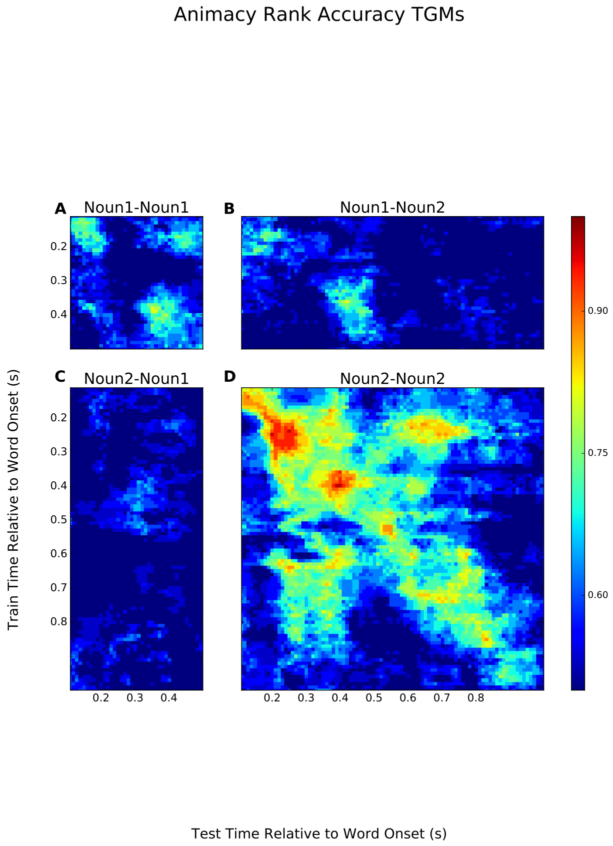


Figure 5.9: **Rank accuracy TGMs crossing noun presentations for animacy decoding.** Each plot shows a rank accuracy TGM for a pair of training and testing scenarios from which we attempted to decode noun animacy. The y axis indicates training time, and the x axis indicates test time, with word onset starting in the upper left corner. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on the pool of both active and passive sentences, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out. **A. Training on first noun, testing on first noun.** Animacy decoding results when training and testing on data from first noun presentation. **B. Training on first noun, testing on second noun.** Animacy decoding results when training on data from first noun presentation (agnostic to sentence length) and testing on data from second noun presentation. **C. Training on second noun, testing on first noun.** Animacy decoding results when training on data from second noun presentation and testing on data from first noun presentation. **D. Training on second noun, testing on second noun.** Animacy decoding results when training on data from second noun presentation and testing on data from second noun presentation.

Chapter 6

Confirmatory Data Analysis

Overview

The neuroscientific goal of this thesis has been to better understand how the brain processes sentences, even when those sentences differ in syntactic structure. To that end we have collected MEG data from participants reading active and passive voice sentences and used a decoding approach to understand the information content in the neural signal both during sentence reading and post-sentence.

In Chapter 5 we discussed a small pilot data set that we used to refine our decoding approach and to conduct an exploratory analysis of the data. From that pilot experiment, we were able to determine several key methodological choices (discussed in more detail in Chapter 3), as well as distill some hypotheses about sentence reading.

We collected a second data set for two reasons:

1. To confirm the (likely optimistic) results found in Chapter 5.
2. To present more balanced stimuli than were presented in the pilot data set.

In spirit, the experiment discussed in this chapter is a replication of the pilot experiment: we again contrasted active and passive sentences. However, there are three key stimulus changes (for example stimuli, see Table 6.1):

1. We present both short and long active and passive sentences.
2. All nouns are animate nouns.
3. The only determiner used is "the."

By adjusting the stimuli, we are able to draw stronger scientific conclusions by accounting for potential confounds. Furthermore, the new set of stimuli enables us to ask additional scientific questions that the pilot data set could not answer.

In this chapter we present two lines of analysis: the confirmation of the Chapter 5 hypotheses, and the exploration of further questions. The first line of analysis is conducted with the decoding approach that we tuned in Chapter 3, while the second uses RSA, modified to be more sensitive to MEG data (as discussed in Chapter 3).

6.1 Questions of Interest

6.1.1 Hypotheses to Confirm

In Chapter 5 we uncovered a post-sentence ‘wrap-up’ period in which the following information is present in the MEG signal:

- The voice of the sentence
- The identity of the verb
- The identity of the agent
- The identity of the patient
- The identity of the first noun (independent of role as agent or patient)
- The identity of the sentence proposition

This can be distilled into the two following hypotheses:

1. A post-sentence wrap-up period exists and contains relevant sentence integration information, and the result of that integration.
2. This wrap-up starts at the presentation of the last word and persists for 1.5s beyond its onset.

Furthermore we observed that the wrap-up period for passive sentences differed from that of active sentences. Specifically, we found that while both agent and patient were decodable from active sentences, only the patient was decodable from post-sentence data in active sentences.

The stimuli in the pilot data set were unfortunately not balanced: sentence voice is confounded both by sentence length and by the fact that all agents were animate nouns and all patients were inanimate nouns. While we seek to confirm the pilot result of the post-sentence wrap-up period, this experiment further refines it by accounting for these stimulus confounds.

6.1.2 Further Exploratory Questions

Because the confirmatory data set is unencumbered by confounds such as sentence length and noun animacy, it can potentially answer additional questions.

To what extent is sentence length decodable from the data? We hypothesize that sentence length (short vs long) will be highly decodable, based on results showing that the number of open nodes in a syntactic tree correlates with neural activity [25]. This is an additional classification task that can only be applied to the confirmatory data set.

An additional question of interest that is corollary to that of proposition identity: how well can we decode argument binding? That is, how well can a classifier distinguish “The man approached the woman” from “The woman approached the man”? This is related to decoding proposition identity insofar as succeeding at this task will improve proposition decoding performance. However, it is possible to simply decode the tuple of (first noun, verb, second noun), and *not* the desired tuple of (agent, verb, patient) and still perform well at the proposition identity decoding task. For example, the classifier gets credit for distinguishing “The boy kicked the girl” from “The man approached the woman,” but that does not require getting the argument binding right. Decoding accuracy for argument binding can be derived from the proposition identity de-

Syntactic Category	Example Sentence
Long Active	The man approached the woman.
Short Active	The man approached.
Long Passive	The woman was approached by the man.
Short Passive	The woman was approached.

Table 6.1: **Sample Stimuli from Confirmatory Experiment**

coding task by rescoring the classifier's predictions to only count whether the correct binding is ranked above the incorrect binding.

As discussed in Chapter 3, RSA is a useful analysis technique that, when optimized for MEG, can help us evaluate models of sentence processing based on their ability to correlate with neural activity. In this chapter we present an application of RSA to the confirmatory data set that contrasts the following:

1. A pure syntax model, in which sentences are considered similar if they have similar syntactic structure.
2. A pure semantic model, in which a bag-of-words semantic representation is created for each sentence.
3. An integration model, that uses both syntactic and semantic information to compute the similarity between sentences.

By comparing these three models in terms of their ability to explain post-sentence activity, we can complement our decoding analysis. The decoding analysis shows which components of the sentence are distinguishable post-sentence, whereas the RSA analysis shows which type of global information (syntax, semantics, an integration of both) best characterizes the neural activity.

6.2 Methods

6.2.1 Data Collection

26 neurologically healthy, native English speakers recruited from the city of Pittsburgh, PA read 32 simple sentences that varied in voicing. The sentences are composed of combinations of 4 nouns (man, woman, girl, boy) and 4 verbs (helped, approached, kicked, punched). Sentences belonged to four syntactic categories, with 8 sentences in each category, summarized in Table 6.1. See Appendix A for data recording and preprocessing details and C for complete stimulus set.

Each sentence was presented a total of 10 times over the course of 5 blocks. Each word was presented for 300ms with 200ms rest in between, and there was a 3s rest period between sentences. Comprehension questions (e.g. "Did she do nothing?") followed 25% of sentences, to ensure semantic engagement.

All but one of the subjects were right-handed; the left-handed participant was excluded from analysis. Of the remaining 25 subjects, one had data quality that was too poor to use, and 4 failed to answer greater than 50% of the engagement questions correctly. These 5 were excluded from analysis, for a total of 20 subjects remaining.

Of the 20 remaining subjects, 10 returned for structural scans that enabled us to source-localize their data. Data were source-localized using Minimum Norm Estimation (MNE) [85]. Sources were spaced 7mm apart using an icosahedral structure. The source-localized data was then parcellized using the Freesurfer atlas so that decoding results could be obtained for each region-of-interest (ROI).

6.2.2 Decoding Experiments

In Chapter 5 we optimized our decoding approach for the classification tasks of interest. Here we use the same parameters (with no adjustment) to cleanly attempt to replicate the pilot results, namely an ℓ_2-penalized, one-vs-all logistic regression classifier is trained and tested in a sliding window fashion over the MEG data timeseries. Each window is the mean activity at 306 sensors over 100ms, concatenated over subjects to form a $306 \times 20 = 6120$ dimensional feature vector. The lack of parameter exploration on our part is the crucial element that allows this study to constitute a replication of the pilot study. For each task we computed the temporal generalization matrix (TGM), training and testing our classifier on all pairs of timepoints, cross-validating in a leave-one-sentence-out manner.

To replicate the results found in Chapter 5 that showed that the constituent words of a sentence are decodable as they are being read, we attempted to decode the first noun, verb, and second noun (when applicable) of each sentence, training and testing on active and passive sentences separately, during sentence reading and post-sentence.

We also attempted to decode the following post-sentence: agent, patient, verb, first noun, voice, and proposition identity. These are replications of the results presented in Chapter 5 . In addition to these tasks we attempted to decode sentence length (long vs short) from the post-sentence activity.

Because some of the sentences presented in the experiment did not contain a second noun, that decoding task, as well as the agent, patient and proposition decoding tasks, was run only on the long sentences from this experiment.

Significance is established via a per-timepoint permutation test with 100 permutations (note that due to computational complexity, permutation results are only available for the post-sentence decoding experiments). Multiple comparisons are corrected via controlling the false discovery rate [42]. To determine whether the effect at a given timepoint (as measured by decoding accuracy) is representative of the subject population or due to the data from only a small subset of subjects, we re-run the analysis in a leave-one-subject-out fashion, using the data from all but one subjects (thus generating a $306 \times 19 = 5814$ length feature vector). We then report the fraction of folds (out of 20) for which the accuracy was above chance. If that fraction is high, then the decoding accuracy comes from a signal that is consistently present in the subject population. If not, then only a few subjects carry the effect. As in Chapter 5, for a given timepoint to be counted as “significant”, 100% of the subject folds must be above chance and it must have a corrected $p < 0.05$.

As an additional analysis, we ran six of the post-sentence decoding tasks (voice, verb, agent, patient, and argument binding) on a per-ROI basis using the 10 source-localized subjects. For each ROI, the sources from each subject were concatenated together (as we did with the sensors) to form the feature vector. Again, a 100ms window average was used.

6.2.3 RSA Experiments

Representational Similarity Analysis (RSA) is an alternative approach to analyzing neural data, with its own distinct advantages over decoding [7]. RSA can allow us to contrast different theoretical models of sentences in terms of their ability to capture the similarity structure of the stimuli in the neural activity. We contrasted three models of sentence similarity in order to try and disentangle different aspects of sentence reading: a pure syntax model, a pure semantics model, and a hierarchical integration model. The model RDMs, with sample stimuli labeled, are shown in Fig. 6.1

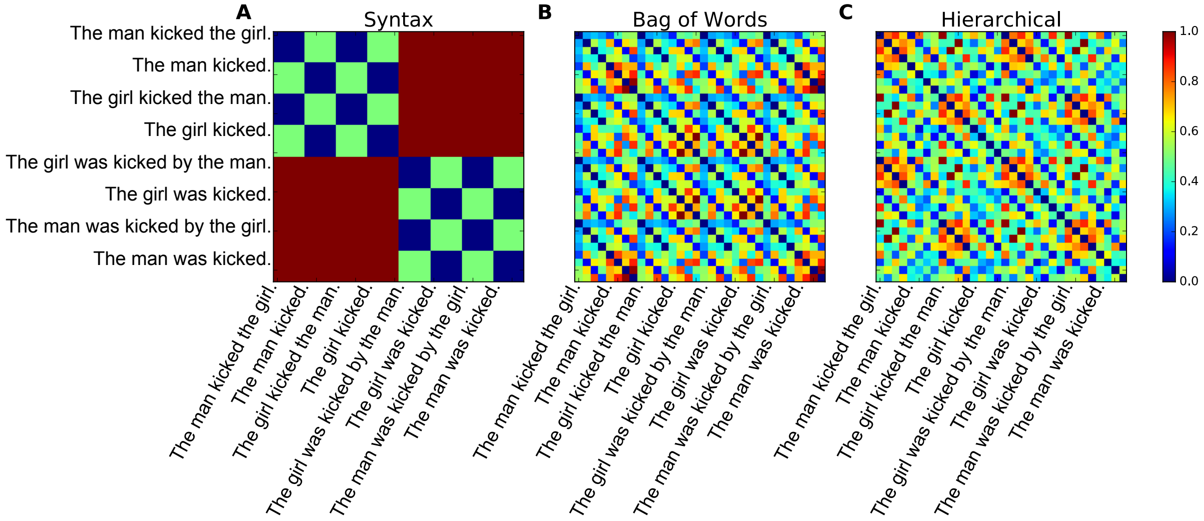


Figure 6.1: **Model RDMs.** Model Representational Dissimilarity Matrices (RDMs) for exploratory RSA analysis. Each element of an RDM is the distance between a pair of stimuli according to a theoretical model. The rows are ordered active sentences first, followed by passive sentences. Sentence length alternates every four rows. **A. Syntax RDM.** Each entry is determined by the voice and the length of the given sentence pair. A different in voice is assigned a distance of 1, while a difference in length is assigned a distance of 0.5 if the voices are the same. **B. Bag of Words RDM.** Each sentence is represented by the average of the GloVe vectors for the constituent words. The entries of this RDM are the euclidean distances between these sentence representations. **C. Hierarchical RDM.** Each entry is the average of the pairwise distances between the GloVe vectors for the agent, patient, and verb. For short sentences, the missing noun (agent or patient) is represented by the average over the nouns in the experiment

The first model, referred to as the Syntax model (Fig. 6.1A), attempts to capture structural information about the sentences such as length and voice. If two sentences have the same voice and are the same length, they are assigned a distance of 0. If they are the same voice but a different length, they are assigned a distance of 0.5. Sentences that differ in voice are assigned a distance of 1.0.

The second model, referred to as the Bag of Words model (Fig. 6.1B), is intended to capture the semantic content of the sentence without making use of any syntactic or structural information (hence the name Bag of Words). First, we represented each verb and noun in the experiment by its GloVe vector [77], which captures the semantic content of the word from its usage in a large corpus. For a given sentence, its Bag of Words representation is the average of the vectors for its constituent content words. An RDM element is the euclidean distance between these Bag of Words representations for two sentences.

The third model, referred to as the Hierarchical model (Fig. 6.1C), is intended to capture the fully integrated meaning of the sentence, using both semantic and syntactic cues. Again, we represented each verb and noun in the experiment by its GloVe vector [77]. Instead of building a sentence-level representation from these vectors, we computed the sentence distances directly. For a given pair of sentences, the euclidean distance between the agent vectors, verb vectors, and patient vectors were each computed separately. The sentence distance is the average of these three distances. For short sentences, which only have one noun, the second noun was represented for the purpose of distance computation by the average GloVe vector of all 4 nouns in the experiment.

For the whole brain analysis, sensor-level data was used to compute MEG RDMs in a sliding-window fashion over time. We focused on the post-sentence time period for this analysis, aligning the sentences by the presentation of the last word.

As discussed in Chapter 3, we computed the rank (Spearman) correlation at each timepoint between the MEG RDM and each of the model RDMs. The three models are not correlated with one-another (max correlation: 0.08), so zero-order correlations are all that is needed. Significance was evaluated via the Mantel test, which amounts to a label permutation test, with 10,000 permutations [48]. Multiple comparisons correction was done by controlling the false discovery rate [42]. The noise ceiling was computed by repeatedly splitting the trials in half, as described in Chapter 3. The lower bound is the correlation between disjoint sets of trials, and the upper bound is the correlation between the average over half the trials and the average over all the trials.

6.3 Results

6.3.1 Confirmatory Decoding Results

As in Chapter 5, our first goal was to see whether we can decode the constituent words of the sentence as they are being read, and how stable the neural representations of these words are over time. To that end, we computed temporal generalization matrices for the first noun, the verb and the second noun of the sentence, training and testing on active and passive sentences separately. For verb and first noun decoding, the pool of long and short sentences were used. Since only the long sentences contained a second noun, that decoding task was restricted to long sentences only. The resulting TGMs, in Fig. 6.2, demonstrate that once again, we can decode these constituent words during sentence reading, and that the representations of these words change rapidly and evolve over time.

Note that these plots by and large are similar to those shown in Fig. 5.1. A key difference, however, is verb decoding from active sentences (Fig. 6.2B). Whereas the pilot experiment results showed a robust reactivation of the verb post-sentence in active sentences, that is not replicated in the confirmation experiment. However, there is a strong symmetric pattern of off-diagonal decoding accuracy.

The diagonal entries of the TGMs, shown in Fig. 6.3, show remarkably similar patterns to those computed from the pilot data, albeit with a much smaller post-sentence effect. An additional difference is that while during-sentence verb decoding accuracy was higher for passive sentences in the pilot data set, it is higher for active sentences in the confirmation data set.

From which brain regions are we decoding the identities of these words and at which times? To examine this question, we generated importance maps from the learned classifier weights as described in [2] at 0.15s post word onset (for the word of interest) and again at 0.3s post word onset. The feature vector used to train the classifier consisted of the concatenation across subjects of mean sensor activity over 100ms. The resulting transformed importance map was averaged over subjects, giving an importance weight to each of 306 sensors. Plots of the importance of the lateral gradiometer are shown in Figs. 5.3 for the 0.15-0.25s time window and 5.4 for the 0.3-0.4s time window.

Again the early decoding period is dominated by occipital electrodes, while late decoding accuracy is supported by a combination of frontal and temporal electrodes.

To boost our sensitivity for post-sentence decoding, we again aligned the sentences by the presentation of the last word and attempted to decode agent, verb, and patient, from the long active sentences and long passive sentences separately. These accuracy traces are shown in Fig. 6.6, panels A and B. Here we see a replication of a key result from the pilot experiment: both the agent and the patient are decodable post-sentence from passive voice sentences, but not from active voice sentences.

On the pool of all sentences (active and passive, long and short), we decoded sentence length (long vs short), sentence voice (active vs passive), and verb identity. From the pool of long active and passive sentences we attempted to decode agent, patient, and proposition identity. These accuracy traces are shown in Fig. 6.6C.

We see a confirmation of the post-sentence wrap-up period on the pooled set of sentences. First sentence length and sentence voice are highly decodable from last word onset until around 1.4s post onset. Once voice and sentence length decoding accuracy start to decline, the proposition can be decoded with high accuracy. The components of the proposition can be decoded as well. This confirms our observation that structural and syntactic information is processed first, followed by a semantic integration of the sentence.

How stable are the representations of the concepts decodable in Fig. 6.6C? We examined this question using temporal generalization matrices for each task, shown in Fig. 6.7.

Sentence length (Fig. 6.7D) is strongly decodable throughout the post-sentence time period, but the representation seems to evolve quite rapidly in that there are very few off-diagonal points of high accuracy. Sentence voice (Fig. 6.7F) is also decodable throughout but peaks just after sentence offset, persisting until 1.4s post last word onset. Again, the neural representation evolves quite rapidly over this time period.

The proposition, verb and agent are decodable only after sentence voice and sentence length are no longer decodable. Patient does not seem very decodable but is another example of off-diagonal accuracy without a corresponding on-diagonal accuracy.

Again, we can examine the importance map over the brain underlying the decoding accuracies that we see, using the method described in [2] at 1.44s post last word onset (the time at which proposition decoding accuracy peaks). The feature vector used to train the classifier consisted of the concatenation across subjects of mean sensor activity over 100ms. The resulting transformed importance map was averaged over subjects, giving an importance weight to each of 306 sensors. Helmet plots of the importance of the lateral gradiometer for each classification task are shown in Fig. 5.8.

Just as we observed in the pilot experiment, proposition decoding is supported by left temporal and parietal sensors. Sentence voice decoding is also supported by similar regions.

6.3.2 Source Localized Decoding Results

While helmet plots such as those in Fig. 6.8 can be helpful in understanding which sensors underlie the observed decoding accuracy, but they fail to provide strong answers regarding which brain regions are responsible for which computational tasks. By source-localizing MEG data, we can estimate the cortical dipoles that produced the observed sensor activity. Using those source-activations, we can better understand the neurobiological underpinnings of the task of interest. To better understand the regions underlying the post-sentence decoding results, we applied six of the post-sentence decoding tasks (voice, verb, agent, patient, proposition, and argument binding) to source-localized data on a per-region basis.

Voice decoding, summarized in Fig. 6.9 is supported primarily by bilateral occipital cortex activation. Additionally, early in the post-sentence time period we see left pars opercularis (Broca's Area) as yielding high decoding accuracy as well. Surprisingly, bilateral tempero-parietal junction and more parietal areas also yielded high decoding accuracy.

Verb identity was much less decodable on a per-region basis, as is evident from Fig. 6.10, but was supported by supported by left superior temporal regions and Broca's area. Of additional interest is the strong decodability from the premoter strip, bilaterally.

A very striking result from the source-localized analyses is the difference in accuracy patterns for agent and patient decoding. On the left hemisphere, early agent decoding is supported largely by frontal regions, whereas early patient decoding is weakly supported by the temporal lobe and inferior parietal cortex. However, late post-sentence, agent and patient are both decodable from roughly the same set of left hemisphere regions, centering on parietal cortex.

Proposition identity (see Fig. 6.13) and argument binding (see Fig. 6.14) show, as predicted, a similar pattern of decodability over the brain. In the period immediately following the sentence, left superior temporal cortex dominates, but at the very end of the sentence, as with agent and patient decoding, a diffuse set of parietal regions contain proposition and argument binding information.

6.3.3 Exploratory RSA Results

The whole brain scores over time for each model are shown in Fig. 6.15. As could be predicted from the decoding results, the syntax model correlates strongly with the MEG data for nearly the entire time period. The Bag of Words model fails to correlate significantly with the neural activity at any timepoint.

The Hierarchical model correlates with the neural activity about 1 second after the onset of the last word, which is the same time during which the proposition identity is decodable from the neural activity (see Fig. 6.6).

Conclusions

In this study we were able to replicate several key results from Chapter 5. During sentence reading, the identities of the constituent words of the sentence are decodable. Because the decoding accuracy for each of these tasks persist beyond word presentation and have a rapidly evolving representation (see Fig. 6.2), the decoding of these words is unlikely to be due solely to the visual aspect of the stimulus.

After the sentence has been presented and the subject is looking at a blank screen (and awaiting a question), we again observed a post-sentence wrap-up period. This wrap-up period contained the following information:

- Sentence Length
- Sentence Voice
- Proposition Identity
- Argument Binding
- Agent Identity
- Patient Identity
- Verb Identity

Sentence length is decodable from two sources: the fact that short sentences end with verbs and long sentences end with nouns, but also the true length of the sentence. This is reflected in the TGM in Fig. 6.7A. Unlike in the pilot experiment, sentence length is distinct from sentence voice, which is also decodable post-sentence. Sentence voice in this experiment is only discernible from true syntactic differences between the sentences.

Early (relative to sentence end) voice decodability is supported by bilateral occipital cortex and left pars opercularis (Broca's area). In the second 500ms of post-sentence time, the verb and the proposition identity (and the arugment binding) are decodable from superior temporal cortex. In the late portion of the post-sentence time period, agent, patient, verb, and proposition are all decodable from parietal cortex.

Using RSA, we were able to confirm and complement our decoding result. As was the case in decoding, syntactic and sentence length information were highly decodable post-sentence. Purely semantic information (as represented both by the word decoding tasks and the Bag of Words RSA model), could not be recovered.

However, an integration of semantic and syntactic information is detectable in the neural signal. We were able to significantly but not strongly decode the proposition identity from the post-sentence neural activity, about 1 second after the onset of the last word of the sentence. Additionally, an integration-based RSA model significantly correlated with neural activity *and* this correlation was within noise tolerance. Again, this significant correlation occurred 1 second after the onset of the last word of the sentence.

Taking all of these results together, we can refine our picture of the results uncovered in the pilot experiment in Chapter 5. We can confirm that there is a post-sentence wrap-up period. This post-sentence wrap-up period contains syntactic/structural information about the sentence. Additionally, 1s after the onset of the last word, we see, using two separate analysis approaches, that the integrated meaning of the sentence is present in the post-sentence neural activity.

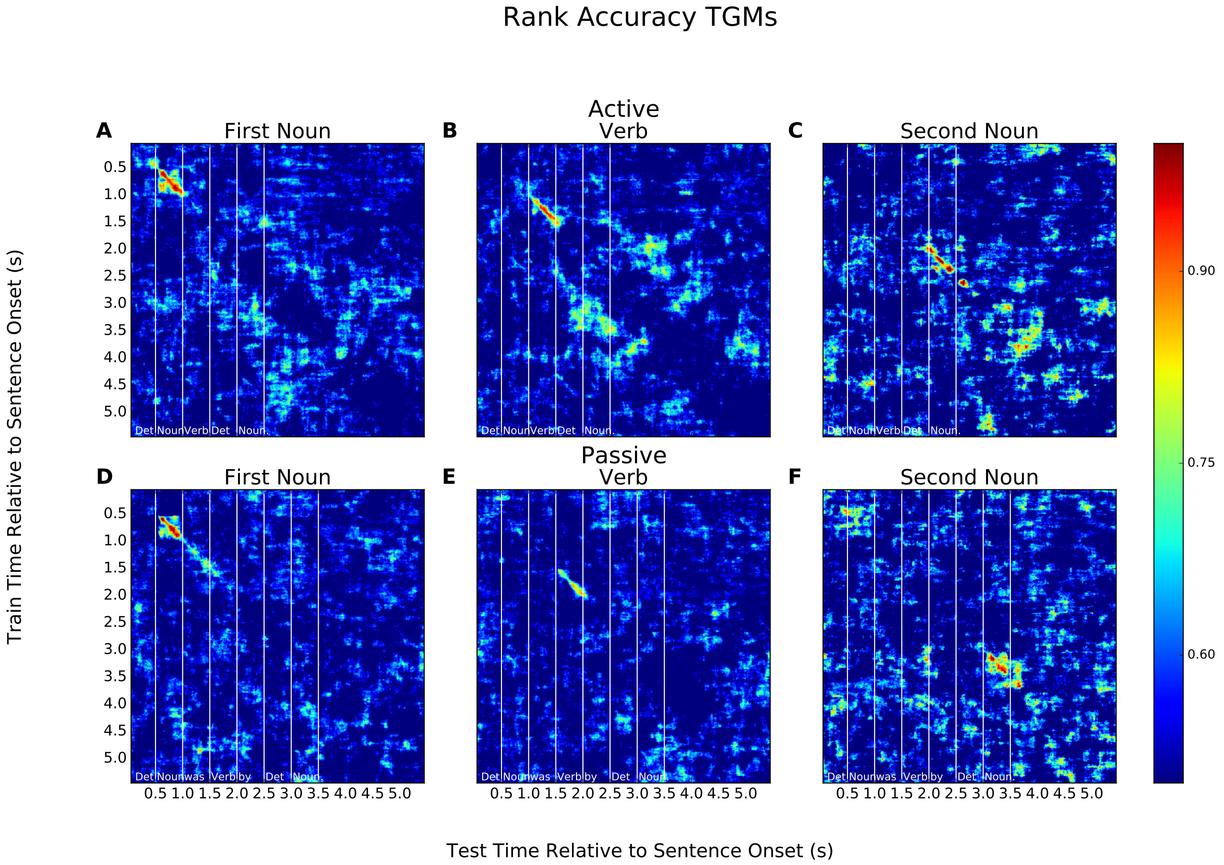


Figure 6.2: **Rank accuracy TGMs for each during-sentence classification task.** Each plot shows all pairs of training and testing timepoints over sentence presentation. The y axis indicates training time, and the x axis indicates test time, with sentence onset starting in the upper left corner. White vertical lines indicate word onsets and final sentence offset. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on either the set of active or passive sentences separately, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out. **A. TGM for decoding the first noun from active-voice sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on active-voice sentences. **B. TGM for decoding the verb from active-voice sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on active-voice sentences. **C. TGM for decoding the second noun from active-voice sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on long active-voice sentences. **D. TGM for decoding the first noun from passive-voice sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on passive-voice sentences. **E. TGM for decoding the verb from passive-voice sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on passive-voice sentences. **F. TGM for decoding the second noun from passive-voice sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on long passive-voice sentences.

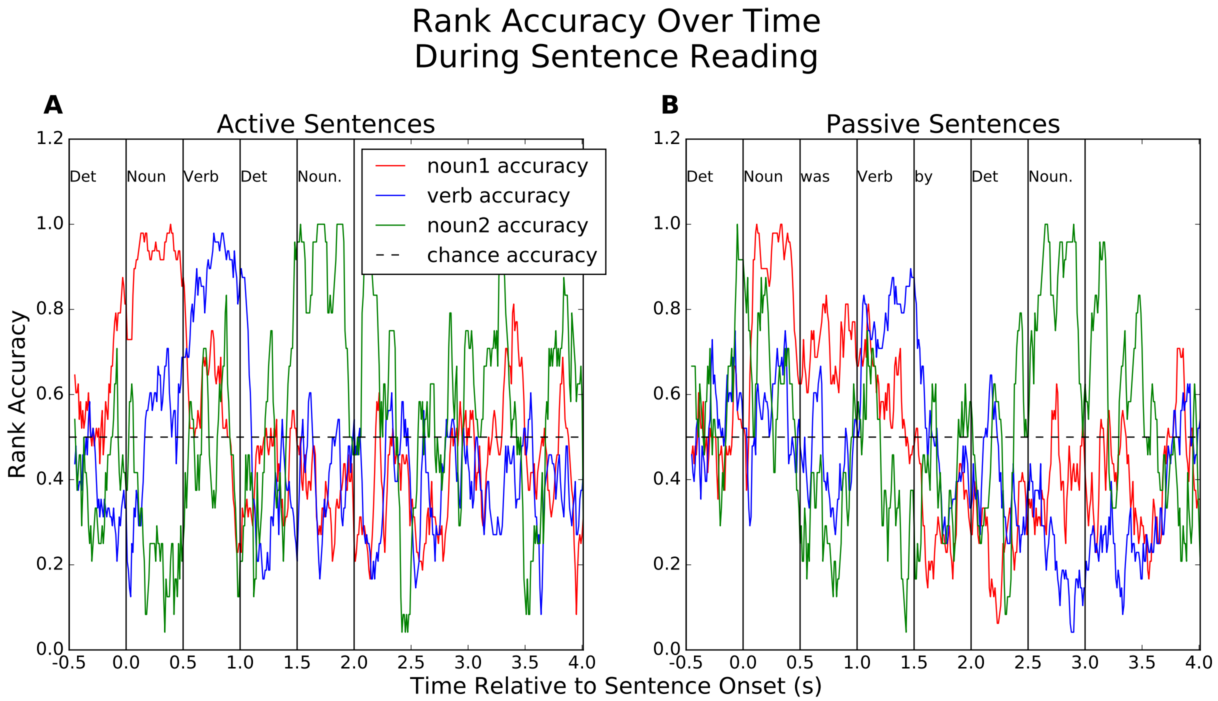


Figure 6.3: **Rank accuracy over time during sentence reading.** Each plot shows rank accuracy over time for each word in the sentence. Black vertical lines indicate word onsets and final sentence offset. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on either the set of active or passive sentences separately, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out. **A. Information flow during active-voice sentence reading.** Accuracy for decoding each of the constituent words from active sentences. **B. Information flow during passive-voice sentence reading.** Accuracy for decoding each of the constituent words from passive sentences.

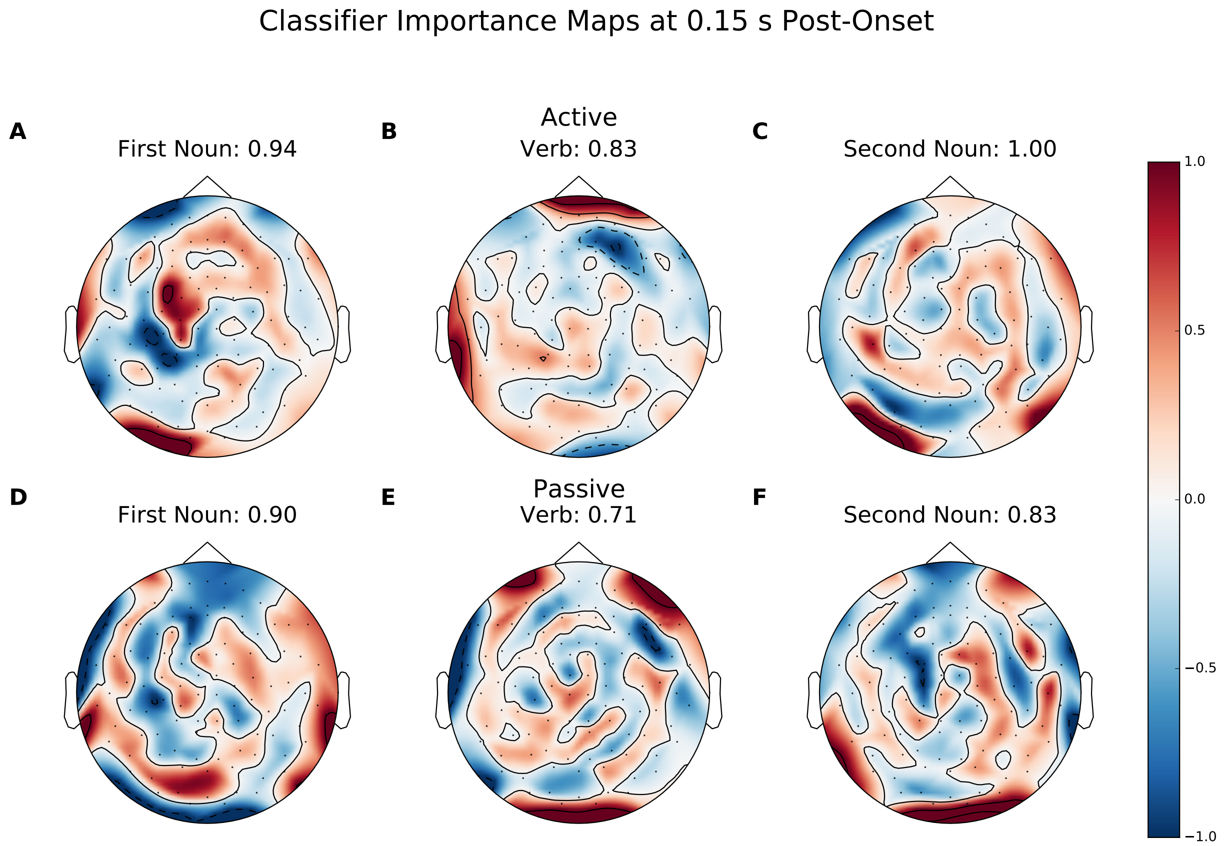


Figure 6.4: **Helmet plots for each during-sentence classification task at 0.15s post word onset.** Each plot shows the resulting importance map from training a classifier 0.15s post onset of the word of interest. Importance maps were computed from the classifier weights on the concatenation of all subjects' mean sensor activity from 0.15 to 0.25 s post onset as described in [2]. Maps were then averaged over subjects. Importance values for a single gradiometer are shown. Titles give the rank accuracy at the time examined. **A. Importance map for decoding first noun from active sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on active-voice sentences. **B. Importance map for decoding verb from active sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on active-voice sentences. **C. Importance map for decoding second noun from active sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on long active-voice sentences. **D. Importance map for decoding first noun from passive sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on passive-voice sentences. **E. Importance map for decoding verb from passive sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on passive-voice sentences. **F. Importance map for decoding second noun from passive sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on long passive-voice sentences.

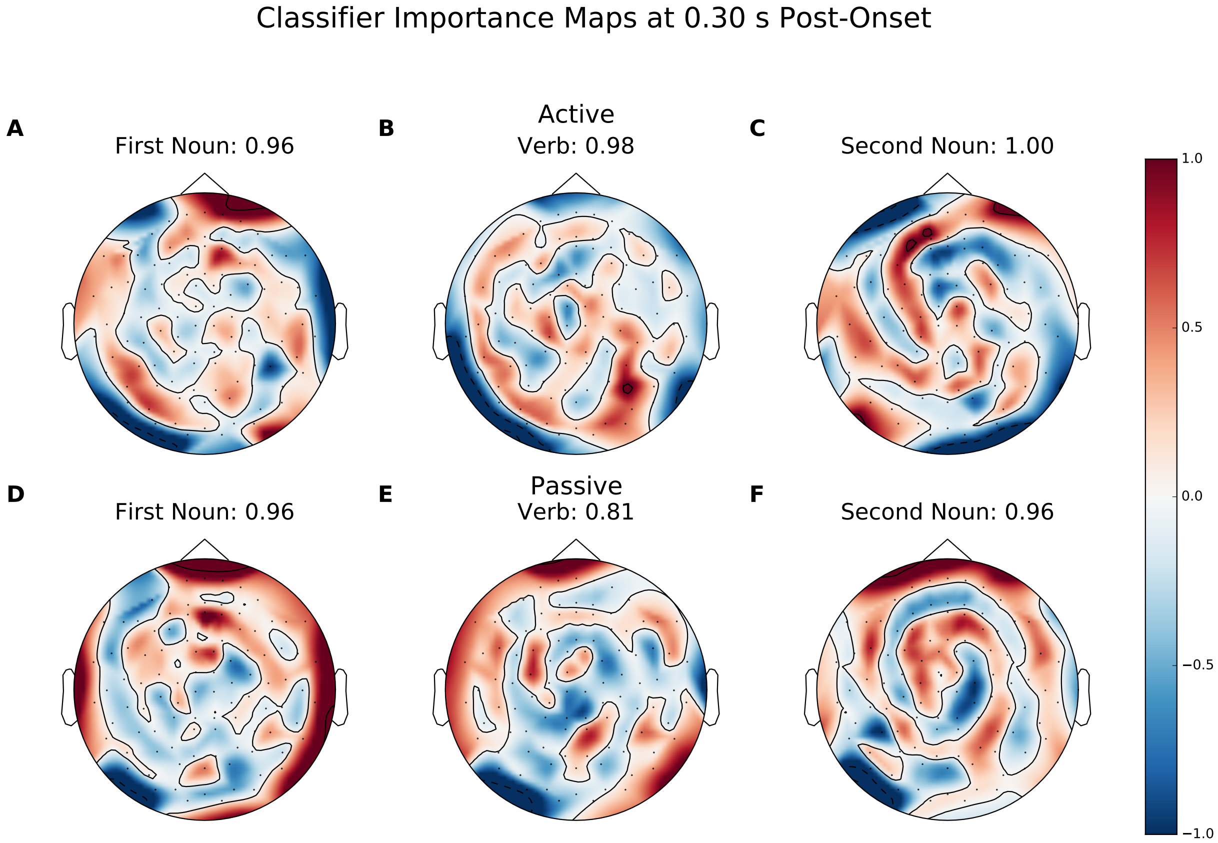


Figure 6.5: **Helmet plots for each during-sentence classification task at 0.3s post word onset.** Each plot shows the resulting importance map from training a classifier 0.3s post onset of the word of interest. Importance maps were computed from the classifier weights on the concatenation of all subjects' mean sensor activity from 0.3 to 0.4 s post onset as described in [2]. Maps were then averaged over subjects. Importance values for a single gradiometer are shown. Titles give the rank accuracy at the time examined. **A. Importance map for decoding first noun from active sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on active-voice sentences. **B. Importance map for decoding verb from active sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on active-voice sentences. **C. Importance map for decoding second noun from active sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on long active-voice sentences. **D. Importance map for decoding first noun from passive sentences.** Classification task was to detect the identity of the first noun of the sentence, training and testing only on passive-voice sentences. **E. Importance map for decoding verb from passive sentences.** Classification task was to detect the identity of the verb of the sentence, training and testing only on passive-voice sentences. **F. Importance map for decoding second noun from passive sentences.** Classification task was to detect the identity of the second noun of the sentence, training and testing only on long passive-voice sentences.

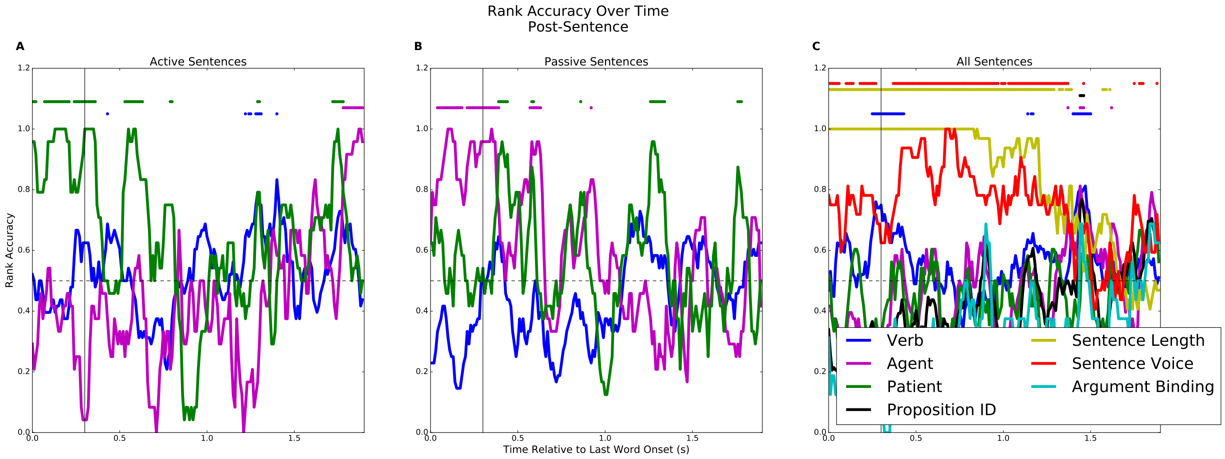


Figure 6.6: **Information flow post-sentence.** Each plot shows rank accuracy over time for each post-sentence classification task. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on either the set of active and passive sentences separately as well as the pool of all sentences, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out. The black vertical line indicates the offset of the last word (the end of the sentence). **A. Decoding accuracy post active-sentences.** Classification accuracy of agent, verb and patient training and testing on long active sentences only. **B. Decoding accuracy post passive-sentences.** Classification accuracy of agent, verb and patient training and testing on long passive sentences only. **C. Decoding accuracy post all sentences.** Classification accuracy of sentence length, agent, verb, patient, voice, first noun, and proposition, training and testing on the pool of all sentences. Agent, patient and proposition decoding tasks were conducted only on the set of long sentences.

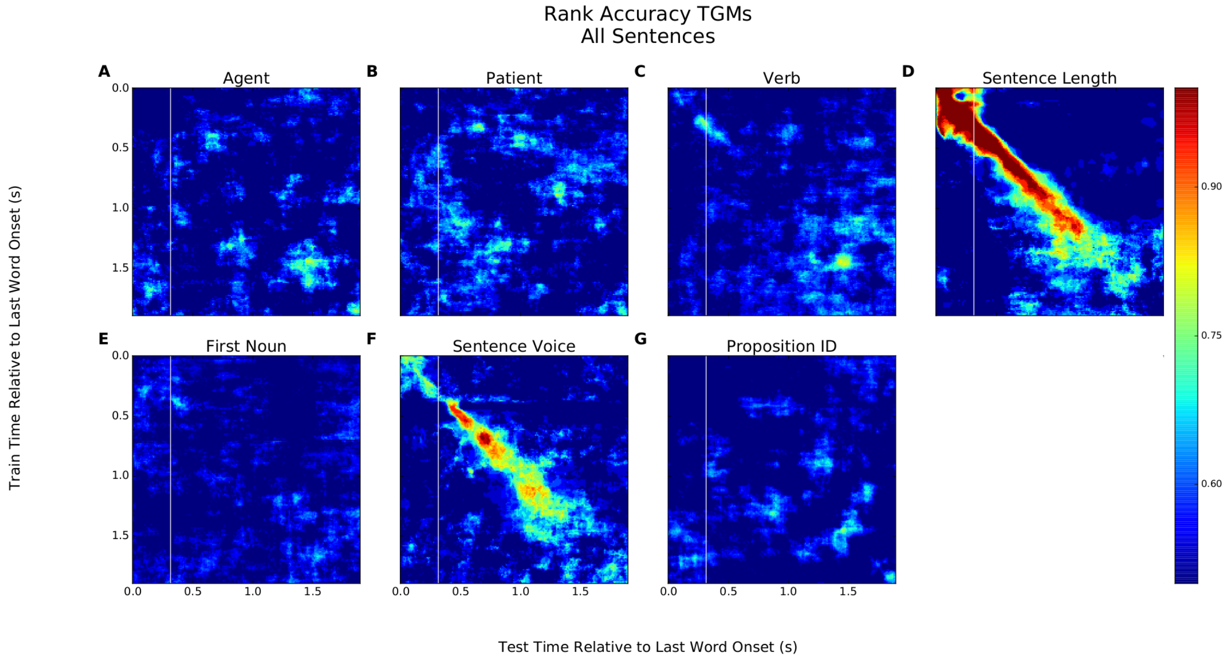


Figure 6.7: **Rank accuracy TGMs for each post-sentence classification task.** Each plot shows all pairs of training and testing timepoints over sentence presentation. The y axis indicates training time, and the x axis indicates test time, with sentence onset starting in the upper left corner. Chance rank accuracy is 0.5, with a maximum value of 1.0. Classification was performed on the pool of both active and passive sentences, using the mean sensor activation over a 100ms window concatenated across subjects as the feature vector. Cross-validation was leave-one-sentence out.The white line indicates the offset of the second noun (the end of the sentence). **A. TGM for decoding the agent of the sentence.** Classification task was to detect the identity of the agent from the pool of long sentences. In active sentences, this is the first noun, and in passive sentences it is the second noun. **B. TGM for decoding the patient of the sentence.** Classification task was to detect the identity of the patient from the pool of long sentences. In active sentences, this is the second noun, and in passive sentences it is the first noun. **C. TGM for decoding the verb of the sentence.** Classification task was to detect the identity of the verb from the pool of all sentences (short and long). **D. TGM for decoding sentence length** Classification task was to detect whether the sentence was long or short (two nouns or one noun). **E. TGM for decoding the first noun of the sentence.** Classification task was to detect the identity of the first noun from the pool of all sentences (short and long). In active sentences this is the agent, and in passive sentences it is the patient. **F. TGM for decoding the voice of the sentence.** Classification task was to detect the voice of the sentence (active or passive) from the pool of all sentences (short and long). **G. TGM for decoding the proposition of the sentence.** Classification task was to detect the proposition of the sentence (where the active and passive version of the proposition were given the same class label) from the pool of long sentences.

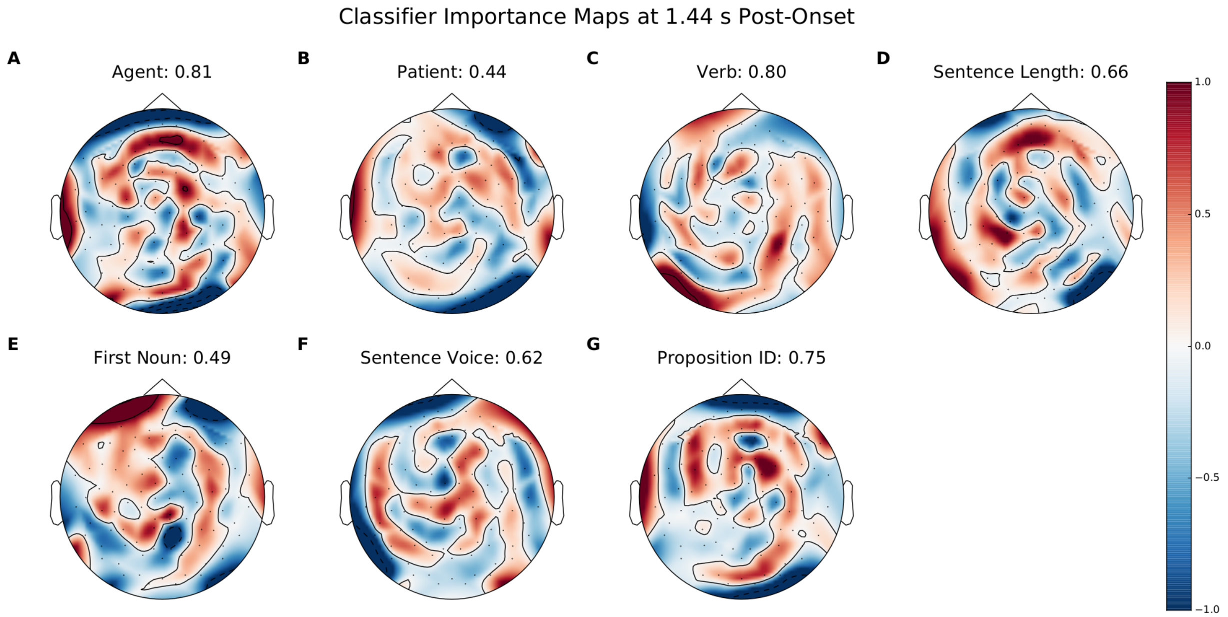


Figure 6.8: **Helmet plots for each during-sentence classification task at 1.44s post last word onset.** Each plot shows the resulting importance map from training a classifier 1.44s post onset of the last word in the sentence. Importance maps were computed from the classifier weights on the concatenation of all subjects' mean sensor activity from 1.44 to 1.54 s post onset as described in [2]. Maps were then averaged over subjects. Importance values for a single gradiometer are shown. Classification tasks were conducted on the pool of both active and passive sentences. Titles give the rank accuracy at the time examined. **A. Importance map for decoding the agent.** Classification task was to detect the identity of the agent of the sentence, which in passive sentences is the second noun and in active sentences is the first noun, from the pool of long sentences. **B. Importance map for decoding the patient.** Classification task was to detect the identity of the patient of the sentence, which in passive sentences is the first noun and in active sentences is the second noun, from the pool of long sentences. **C. Importance map for decoding verb.** Classification task was to detect the identity of the verb of the sentence, from the pool of long sentences. **D. Importance map for decoding sentence length** Classification task was to detect whether the sentence was long or short (two nouns or one noun). **E. Importance map for decoding first noun.** Classification task was to detect the identity of the first noun of the sentence. **F. Importance map for decoding sentence voice.** Classification task was to detect the voice (active or passive) of the sentence. **G. Importance map for decoding proposition.** Classification task was to detect the identity of the proposition of the sentence from the pool of long sentences.

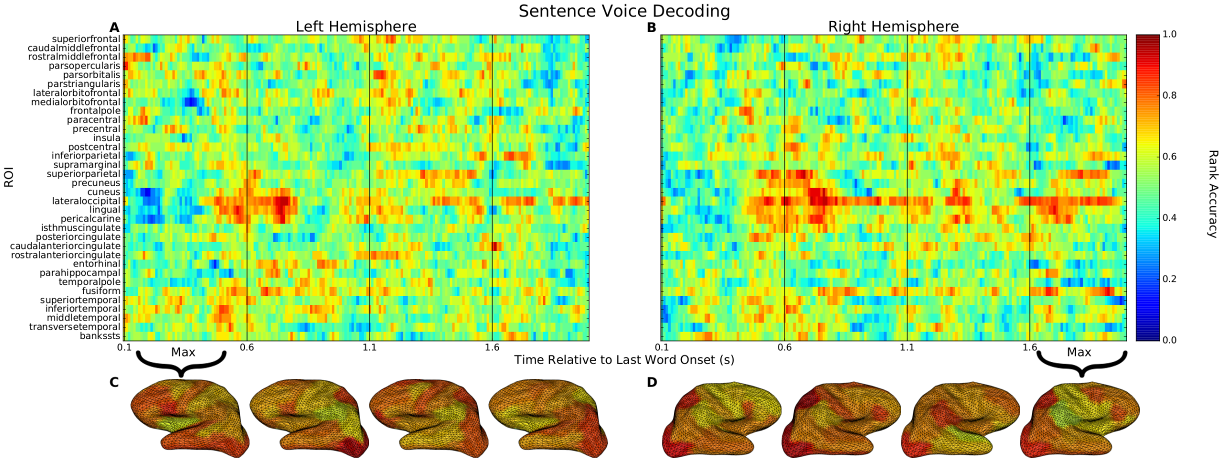


Figure 6.9: **Voice Decoding Accuracy by Region and Time.** Rank Accuracy for decoding per-ROI over time for the voice decoding task. **A. Heatmap of accuracy by ROI and Time, Left Hemisphere** Each point in the grid is the rank accuracy for voice decoding at a given region and time point. The y-axis lists left hemisphere regions (from the Freesurfer atlas) and the x-axis shows time relative to last word onset. The vertical lines divide the post-sentence time period into 4 500ms periods. **B. Heatmap of accuracy by ROI and Time, Right Hemisphere** Same as in **A** but for right hemisphere regions. **C. Inflated Left Hemisphere Brain Maps of Accuracy for 500ms Periods.** For each 500ms window indicated in **A**, we took the max over time for each region and plotted it on the inflated brain map. **D. Inflated Right Hemisphere Brain Maps of Accuracy for 500ms Periods.** Same as in **C** but for right hemisphere regions.

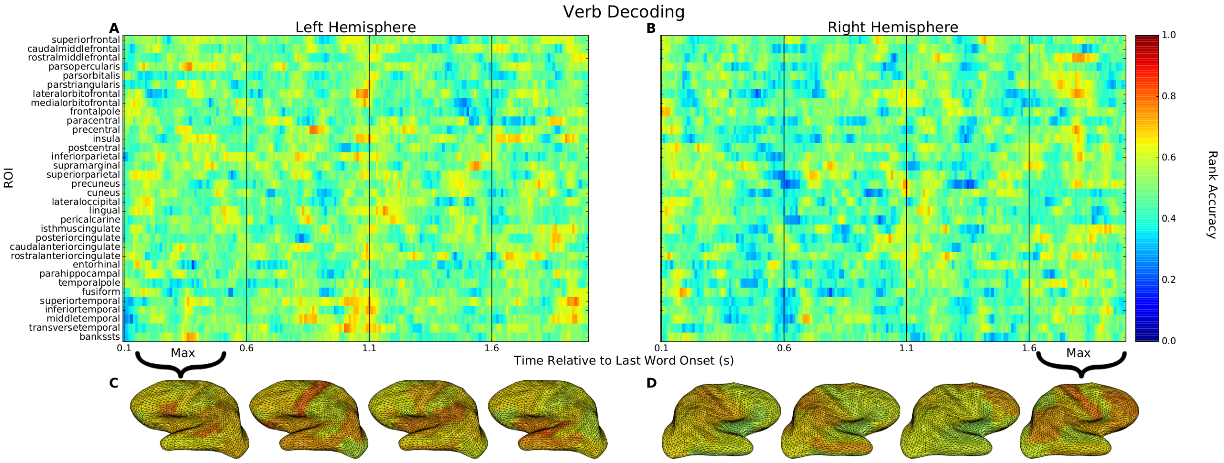


Figure 6.10: **Verb Decoding Accuracy by Region and Time.** Rank Accuracy for decoding per-ROI over time for the verb decoding task. **A. Heatmap of accuracy by ROI and Time, Left Hemisphere** Each point in the grid is the rank accuracy for verb decoding at a given region and time point. The y-axis lists left hemisphere regions (from the Freesurfer atlas) and the x-axis shows time relative to last word onset. The vertical lines divide the post-sentence time period into 4 500ms periods. **B. Heatmap of accuracy by ROI and Time, Right Hemisphere** Same as in **A** but for right hemisphere regions. **C. Inflated Left Hemisphere Brain Maps of Accuracy for 500ms Periods.** For each 500ms window indicated in **A**, we took the max over time for each region and plotted it on the inflated brain map. **D. Inflated Right Hemisphere Brain Maps of Accuracy for 500ms Periods.** Same as in **C** but for right hemisphere regions.

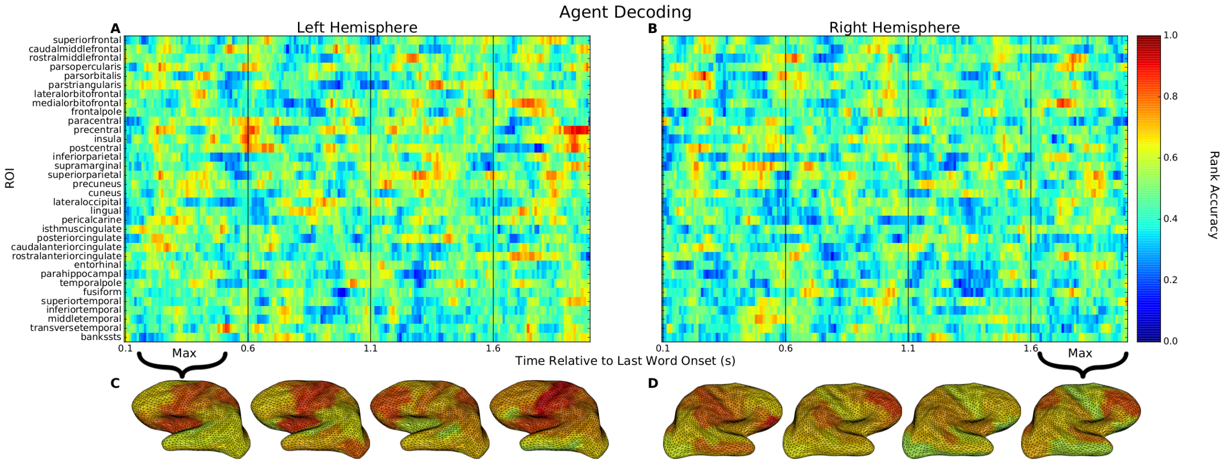


Figure 6.11: **Agent Decoding Accuracy by Region and Time.** Rank Accuracy for decoding per-ROI over time for the agent decoding task. **A. Heatmap of accuracy by ROI and Time, Left Hemisphere** Each point in the grid is the rank accuracy for agent decoding at a given region and time point. The y-axis lists left hemisphere regions (from the Freesurfer atlas) and the x-axis shows time relative to last word onset. The vertical lines divide the post-sentence time period into 4 500ms periods. **B. Heatmap of accuracy by ROI and Time, Right Hemisphere** Same as in **A** but for right hemisphere regions. **C. Inflated Left Hemisphere Brain Maps of Accuracy for 500ms Periods.** For each 500ms window indicated in **A**, we took the max over time for each region and plotted it on the inflated brain map. **D. Inflated Right Hemisphere Brain Maps of Accuracy for 500ms Periods.** Same as in **C** but for right hemisphere regions.

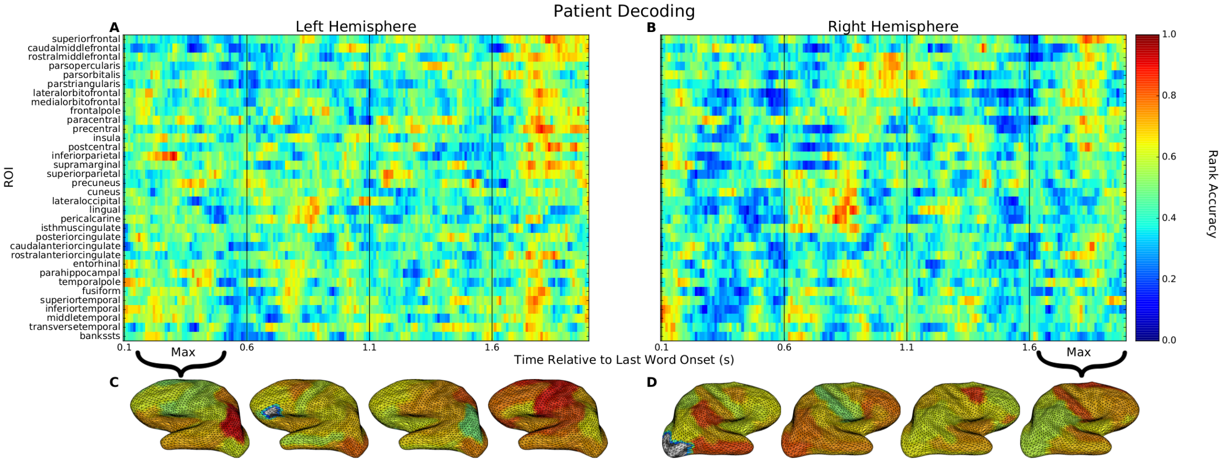


Figure 6.12: **Patient Decoding Accuracy by Region and Time.** Rank Accuracy for decoding per-ROI over time for the patient decoding task. **A. Heatmap of accuracy by ROI and Time, Left Hemisphere** Each point in the grid is the rank accuracy for patient decoding at a given region and time point. The y-axis lists left hemisphere regions (from the Freesurfer atlas) and the x-axis shows time relative to last word onset. The vertical lines divide the post-sentence time period into 4 500ms periods. **B. Heatmap of accuracy by ROI and Time, Right Hemisphere** Same as in **A** but for right hemisphere regions. **C. Inflated Left Hemisphere Brain Maps of Accuracy for 500ms Periods.** For each 500ms window indicated in **A**, we took the max over time for each region and plotted it on the inflated brain map. **D. Inflated Right Hemisphere Brain Maps of Accuracy for 500ms Periods.** Same as in **C** but for right hemisphere regions.

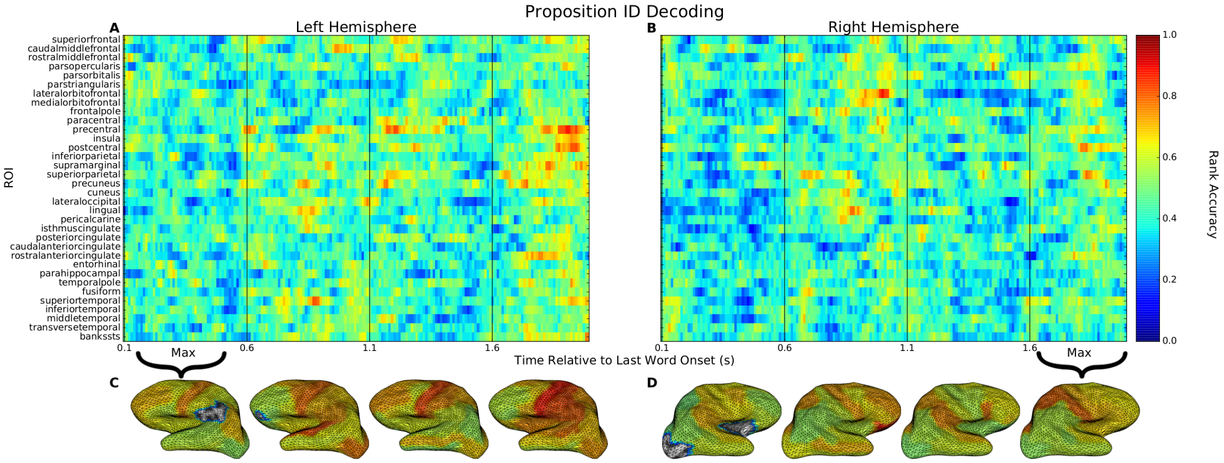


Figure 6.13: **Proposition Decoding Accuracy by Region and Time.** Rank Accuracy for decoding per-ROI over time for the proposition decoding task. **A. Heatmap of accuracy by ROI and Time, Left Hemisphere** Each point in the grid is the rank accuracy for proposition decoding at a given region and time point. The y-axis lists left hemisphere regions (from the Freesurfer atlas) and the x-axis shows time relative to last word onset. The vertical lines divide the post-sentence time period into 4 500ms periods. **B. Heatmap of accuracy by ROI and Time, Right Hemisphere** Same as in **A** but for right hemisphere regions. **C. Inflated Left Hemisphere Brain Maps of Accuracy for 500ms Periods.** For each 500ms window indicated in **A**, we took the max over time for each region and plotted it on the inflated brain map. **D. Inflated Right Hemisphere Brain Maps of Accuracy for 500ms Periods.** Same as in **C** but for right hemisphere regions.

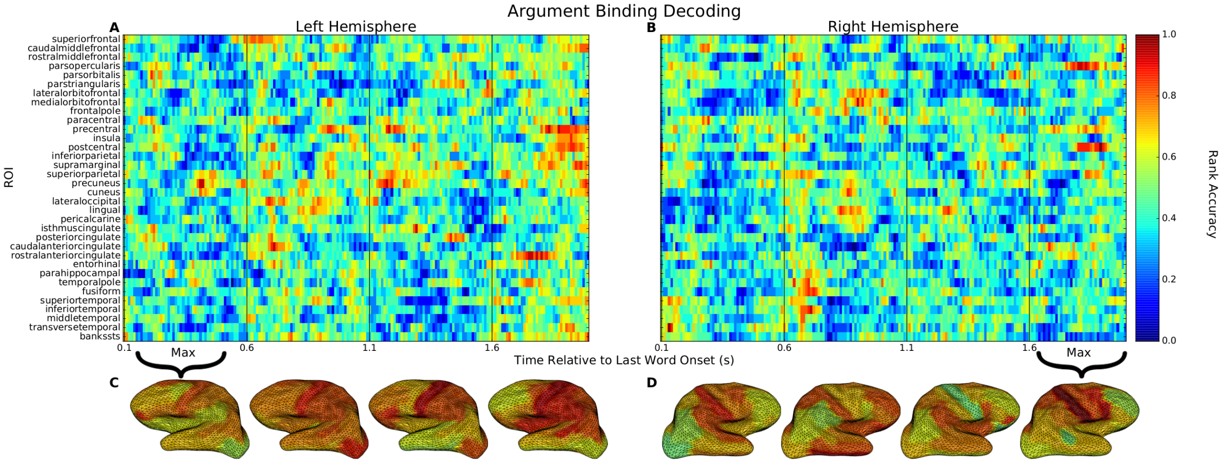


Figure 6.14: **Argument Binding Decoding Accuracy by Region and Time.** Rank Accuracy for decoding per-ROI over time for the argument binding decoding task. **A. Heatmap of accuracy by ROI and Time, Left Hemisphere** Each point in the grid is the rank accuracy for argument binding decoding at a given region and time point. The y-axis lists left hemisphere regions (from the Freesurfer atlas) and the x-axis shows time relative to last word onset. The vertical lines divide the post-sentence time period into 4 500ms periods. **B. Heatmap of accuracy by ROI and Time, Right Hemisphere** Same as in **A** but for right hemisphere regions. **C. Inflated Left Hemisphere Brain Maps of Accuracy for 500ms Periods.** For each 500ms window indicated in **A**, we took the max over time for each region and plotted it on the inflated brain map. **D. Inflated Right Hemisphere Brain Maps of Accuracy for 500ms Periods.** Same as in **C** but for right hemisphere regions.

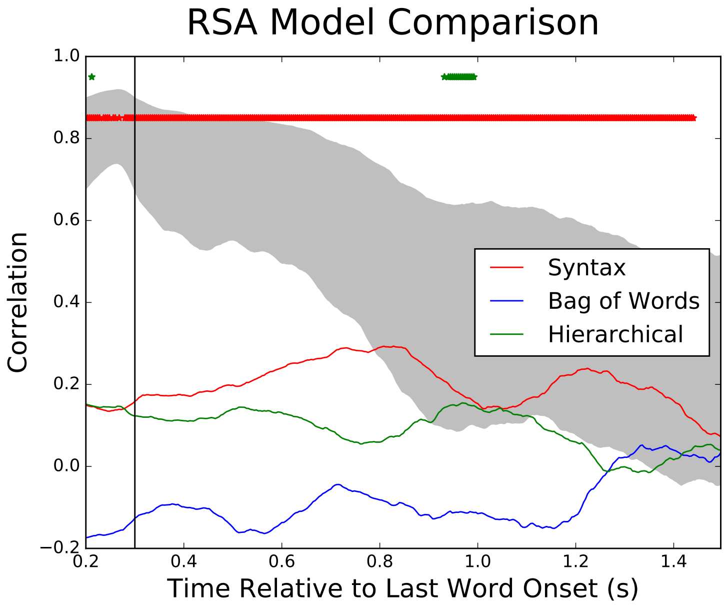


Figure 6.15: **Whole brain RSA model comparison.** Spearman correlation between each model RDM in Fig. 6.1 and a MEG data RDM constructed from whole brain sensor-level data. Noise ceiling is shown in gray. Corrected significance is indicated via starred points.

Chapter 7

Conclusion and Future Work

Overview

In order to better understand sentence comprehension in the brain, we recorded magnetoencephalography (MEG) data from humans while they read active and passive sentences. This choice of stimuli allowed us to dissociate syntax and semantics, since the same semantic proposition can be expressed in both the active and passive voice. By using MEG data, we were able to determine the information flow in the brain during sentence reading, as well as during the post-sentence time period.

In order to draw scientific conclusions from these data, we require advanced analysis techniques. In the service of our neuroscientific goal, we optimized two common analysis approaches for neural data, decoding and representational similarity analysis (RSA). Thus the contributions of this thesis are two-fold: we have contributed methodologically by improving existing analysis approaches for MEG data. We have then used these approaches to gain insight into sentence comprehension in the brain.

We determined that the best way to improve performance for both decoding and RSA is to combine data from multiple subjects. The standard decoding approach is to classify using data from only a single subject, and then average the resulting decoding accuracies across subjects. Similarly, RSA is typically applied per-subject, and with correlations then averaged over subjects. In the case of decoding, by concatenating data from multiple subjects as additional features, we were able to improve performance significantly, despite having the same number of data samples. This of course was contingent upon using classifiers with regularization. In the case of RSA, we used the average representational dissimilarity matrix (RDM) over subjects for comparison with each model's RDM. The correlation was significantly higher for all models. However, this multi-subject RSA approach required us to develop a novel method for computing the noise ceiling that splits data over trials (as opposed to subjects) to assess data repeatability.

Our improved approaches to decoding and to RSA provide consistent and converging evidence that after a sentence is read, relevant information is present in the neural signal. Neural activity post-sentence is dominated first by syntactic information such as sentence voice and length. Then, semantic information *integrated with* syntactic information is detectable from the neural signal. Specifically, we can determine the sentence proposition ("Who did what to whom?") independent of syntactic voice. This constitutes the first such characterization of the information content of post-sentence neural activity using the temporal resolution of MEG data.

Methodological Contributions

Traditional decoding approaches that rely on averaging accuracy timeseries from single-subject models lose a great deal of signal due to temporal misalignment across subjects. By combining data from multiple subjects into a large feature vector, we are able to greatly boost accuracy. We can still retain the power to perform population-level inference by cross-validating over subjects and examining the fraction of folds where above chance accuracy is achieved. Additionally we demonstrated that regularization and test set SNR are the most important factors in determining classifier performance.

We contributed a novel approach to noise ceiling computation for RSA so that we can make use of data from multiple subjects to boost SNR. We additionally showed how the noise ceiling height can be used to optimize the creation of MEG RDMs. Lastly, we recommend the use of partial correlations to account for model confounds, although none of the models explored in this work suffer from such a problem.

Neuroscientific Contributions

To our knowledge, this work constitutes the first ever demonstration of the decoding stimuli while they are not currently visually present on the screen, without asking the subject explicitly to hold the stimuli in memory. We demonstrate that during sentence reading, each content word (noun or verb) in a sentence is decodable as it is presented. Furthermore, we show that determiners are uniquely represented in the brain and can be decoded with high accuracy.

We decode sentence voice and sentence length from whole-brain post-sentence activity with near-perfect accuracy. Both in our pilot and our confirmation data set, syntactic and structural sentence information was highly decodable from the MEG signal, and this result was additionally shown using RSA on our confirmation data set, demonstrating that a model of sentence similarity based on syntax correlated significantly with neural activity.

The neural activity during this post-sentence time period also contains the information needed for sentence integration, such as the identities of the agent, patient, and verb of the sentence, which can be reliably decoded above chance. Furthermore, a model of sentence similarity that uses both the semantic word content and the syntactic structure correlates significantly with the MEG data post-sentence. This model computes the similarity between two sentences by comparing their respective agents, verbs, and patients, thus correctly understanding that sentences with different voices can still convey the same meaning.

A Refined Theory of Sentence Processing

In Chapter 2 we explored the existing literature on sentence processing in the brain, finding results that supported the idea that processing a passive voice sentence involves *post hoc* reanalysis, in which the roles of agent and patient are swapped, in order for the proposition to be fully understood. Left pars opercularis (Broca's area) supposedly supports the syntactic aspect of the computation [1, 12, 13, 16, 18, 23, 25], whereas superior temporal lobe is implicated in semantic integration [12, 30, 34].

Past work has only demonstrated that there exists a neural activity difference between active and passive sentences during the post-sentence time period, which we confirmed by demonstrating that sentence voice (active or passive) is highly decodable from the post-sentence signal. In this thesis we sought to determine the information content of the reanalysis signal, specifically, what computations the brain is performing in passive sentences but not active sentences that would lead to the observable signal difference. To that end, we decoded content words (agent, verb, patient) from active and passive sentences to search for informational differences.

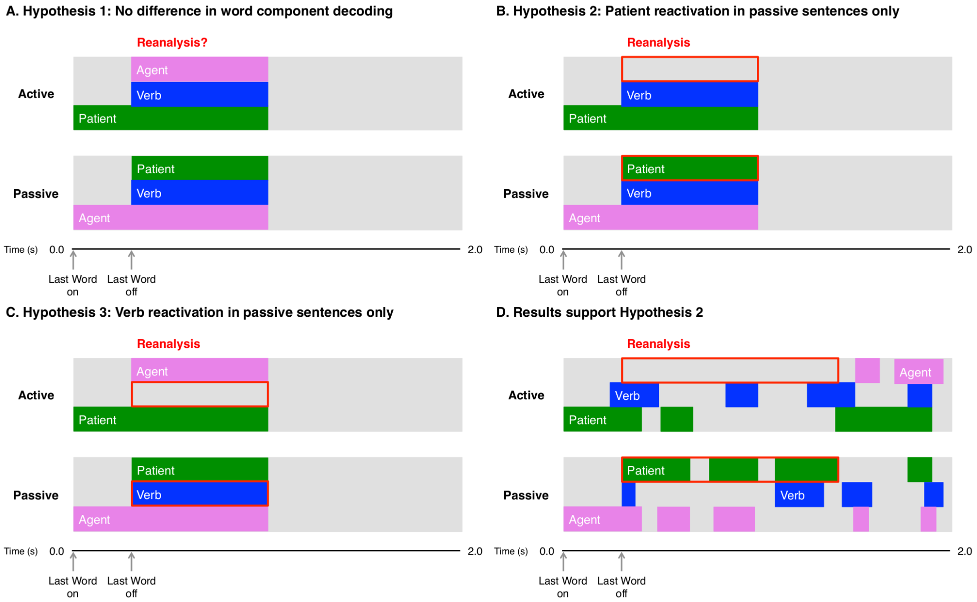


Figure 7.1: **Computations underlying the reanalysis signal.** Diagrams demonstrating the expected decoding result under several hypotheses for reanalysis, as well as the true decoding result. Colored blocks indicate significantly above-chance decoding. **A. Hypothesis 1: No difference in post-sentence word decoding.** Under this hypothesis, there is no difference between active and passive sentences in terms of which of the constituent sentence words can be decoded. It is unclear what the reanalysis signal can be attributed to in this case aside from some purely syntactic function. **B. Hypothesis 2: Patient reactivation in passive sentences only.** Under this hypothesis, part of the reanalysis signal can be explained by the fact that the first noun in passive sentences must switch its role from agent to patient, leading to a reactivation. Since no reactivation is necessary in active sentences, we will not see any decodability of the first noun in that case. **C. Hypothesis 3: Verb reactivation in passive sentences only.** Under this hypothesis, it is not the noun's role that must be switched, but rather the argument bindings of the verb, leading to selective verb reactivation. **D. Results.** Results show that the patient is reactivated in passive sentences, while the agent is not reactivated in active sentences, conforming to Hypothesis 2 (panel B).

Figure 7.1 shows three candidate hypotheses regarding reanalysis, as well as a summary of our true result. In the first hypothesis (Fig. 7.1A), there is no difference between active and passive sentences in terms of which aspects of the sentence can be decoded. Under that circumstance, we would see a similar decodability profile for both types of sentences across classification tasks, and we would attribute the reanalysis signal to working memory or some purely syntactic process. An alternative hypothesis, shown in Fig. 7.1B, is that first noun information (the patient, for passive voice sentences) is selectively reactivated in passive-voice sentences. This theory corresponds best with what would be predicted by the literature, which points to separate storage of agent and patient [30, 34]. In this case, we should be able to decode the patient from passive-voice sentences but we should not be able to decode the agent in active voice sentences (because there is no need for it to be reactivated). A third hypothesis (Fig. 7.1C) centers on the verb, which may require selective reactivation so that the argument binding can be reversed. Under this hypothesis we would selectively decode the verb in passive voice sentences. These three hypotheses are only a few of many potential explanations for reanalysis. Our results, shown in Fig. 7.1D, confirm the second hypothesis. We observed selective decodability of the first noun in passive voice sentences, thus indicating that the reanalysis signal can be attributed in part to a reactivation of the first noun (the patient) in passive voice sentences.

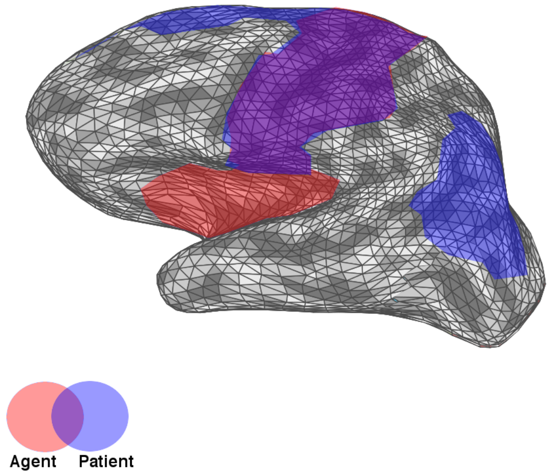
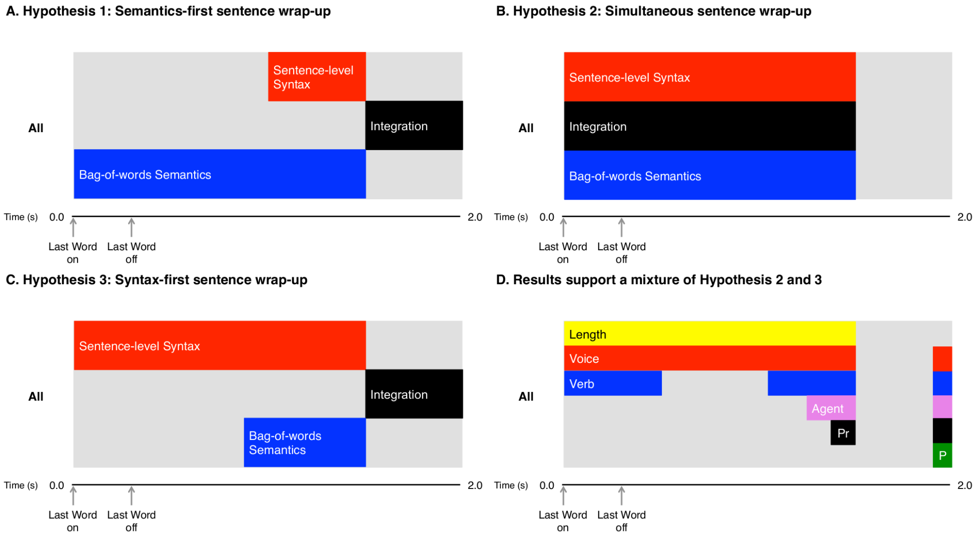


Figure 7.2: **Regions that maximally decode agent and patient identity.** During the post-sentence time period, we decoded agent and patient identity from the pool of all sentences (active and passive) for each ROI and each time point. In red are the regions for which agent decoding was maximal over the time period, and in blue are the regions where the patient was most decodable, with purple indicating overlap. Accuracy was thresholded at 0.8 for inclusion in this map.

Recall the claims from the fMRI literature that agent and patient are stored in proximal but distinct regions of the brain, specifically in inferior frontal cortex [30, 34]. Our whole-brain decoding results regarding reanalysis are consistent with what one would expect under this theory (namely that the patient must be reactivated in passive voice sentences in order to be "transferred" to its true location). We can further investigate the truth of this claim using our source-localized decoding results. Figure 7.2 shows the regions of maximal decoding accuracy (of accuracy at least 0.8) for decoding agent and patient identity from the post-sentence time period, using the pool of all sentences (active and passive). Note that there is large overlap in parietal and inferior frontal cortex, while there are also a few completely separate regions (e.g. the parieto-occiptal junction). While source-localized MEG data lacks the resolution to truly verify the existing fMRI results, these findings are largely consistent with the idea that there are proximal regions encoding agent and patient identity, with some separation.


A. Hypothesis 1: Semantics-first sentence wrap-up

B. Hypothesis 2: Simultaneous sentence wrap-up

C. Hypothesis 3: Syntax-first sentence wrap-up

D. Results support a mixture of Hypothesis 2 and 3

Figure 7.3: **Post-sentence wrap-up.** Diagrams demonstrating the expected decoding result under several hypotheses of post-sentence wrap-up, as well as the true decoding result. Colored blocks indicate significantly above-chance decoding. **A. Hypothesis 1: Semantics-first sentence wrap-up.** Under this hypothesis, the a bag-of-words representation of the sentence is detectable (e.g. the component words), followed by the syntax and then the integration of syntax and semantics to get the proposition. **B. Hypothesis 2: Simultaneous sentence wrap-up.** Under this hypothesis, syntax and semantics are integrated simultaneously to generate the proposition. **C. Hypothesis 3: Syntax-first sentence wrap-up.** Under this hypothesis, syntax information is processed first, followed by semantic and then the final integration. **D. Results.** Results support a mixture of Hypothesis 2 and 3. Sentence voice and length (syntactic properties) are decodable first, along with the verb identity. Then, the proposition and agent become decodable as well. At the very end of the post-sentence time period, all sentence attributes are decodable.

Decoding and RSA provided complementary and converging evidence in support of the existence of a post-sentence wrap-up period. Through both methods we were able to determine that sentence voice and length were encoded in the neural signal post-sentence. With decoding, we determined this because voice and length could be decoded with high accuracy, while with RSA we showed that a model of sentence similarity based on syntax and length correlated strongly with neural activity.

We can situate these results in terms of a classic psycholinguistic debate: syntax or semantics first. One could theorize, as shown in Fig. 7.3A, a semantics-first sentence wrap-up period, in which the constituent words of a sentence are held in memory in a bag-of-words fashion and then combined syntactically so that full integration can take place. In that case we would expect to see that content words are decodable prior to syntactic properties such as voice or sentence length, ultimately followed by proposition identity decoding. An alternative, shown in Fig. 7.3B, is that all processing occurs simultaneously, with every attribute decodable during the same post-sentence window. Yet another alternative is a syntax-first model, shown in Fig. 7.3C, in which syntactic attributes are first decodable, followed by semantics and integration.

Both our RSA and decoding results (decoding results summarized in Fig. 7.3D), support a mixture of the last two hypotheses, with a largely syntax-first presentation but with the verb decodable throughout the post-sentence time period. Our RSA model of sentence meaning that incorporated both syntax and semantics, i.e. our hierarchical model, correlated significantly with the neural activity *after* correlation between the neural activity and the syntax model began to decline. Similarly, after sentence voice and length become less decodable, we can decode the sentence proposition and the other pieces of that proposition (e.g. the agent and the patient).

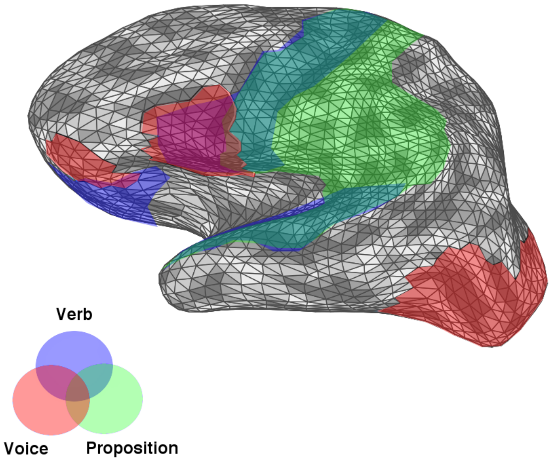


Figure 7.4: **Regions that maximally decode verb, voice and proposition identity.** During the post-sentence time period, we decoded voice, verb, and proposition identity from the pool of all sentences (active and passive) for each ROI and each time point. In red are the regions for which voice decoding was maximal over the time period, and in blue are the regions where the verb was most decodable. Green represents maximal proposition decoding. Accuracy was thresholded at 0.8 for inclusion in this map.

Which regions underlie our ability to decode the voice, the proposition, and the verb of a sentence? We can again examine our source-localized decoding results, summarized in Fig. 7.4. The regions that maximally decoded voice included the left pars opercularis (Broca's area), as would be expected from the literature [1, 12, 13, 16, 18, 23, 25], as well as occipital cortex and frontal cortex. The regions that maximally decoded the proposition include superior temporal cortex (as expected [12, 30, 34]) as well as parietal cortex. Importantly, the verb is decodable in what seems to be the union of these regions, with inferior frontal cortex, Broca's area, parietal cortex, and superior temporal cortex all containing information for decoding. This result implicates verb processing as crucial for the unification of syntactic and semantic information and the formation of the integrated proposition.

Future work

The scientific goal of this thesis was to understand how varied syntax could be reconciled during sentence comprehension. While active and passive sentences are a convenient set of syntactic structures to use (since the same proposition can be expressed in both voices), the logical extension of this work would be to explore a wider array of syntactic structures. This in turn can allow us to test a more complex set of models of sentence composition, beyond the simple syntax, semantic, and hierarchical models explored in Chapter 6.

Additionally, sentences are not typically read in a vacuum, without context, but rather as part of a larger narrative. An important next step for the neuroscience of language is to explore sentences in a natural context and determine whether the post-sentence wrap up period observed here is similarly detectable.

Appendix A

Data Collection and Preprocessing Details

Elekta Neuromag. 102 locations, each location has 3 sensors: a magnetometer that detects magnetic field, and two orthoganal planar gradiometers to detect magnetic field spatial gradient.

The data were spatially filtered using the temporal extension of the signal space separation (tSSS) algorithm, lowpass filtered to 150Hz with notch filters applied at 60 and 120Hz, and downsampled to 500Hz. Artifacts from tSSS-filtered same-day empty room measurements, ocular and cardiac artifacts were removed via signal space projection (SSP).

Appendix B

Pilot Stimuli

Active				
Det-1	Noun-1	Verb	Det-2	Noun-2
A	dog	found	the	peach
The	dog	kicked	a	school
A	dog	inspected	a	hammer
The	dog	touched	the	door
The	doctor	found	a	school
A	doctor	kicked	the	peach
A	doctor	inspected	a	door
The	doctor	touched	the	hammer
The	student	found	a	door
A	student	kicked	the	hammer
The	student	inspected	the	school
A	student	touched	a	peach
A	monkey	found	the	hammer
The	monkey	kicked	a	door
The	monkey	inspected	the	peach
A	monkey	touched	a	school

Table B.1: Active sentences used in the experiment.

Passive						
Det-1	Noun-1	Verb-Aux	Verb	Prep	Det-2	Noun-2
The	peach	was	found	by	a	dog
A	school	was	kicked	by	the	dog
A	hammer	was	inspected	by	a	dog
The	door	was	touched	by	the	dog
A	school	was	found	by	the	doctor
The	peach	was	kicked	by	a	doctor
A	door	was	inspected	by	a	doctor
The	hammer	was	touched	by	the	doctor
A	door	was	found	by	the	student
The	hammer	was	kicked	by	a	student
The	school	was	inspected	by	the	student
A	peach	was	touched	by	a	student
The	hammer	was	found	by	a	monkey
A	door	was	kicked	by	the	monkey
The	peach	was	inspected	by	the	monkey
A	school	was	touched	by	a	monkey

Table B.2: Passive sentences used in the experiment.

Engagement Question

Question	Answer
Was there a vegetable?	N
Could you get hurt doing this?	Y
Was a tool seen?	Y
Was it bouncy?	N
Could this item be put in a pocket?	N
Was fruit damaged?	Y
Was it a hard surface?	Y
Was it bendy?	N
Was this item bigger than a whale?	N
Was it a soft item?	N
Did this involve a building?	Y
Was there something fuzzy?	Y
Was this item smaller than an elephant?	Y
Is this something you could do?	Y
Was something blue seen?	N
Was this item squishy?	N

Table B.3: Comprehension questions used in the Pilot experiment.

Appendix C

Confirmatory Stimuli

The stimuli are listed in Tables C.1 and C.2.

Active						
Art-1	Noun-1	Verb	Det-2	Noun-2	Question	Answer
The	man	kicked	the	girl.	Did he see someone?	Y
The	girl	helped	the	boy.	Did she do nothing?	N
The	woman	approached	the	man.	Was he seen?	N
The	boy	punched	the	woman.	Was she attacked?	Y
The	man	kicked.			Was he sleeping?	N
The	girl	helped.			Did she act?	Y
The	woman	approached.			Did she move?	Y
The	boy	punched.			Was he still?	N
The	girl	kicked	the	man.	Did she behave nicely?	N
The	boy	helped	the	girl.	Did he do something?	Y
The	man	approached	the	woman.	Was she visible?	Y
The	woman	punched	the	boy.	Was he safe?	N
The	girl	kicked.			Was she sleeping?	N
The	boy	helped.			Did he act?	Y
The	man	approached.			Did he move?	Y
The	woman	punched.			Was she sleeping?	N

Table C.1: Active voiced Stimuli

Passive								
Det-1	Noun-1	Verb-Aux	Verb	Prep	Det-2	Noun-2	Question	Answer
The	girl	was	kicked	by	the	man.	Did he see someone?	Y
The	boy	was	helped	by	the	girl.	Did she do nothing?	N
The	man	was	approached	by	the	woman.	Was he seen?	N
The	woman	was	punched	by	the	boy.	Was she attacked?	Y
The	girl	was	kicked.				Was she hurt?	Y
The	boy	was	helped.				Was he ignored?	N
The	man	was	approached.				Was he visible?	Y
The	woman	was	punched.				Was she unharmed?	N
The	man	was	kicked	by	the	girl.	Did she behave nicely?	N
The	girl	was	helped	by	the	boy.	Did he do something?	Y
The	woman	was	approached	by	the	man.	Was she visible?	Y
The	boy	was	punched	by	the	woman.	Was he safe?	N
The	man	was	kicked.				Was he hurt?	Y
The	girl	was	helped.				Was she ignored?	N
The	woman	was	approached.				Was she visible?	Y
The	boy	was	punched.				Was he unharmed?	N

Table C.2: Passive voiced stimuli

Bibliography

[1] Angela D Friederici. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392, 2011. (document), 2.1, 2.2.1, 2.2.1, 7, 7

[2] Stefan Haufe, Frank Meinecke, Kai GÃűrgen, Sven DÃd’hne, John-Dylan Haynes, Benjamin Blankertz, and Felix BieÃ§mann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96 – 110, 2014. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2013.10.067. URL `http://www.sciencedirect.com/science/article/pii/S1053811913010914`. (document), 4.1.2, 5.3.1, 5.3.2, 5.3, 5.4, 5.8, 6.3.1, 6.3.1, 6.4, 6.5, 6.8

[3] Shelagh Brumfitt. Losing your sense of self: What aphasia can do. *Aphasiology*, 7(6): 569–575, 1993. 1

[4] Peter Hansen, Morten Kringelbach, and Riitta Salmelin. *MEG: An introduction to methods*. Oxford university press, 2010. 1

[5] Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. Tracking Neural Coding of Perceptual and Semantic Features of Concrete Nouns. *NeuroImage*, 62(1):451–463, 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2012.04.048. URL `http://linkinghub.elsevier.com/retrieve/pii/S1053811912004442`. 1, 2.2.1, 2.3.1, 3.1.1

[6] Joachim Gross, Sylvain Baillet, Gareth R Barnes, Richard N Henson, Arjan Hillebrand, Ole Jensen, Karim Jerbi, Vladimir Litvak, Burkhard Maess, Robert Oostenveld, et al. Good practice for conducting and reporting meg research. *Neuroimage*, 65:349–363, 2013. 1, 3.1.1

[7] Nikolaus Kriegeskorte and Rogier A Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, 17(8):401–412, 2013. 1, 2.3, 2.3.2, 3, 3.2.1, 6.2.3

[8] Tijl Grootswagers, Susan G. Wardle, and Thomas A. Carlson. Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, 29(4):677–697, 2017. doi: 10.1162/jocn_a_01068. URL `http://dx.doi.org/10.1162/jocn_a_01068`. PMID: 27779910. 1, 2.3.1, 2.3.2, 3.1.1, 3.1.3, 3.1.3, 3.1.4, 3.1.4, 3.2.1

[9] Shevaun Lewis and Colin Phillips. Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*, 44(1):27–46, 2015. 2.2

[10] Friedemann Pulvermüller. Brain embodiment of syntax and grammar: Discrete combinato-

rial mechanisms spelt out in neuronal circuits. *Brain and language*, 112(3):167–179, 2010. 2.2

[11] Noam Chomsky. Problems of projection. *Lingua*, 130:33–49, 2013. 2.2

[12] Peter Hagoort. Muc (memory, unification, control) and beyond. *Frontiers in psychology*, 4: 416, 2013. 2.2, 2.2.1, 2.2.2, 7, 7

[13] Peter Hagoort and Peter Indefrey. The neurobiology of language beyond single words. *Annual Review of Neuroscience*, 37:347–362, 2014. 2.2, 2.2.1, 7, 7

[14] Angela D Friederici and Wolf Singer. Grounding language processing on basic neurophysiological principles. *Trends in cognitive sciences*, 19(6):329–338, 2015. 2.2, 2.2.1

[15] Stanislas Dehaene, Florent Meyniel, Catherine Wacongne, Liping Wang, and Christophe Pallier. The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron*, 88(1):2–19, 2015. 2.2

[16] Yosef Grodzinsky and Angela D Friederici. Neuroimaging of syntax and syntactic processing. *Current opinion in neurobiology*, 16(2):240–246, 2006. 2.2.1, 7, 7

[17] Cathy J. Price. A review and synthesis of the first 20years of pet and fmri studies of heard speech, spoken language and reading. *NeuroImage*, 62(2):816 – 847, 2012. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2012.04.062. URL `http://www.sciencedirect.com/science/article/pii/S1053811912004703`. 20 YEARS OF fMRI. 2.2.1

[18] Stefano F Cappa. Imaging semantics and syntax. *Neuroimage*, 61(2):427–431, 2012. 2.2.1, 7, 7

[19] Jonathan Brennan, Yuval Nir, Uri Hasson, Rafael Malach, David J Heeger, and Liina Pylkkänen. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and language*, 120(2):163–173, 2012. 2.2.1

[20] Corianne Rogalsky and Gregory Hickok. Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. *Cerebral Cortex*, 19(4):786–796, 2008. 2.2.1

[21] Nai Ding, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, 19(1):158, 2016. 2.2.1

[22] Idan Blank, Zuzanna Balewski, Kyle Mahowald, and Evelina Fedorenko. Syntactic processing is distributed across the language system. *NeuroImage*, 127:307–323, 2016. 2.2.1

[23] Emiliano Zaccarella and Angela D Friederici. The neurobiological nature of syntactic hierarchies. *Neuroscience & Biobehavioral Reviews*, 81:205–212, 2017. 2.2.1, 2.2.2, 7, 7

[24] Michiru Makuuchi, Jörg Bahlmann, Alfred Anwander, and Angela D Friederici. Segregating the core computational faculty of human language from working memory. *Proceedings of the National Academy of Sciences*, 106(20):8362–8367, 2009. 2.2.1

[25] Matthew J Nelson, Imen El Karoui, Kristof Giber, Xiaofang Yang, Laurent Cohen, Hilda Koopman, Sydney S Cash, Lionel Naccache, John T Hale, Christophe Pallier, et al. Neuro-

physiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, page 201701590, 2017. 2.2.1, 2.2.2, 3.1.4, 5, 5.3.2, 6.1.2, 7, 7

[26] Mirella Dapretto and Susan Y Bookheimer. Form and content: dissociating syntax and semantics in sentence comprehension. *Neuron*, 24(2):427–432, 1999. 2.2.2

[27] Ina Bornkessel, Matthias Schlesewsky, and Angela D Friederici. Eliciting thematic reanalysis effects: The role of syntax-independent information during parsing. *Language and Cognitive Processes*, 18(3):269–298, jun 2003. ISSN 0169-0965. doi: 10.1080/01690960244000018. 2.2.2, 5, 5.3.2, 5.3.2

[28] Lars Meyer, Jonas Obleser, Stefan J Kiebel, and Angela D Friederici. Spatiotemporal dynamics of argument retrieval and reordering: an FMRI and EEG study on sentence processing. *Frontiers in psychology*, 3(December):523, jan 2012. ISSN 1664-1078. doi: 10.3389/fpsyg.2012.00523. 2.2.2, 5, 5.3.2, 5.3.2

[29] Jennifer E Mack, Aya Meltzer-Asscher, Elena Barbieri, and Cynthia K Thompson. Neural correlates of processing passive sentences. *Brain sciences*, 3(3):1198–1214, 2013. 2.2.2

[30] Masako Hirotani, Michiru Makuuchi, Shirley-Ann Rüschemeyer, and Angela D Friederici. Who was the agent? the neural correlates of reanalysis processes during sentence comprehension. *Human brain mapping*, 32(11):1775–1787, 2011. 2.2.2, 5.3.2, 7, 7, 7, 7

[31] Peter Hagoort. Interplay between syntax and semantics during sentence comprehension: Erp effects of combining syntactic and semantic violations. *Journal of cognitive neuroscience*, 15(6):883–899, 2003. 2.2.2

[32] Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9): 424–430, 2006. 2.2.2, 2.3, 2.3.1, 3

[33] Jing Wang, Vladimir L Cherkassky, Ying Yang, Kai-min Kevin Chang, Robert Vargas, Nicholas Diana, and Marcel Adam Just. Identifying thematic roles from neural representations measured by functional magnetic resonance imaging. *Cognitive neuropsychology*, 33 (3-4):257–264, 2016. 2.2.2

[34] Steven M Frankland and Joshua D Greene. An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 112(37), aug 2015. ISSN 1091-6490. doi: 10.1073/pnas.1421236112. URL `http://www.pnas.org/content/112/37/11732.abstract`. 2.2.2, 5.3.2, 7, 7, 7, 7

[35] Andrew James Anderson, Edmund C Lalor, Feng Lin, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Rajeev DS Raizada, Scott Grimm, and Xixi Wang. Multiple regions of a cortical network commonly encode the meaning of words in multiple grammatical positions of read sentences. *Cerebral Cortex*, 2018. 2.2.2

[36] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, 45(1):S199–S209, 2009. 2.3, 2.3.1

[37] Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L

Malave, Robert A Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008. 2.3.1

[38] Sebastian Weichwald, Timm Meyer, Ozan Özdenizci, Bernhard Schölkopf, Tonio Ball, and Moritz Grosse-Wentrup. Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110:48–59, 2015. 2.3.1, 4.1.2

[39] J. R. King and S. Dehaene. Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210, 2014. ISSN 1879307X. doi: 10.1016/j.tics.2014.01.002. URL `http://dx.doi.org/10.1016/j.tics.2014.01.002`. 2.3.1, 3.1.1, 3.1.4, 3.1.4, 5.2.2

[40] Jean Dickinson Gibbons and Subhabrata Chakraborti. Nonparametric statistical inference. In *International encyclopedia of statistical science*, pages 977–979. Springer, 2011. 2.3.1, 2.3.2

[41] Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25, 2002. 2.3.1

[42] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995. 2.3.1, 5.2.2, 6.2.2, 6.2.3

[43] Qiong Zhang, Jelmer P Borst, Robert E Kass, and John R Anderson. Inter-subject alignment of meg datasets in a common representational space. *Human brain mapping*, 38(9):4287–4301, 2017. 2.3.1

[44] Hao Xu, Alexander Lorbert, Peter J Ramadge, J Swaroop Guntupalli, and James V Haxby. Regularized hyperalignment of multi-set fmri data. In *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, pages 229–232. IEEE, 2012. 2.3.1, 3.1.1, 3.1.2

[45] Alona Fyshe, Gustavo Sudre, Leila Wehbe, Nicole Rafidi, and Tom M Mitchell. The semantics of adjective noun phrases in the human brain. *bioRxiv*, 2016. doi: 10.1101/089615. URL `http://biorxiv.org/content/early/2016/11/25/089615`. 2.3.1, 3.1.1

[46] Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, Doha, Qatar, October 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D14-1030`. 2.3.1, 3.1.1

[47] Brian Murphy, Leila Wehbe, and Alona Fyshe. Decoding language from the brain. *Language, Cognition, and Computational Models*, page 53, 2018. 2.3.1

[48] Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220, 1967. 2.3.2, 6.2.3

[49] Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553, 2014. 2.3.2

[50] Alex Clarke and Lorraine Tyler. Object-specific semantic coding in human perirhinal cor-

tex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 34: 4766–75, 04 2014. 2.3.2

[51] Jens HjortkjÃęr, Tanja Kassuba, Kristoffer Madsen, Martin Skov, and Hartwig Siebner. Task-modulated category representations of natural sound sources in human cortex. *Cerebral Cortex*, 11 2017. 2.3.2

[52] Susan G Wardle, Nikolaus Kriegeskorte, Tijl Grootswagers, Seyed-Mahdi Khaligh-Razavi, and Thomas A Carlson. Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with meg. *NeuroImage*, 132:59–70, 2016. 2.3.2, 3.2.1

[53] Gaël Varoquaux, Pradeep Reddy Raamana, Denis A Engemann, Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, 2017. 3.1.1, 3.1.6

[54] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 3.1.1

[55] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014. 3.1.1

[56] James J Higgins. *Introduction to modern nonparametric statistics*. Cengage Learning, 2003. 3.1.2

[57] Ryszard S Michalski, Jaime G Carbonell, and Tom M Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013. 3.1.3

[58] Harry Nyquist. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2):617–644, 1928. 3.1.4

[59] Marek Kuczma. *An introduction to the theory of functional equations and inequalities: Cauchy's equation and Jensen's inequality*. Springer Science & Business Media, 2009. 3.1.5

[60] Etienne Combrisson and Karim Jerbi. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of neuroscience methods*, 250:126–136, 2015. 3.1.6

[61] Jerome L Myers, Arnold D Well, and Robert F Lorch Jr. *Research design and statistical analysis*. Routledge, 2013. 3.2.4

[62] Sewall Wright. The method of path coefficients. *The annals of mathematical statistics*, 5 (3):161–215, 1934. 3.2.4

[63] William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011. 4

[64] Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535, 2009. 4.1

[65] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning, 2001. 4.1

[66] Joseph Henrich, Steven J Heine, and Ara Norenzayan. Most people are not weird. *Nature*, 466(7302):29, 2010. 4.2.1

[67] Patrick Dupont, Guy A Orban, Rufin Vogels, Guy Bormans, Johan Nuyts, Christiaan Schiepers, Michael De Roo, and Luc Mortelmans. Different perceptual tasks performed with the same visual stimulus attribute activate different regions of the human brain: a positron emission tomography study. *Proceedings of the National Academy of Sciences*, 90 (23):10927–10931, 1993. 4.2.1

[68] Gordon L Shulman, Maurizio Corbetta, Randy L Buckner, Marcus E Raichle, Julie A Fiez, Francis M Miezin, and Steven E Petersen. Top-down modulation of early sensory cortex. *Cerebral cortex (New York, NY: 1991)*, 7(3):193–206, 1997. 4.2.1

[69] Wim Fias, Patrick Dupont, Bert Reynvoet, and Guy A. Orban. The quantitative nature of a visual task differentiates between ventral and dorsal stream. *Journal of Cognitive Neuroscience*, 14(4):646–658, 2002. doi: 10.1162/08989290260045873. URL https://doi.org/10.1162/08989290260045873. 4.2.1

[70] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. ACM, 2010. 4.2.1

[71] Monya Baker. Reproducibility crisis? *Nature*, 533:26, 2016. 4.3

[72] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365, 2013. 4.3, 4.3

[73] Open Science Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015. 4.3

[74] Simona Spinu, Lucia-Elena Enciu, and Radu Mutihac. Confirmatory versus exploratory statistical analysis of functional brain imaging data. *ROMANIAN JOURNAL OF PHYSICS*, 61(7-8):1299–1311, 2016. 3

[75] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1:0021, 2017. 4.3, 4.3

[76] Craig M Bennett, MB Miller, and GL Wolford. Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: an argument for multiple comparisons correction. *Neuroimage*, 47(Suppl 1):S125, 2009. 4.3

[77] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162. 6.2.3

[78] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393, 2007.

[79] Nicole S. Rafidi, Erika J.C. Laing, and Tom Mitchell. The role of syntax in semantic processing: a study of active and passive sentences. *Oral Session at the Organization for Human Brain Mapping Annual Meeting, June 2015*, 2015.

[80] Seyed-Mahdi Khaligh-Razavi, Linda Henriksson, Kendrick Kay, and Nikolaus Kriegeskorte. Fixed versus mixed rsa: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *Journal of mathematical psychology*, 76:184–197, 2017.

[81] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[82] Jerome T Connor, R Douglas Martin, and Les E Atlas. Recurrent neural networks and robust time series prediction. *IEEE transactions on neural networks*, 5(2):240–254, 1994.

[83] Samu Taulu and Riitta Hari. Removal of magnetoencephalographic artifacts with temporal signal-space separation: demonstration with single-trial auditory-evoked responses. *Human brain mapping*, 30(5):1524–1534, 2009.

[84] Mikko A Uusitalo and Risto J Ilmoniemi. Signal-space projection method for separating meg or eeg into components. *Medical and Biological Engineering and Computing*, 35(2): 135–140, 1997.

[85] Matti S Hämäläinen and Risto J Ilmoniemi. Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & biological engineering & computing*, 32(1):35–42, 1994. 6.2.1

[86] Rebecca A Hutchinson, Radu Stefan Niculescu, Timothy A Keller, Indrayana Rustandi, and Tom M Mitchell. Modeling fmri data generated by overlapping cognitive processes with unknown onsets using hidden process models. *NeuroImage*, 46(1):87–104, 2009.

[87] Marco Congedo, Louis Korczowski, Arnaud Delorme, et al. Spatio-temporal common pattern: A companion method for erp analysis in the time domain. *Journal of neuroscience methods*, 267:74–88, 2016.

[88] Omid Sayadi and Mohammad Bagher Shamsollahi. Ecg denoising and compression using a modified extended kalman filter structure. *IEEE Transactions on Biomedical Engineering*, 55(9):2240–2248, 2008.

[89] John R Anderson. Tracking problem solving by multivariate pattern analysis and hidden markov model algorithms. *Neuropsychologia*, 50(4):487–498, 2012.

[90] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.

[91] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.