Bridging Language in Machines with Language in the Brain

Mariya Toneva

July 2021
CMU-ML-21-106

Joint Program in Machine Learning and Neural Computation
School of Computer Science; Neuroscience Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Leila Wehbe, Co-chair
Tom Mitchell, Co-chair
Michael Tarr
Chris Dyer (DeepMind)
Tal Linzen (New York University)

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

This material is based upon work supported by NIH U01NS098969, NIH TDA022762B, NIH RHD075328B, NSF DGE1252522, NSF DGE1745016, NSF DGE1745016, AFRL FA95501710218, AFRL FA865013C7360, and AFOSR FA95502010118. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the United States Air Force, the National Institutes of Health, or the National Science Foundation.

Keywords: Artificial Intelligence, Machine Learning, Deep Learning, Natural Language Processing, Interpretability, Cognitive Neuroscience, Computational Neuroscience, fMRI, MEG, Naturalistic Experiments, Encoding Models, Computational Controls

Abstract

Several major innovations in artificial intelligence (AI) (e.g. convolutional neural networks, experience replay) are based on findings about the brain. However, the underlying brain findings took many years to first consolidate and many more to transfer to AI. Moreover, these findings were made using invasive methods in non-human species. For brain functions that are uniquely human, such as understanding complex language, there is no suitable animal that can serve as a model organism and thus a mechanistic understanding is that much farther away.

In this dissertation, we present a data-driven framework that circumvents these limitations by establishing a direct connection between brain recordings of people comprehending language and natural language processing (NLP) computer systems. We present evidence that this connection can be beneficial for both neurolinguistics and NLP. Specifically, we show that this framework can utilize recent successes in neural networks for NLP to enable scientific discovery about context- and task-dependent meaning composition in the brain, and we present the first evidence that brain activity measurements of people reading can be used to improve the generalization performance of a popular deep neural network language model. These investigations also contribute advances in cognitive modeling that may be useful beyond the study of language. In short, this dissertation involves multidisciplinary investigations and makes contributions to cognitive neuroscience, neurolinguistics, and natural language processing.

Acknowledgments

I am indebted to the many people who have contributed to this thesis directly and indirectly. Firstly, this thesis would not have been possible without the support of my advisors Tom Mitchell and Leila Wehbe. I remember reading about Leila and Tom's work that used machine learning to study the brain processes that underlie story reading during my third year of college, which first inspired me to pursue research at the intersection of language in the brain and machines. I feel incredibly fortunate to have the opportunity to learn from both of them. Tom gave me the freedom to develop my own scientific interests and has been a paramount example of someone who can communicate ideas and findings clearly–a skill, which I strive to emulate. Leila was extremely patient and generous with her time in helping me grow as a researcher with many brilliant ideas and late-night writing sessions, and connected me with multiple collaborators who contributed to this thesis and beyond.

I am also grateful for the support of the rest of my thesis committee–Michael Tarr, Tal Linzen, and Chris Dyer. Their feedback has been helpful in shaping the main arguments in the thesis and has helped me better communicate our findings to multi-disciplinary audiences. I'd also like to thank Mike for his role in my ending up at CMU. During my bachelors, I had the opportunity to do a research internship in Mike's lab as part of the undergraduate program in neural computation. This internship played a huge part in my joining CMU as a PhD student one year later. A big thank you also to Elissa Aminoff and Ying Yang, who also mentored me during this undergraduate research opportunity, and to Brian Scassellati and Henny Admoni who first introduced me to research as a first year undergraduate student.

Many thanks also to Geoff Gordon who took a chance on a neuroscience student and gave me the opportunity to do a research internship at Microsoft Research Montreal. My time at MSR Montreal showed me that my scientific mindset can also be useful in core machine learning problems, and helped me pursue these new areas of research. A huge thank you to Alessandro Sordoni for being a supportive mentor during this internship and beyond. I would like to also thank a number of faculty and staff who have hugely contributed to my experience at CMU: Diane Stidle for her significant part in making MLD a supportive and welcoming place, Aaditya Ramdas and Ryan Tibshirani for being incredible teachers and thoughtful members of the community, Ameet Talwalker for helping start and support the ML@CMU blog, and Rob Kass for being an enthusiastic proponent of computational neuroscience and the joint programs between neural computation, statistics, and machine learning.

Much of the work during my PhD has been in collaboration with many talented researchers. I'd like to thank Otilia Stretcu for her significant contributions to Chapter 5 which provided a unified framework that hugely simplifies implementing and testing different computational hypotheses, and Anand Bollu, Jenn Williams, and Christoph Dann for their contributions to the work in Chapter 3, which was truly a team effort. Many thanks also to the Courtois NeuroMod group at the University of Montreal for collecting and sharing such high-quality fMRI recordings with us, which made the

work in Chapter 3 possible. I've also really enjoyed collaborating with Dan Schwartz, who is not only a great thinker, but also a fantastic friend and one of my favorite conference travel buddies. Conversations with Dan have influenced much of my research, including work that became the basis for Chapters 4 and 6.

Throughout moves to many new places, Pittsburgh was the first one to feel like home in large part due to the many amazing people there. Thank you to current and previous members of Tom and Leila's groups–Alona Fyshe, Nicole Rafidi, Tara Pirnia, Aniketh Reddy, Aria Wang, Nidhi Jain, Maggie Henderson, Ruogu Lin, and Srini Ravishankar–for the sense of solidarity that comes from working towards a common goal. To my close friends–Abu Saparov, Anthony Platanios, Avi Dubey, Christoph Dann, Dan Schwartz, Maruan Al-Shedivat, Mrinmaya Sachan, Otilia Stretcu, and Xun Zheng–I feel grateful to have shared the highs and lows of our PhDs together. To the many more amazing people I crossed paths with–Lacey Konopasek, Chenghui Zhou, Robin Schmucker, Ezra Winston, Tolani Olarinre, Renato Negrinho, Han Zhao, Simon Du, Kenny Marino, Adarsh Prasad, Arun Suggala, Jonathan Mei, Nick Blauch, Qiong Zhang, Kirstin Early, Ben Cowley, Calvin Murdock, Micol Marchetti-Bowick, Tim Hyde, Willie Neisswanger, Junier Oliva, Conor Igoe, Ian Char, Jake Tyo, and more–thank you for the many laughs throughout the years. Thank you to Pinky, who provided companionship to several generations of graduate students and lived well beyond the life expectancy of a Kissing Gourami.

Lastly, I would like to express my gratitude to my family. To my grandparents–Mariya, Stoyan, Tsvetana, and Georgi–thank you for teaching me the importance of education and of being connected to my roots. To the more recent additions to my family: Markus and Mahsa–thank you for your advice along the way, and Elfriede and Helmut–thank you for the warm welcome and for the delicious cake recipes that fueled much research progress. And to my parents–Ivanka and Krasimir–thank you for the countless sacrifices that included uprooting your lives twice for better opportunities for me. You've taught me, by example, to be resilient and to believe in myself in the face of uncertainty, and I would not be where I am today without these values. To my partner–Christoph–thank you for your patience and kindness, and for making the bad times less bad, and the good times much better. I look forward to our next chapter.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The human brain is a remarkable information processing system that is able to learn from few examples, retain previously learned facts and skills while learning new ones, and comprehend text of any length, in any language. It has long served as inspiration to the fields of artificial intelligence (AI) and machine learning (ML). For example, the discovery of cell receptive fields and the hierarchy of information processing in the early visual system (Hubel et al., 1968) led to the invention of convolutional neural networks (Fukushima et al., 1982), which revolutionized computer vision, and discovering that replaying prior experiences in the hippocampus aids memory consolidation (McNaughton, 1983) inspired the development of experience replay (McClelland et al., 1992) (see also (Lin, 1992)), which has become a staple in deep reinforcement learning after its role in DeepMind's AlphaGo's (Silver et al., 2016) first win over the human Go world champion. In these cases, the underlying finding about brain function took many years to first consolidate and many more to transfer to AI (e.g. the visual system findings took 10 years to consolidate and more than 10 additional years to transfer to AI). Moreover, these findings were made using invasive methods in non-human species. For brain functions that are uniquely human, such as understanding complex language, there is no suitable animal that can serve as a model organism and thus a mechanistic understanding is that much farther away.

In this dissertation, we present a data-driven framework that circumvents these limitations by establishing a direct connection between the brain and natural language processing (NLP) computer systems. We visualize the basis of this data-driven framework in Figure 1.1. We present evidence that this data-driven connection can be beneficial for both neurolinguistics and NLP. Specifically, we show that this framework can utilize recent successes in neural networks for NLP to enable scientific discovery about context- and task-dependent meaning composition in the brain, and present the first evidence that brain activity measurements of people reading can be used to improve the generalization performance of a popular deep neural network language model.

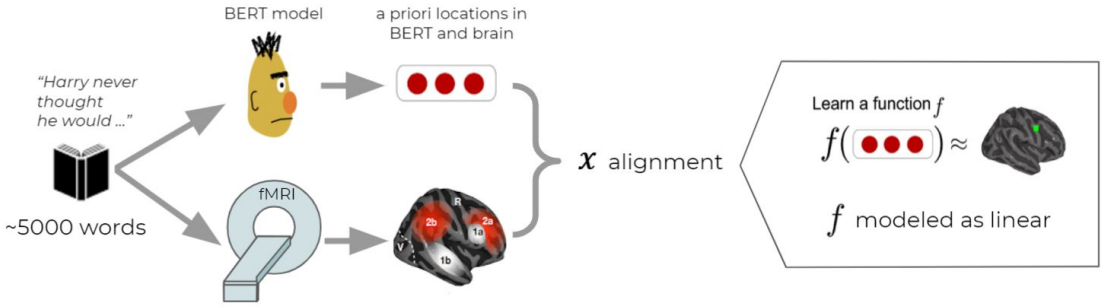


Figure 1.1: Visualization of our data-driven approach. We first present naturalistic text, such a chapter of a book, to a person word-by-word while her brain activity is recorded by a brain imaging device, such as fMRI. We present the same text as input to a natural language processing (NLP) computer system, such as BERT (Devlin et al., 2018). Next, we extract representations of this text from specific intermediate layers of the NLP system and we observe the brain recordings that are elicited by the same text. Lastly, we compute the alignment between these two representations of text–one from the NLP system and one from the brain recordings. This alignment between the brain recordings and the NLP system is estimated by learning a function f, which predicts the activity in each brain source (e.g. voxel, sensor-timepoint, etc.) as a function of the NLP system's representation of the presented text. This function is often modeled as a linear function and learned using standard machine learning techniques. Please refer to Chapter 3 for more details about this method.

NLP Systems as Model Organisms for Human Language Comprehension

When reading the sentence "The trophy doesn't fit into the brown suitcase because it's too big", we understand the meaning of this sentence despite of the ambiguous pronoun "it", which may refer to either the trophy or the suitcase (Levesque et al., 2012). We know that the referent is the trophy. If the sentence instead was "The trophy doesn't fit into the brown suitcase because it's too small", then we would have inferred that the referent is the suitcase. How does the brain process these sentences and attribute real-world meaning to them? To address this, there are some fundamental preliminary questions to answer about *what* information is processed *where* and *when* in the brain, in order to understand *how* this information is aggregated across different locations and time points.

Using neuroimaging devices that record human brain activity during language processing, neuroscientists have made progress towards answering the *what*, *where*, and *when* questions. For instance, researchers have found that the meaning of individual words is distributed across the cortex but consistent across different people (Mitchell, Shinkareva, et al., 2008; Wehbe, Murphy, et al., 2014; Huth, Heer, et al., 2016), that a certain set of brain regions termed the "language network" supports language comprehension (Fedorenko, Hsieh, et al., 2010; Fedorenko and Thompson-Schill, 2014), and that the meaning of a word is processed between 200 and 600ms after it is first

read (Salmelin, 2007; Skeide et al., 2016). However, *how* information is aggregated by the brain across different locations and time points during language comprehension is still elusive.

Meanwhile, the field of natural language processing (NLP) has created computational systems that aggregate the meaning of words in specific ways in order to perform a specific linguistic task, such as predicting the upcoming word in a sentence. However, it is not clear whether these computational systems truly understand the meaning of a sentence, and whether the "how" of an NLP system is the same as the "how" of the brain. In this dissertation, we argue that neurolinguistics can benefit from using NLP systems as *model organisms* for how information is aggregated during language comprehension in the human brain, despite NLP systems' differences from the human brain. Model organisms make it easier to study a specific brain function because they allow for direct interventions, which are difficult to do in humans due to ethical or practical reasons. For example, rats are used as model organisms for studying the neural components of spatial representation because of their exploration behavior and because of their size, which enables neural recordings during free behavior (Yartsev, 2017). But what makes a specific organism a good model for a specific function? We argue that a good model organism satisfies three properties: 1) it performs the specific function, 2) it provides an easier platform for studying this function than studying it directly in the human brain, and 3) the function in the model organism can be related to the analogous function in the human brain.

In this dissertation, we provide evidence that NLP systems satisfy the second and third properties of good model organisms for language comprehension in the brain. The second property is arguably the most important benefit of NLP systems to neurolinguistics–the ability to make specific interventions in the NLP system and observe how these interventions affect its alignment with the brain. Much like in an animal model organism, researchers can ablate existing information pathways in an NLP system or introduce new ones. These interventions allow the researcher to make causal claims about the information contained in the NLP system, that are not possible in non-invasive neuroscience studies with real brains. In this dissertation, we provide a case study of using an intervention in an NLP system to study the neural basis of supra-word meaning: the multi-word meaning of language that is beyond the meaning of individual words. To measure the alignment with the brain of the pre- and post-intervention NLP system, we show that we can use encoding models, which are trained to predict brain recordings as a function of representations of text obtained from an NLP system. Encoding models offer one way to relate representations of words from NLP systems to brain recordings of people comprehending language, thereby satisfying the third property of a good model organism.

Brain-Guided NLP Systems

What about the first property of a good model organism? A *good* model organism of language comprehension in the brain must comprehend language. Does an NLP system truly comprehend language? To answer this question, we can examine different ways to quantify language comprehension and observe that NLP systems perform well on some of these metrics, but not others. For example, an NLP system is very good at predicting the upcoming word in a sentence, but may wrongly conclude that "it" in the sentence "I put the heavy table on the book and it broke" refers to

"table" rather than to "book" (Trichelair et al., 2018) because it has presumably learned that tables and breaking appear in context more frequently than books and breaking. This difficulty to reason beyond word associations is a current challenge for NLP systems. Similarly, if we input a chapter of a book in an NLP system, sentence-by-sentence, the NLP system's internal states would be biased towards the recently-processed text (Goodfellow et al., 2016). The farther back we go into the chapter, the worse the NLP system's memory is (Khandelwal et al., 2018; Dai et al., 2019). This is an issue because understanding realistic language requires us to sometimes resolve long-term dependencies. Evaluating the ability of an NLP system to encode long-range context, as well as increasing that ability, is an active area of research.

In contrast to the NLP system, a human who reads a chapter of a book remembers information from the very beginning of the chapter (e.g. contextual information that helps understand how characters relate to one another). So a key question that we ask is: can we use the only processing system that we have that truly understands language–the human brain–to evaluate, and perhaps even improve, the information that these NLP systems are able to encode? To probe the information in the human brain during language comprehension, we follow decades of work in neuroscience and use brain imaging devices (fMRI and MEG) to sample the brain activity as people read text word-by-word. We then conduct different interventions in an NLP system and observe how its alignment with the brain recordings changes. We observe that a specific intervention, which forces the NLP system to place equal weight to all words when aggregating their meanings, improves the alignment with the brain recordings and also improves the performance of the NLP system on a new data distribution. This is the first evidence that brain recordings of people comprehending language can be used to improve the generalization performance of a popular neural network NLP system.

While NLP systems do not yet perfectly comprehend language, they have significantly improved over the last three years along many of the relevant metrics of language comprehension. Importantly, NLP systems are not static, and future improvements that can lead to a more human-like understanding of language will result in even better model organisms. Future work that encourages more human-like language comprehension in NLP system may investigate the role of explicit memory modules in integrating and maintaining linguistic information and may integrate information from multiple sensory modalities, such as vision and audition.

1.1 Thesis Statement and Outline

This dissertation is centered around the following thesis statement:

> *Establishing a data-driven connection between language processing in the brain and language processing in machines can improve: 1) our mechanistic understanding of language processing in the brain through computational modeling, and 2) the generalization performance of natural language processing models through transfer of insight from the brain.*

Chapter 2 contains details about the relevant previous neurolinguistic findings, the brain imaging recording modalities and datasets used in the dissertation, and the NLP systems that are com-

mon across multiple thesis chapters.

Chapter 3 details existing methods for training encoding models, which are used to establish the data-driven connection between language in the brain and language in machines and measure their alignment. It further discusses the limitations of encoding models, specifically when trained as a function of complex representations, such as those obtained from NLP systems. Lastly, it presents two new metrics that enable more precise scientific inferences about information processing in the brain, and validates them in two naturalistic fMRI datasets.

Chapter 4 provides support for the first claim of the thesis by using an intervention in an NLP system to study the neural basis of the multi-word meaning of language that is beyond the meaning of individual words, that we term supra-word meaning (Toneva, Mitchell, et al., 2020). We intervene on an NLP system by isolating this supra-word meaning from the meaning of individual words. Using fMRI recordings, we reveal that hubs thought to process lexical-level meaning also maintain supra-word meaning, suggesting a common substrate for lexical and combinatorial semantics. However, surprisingly, we find that supra-word meaning is difficult to detect in MEG. Instead, the MEG recordings are significantly predicted by information that is unique to the individual recently-read words. The difference between the fMRI and MEG results suggests that the processing of supra-word meaning may be based on neural mechanisms that are not related to synchronized cell firing, as is the MEG signal.

Chapter 5 also provides support for how computational modeling can lead to scientific discovery about language in the brain. In this chapter, we investigate the effect of a question task on the processing of a concrete noun by predicting the millisecond-resolution MEG brain activity as a function of both the semantics of the noun and the task (Toneva, Stretcu, et al., 2020). This work provides the first methodology that predicts brain recordings as a function of both the observed stimulus and question task. Using our proposed approach, we show that incorporating the task semantics (i.e., the specific question asked) significantly improved the prediction of MEG recordings, across participants. The improvement occurs $475 - 550$ms after the participants first see the word, which corresponds to what is considered to be the ending time of semantic processing for a word. These results suggest that only the end of semantic processing of a word is task-dependent.

Chapter 6 provides support for the second claim of the thesis. In this chapter, we develop a method that uses prior neurolinguistic evidence to evaluate the presence of specific brain-relevant information in the representations of an NLP model (Toneva and Wehbe, 2019). The method presents the same text word-by-word to a person in a neuroimaging device and an NLP model, and measures how well the network-derived representations align with the brain recordings in relevant brain regions. This work showed that we can use this method and a snapshot of brain activity, captured by functional magnetic resonance imaging, to reveal how much context is encoded in the representations derived from 4 popular pretrained NLP models. We further showed that altering a state-of-the-art pretrained model to better predict fMRI recordings also significantly improved its generalization performance to a new data distribution. These results are the first evidence that fMRI recordings of people reading can be used to improve a neural network NLP model.

1.2 Summary of Contributions

The contributions of this dissertation can be summarized as follows:

- **Chapter 3:** We conceptually break down the possible underlying relationships for the shared variance between two brain sources, the experimental stimulus, and the selected stimulus representation.
- **Chapter 3:** We present limitations of commonly used methods for disambiguating these different relationships and propose two new approaches that can distinguish them, providing evidence using both simulated data and fMRI data from two naturalistic experiments.
- **Chapter 4:** We introduce a new approach based on computational modeling that makes interventions in NLP systems in order to capture the meaning of the whole as separate from the meaning of the parts. This approach allows the study of complex and composed multi-word meanings in the brain in a way not previously possible.
- **Chapter 4:** We identify potential limitations on the type of information that is detectable in MEG. While high temporal imaging resolution is key to reaching a mechanistic level of understanding of language processing, our findings suggest that a modality other than MEG may be necessary to detect long-range contextual information.
- **Chapter 5:** We provide the first methodology that can predict brain recordings as a function of *both* the observed stimulus and question task. This is important because it will not only encourage neuroscientists to formulate mechanistic computational hypotheses about the effect of a question on the processing of a stimulus, but also enable neuroscientists to test these different hypotheses against each other by evaluating how well they can align with brain recordings.
- **Chapter 5:** We show that models that integrate task and stimulus representations have significantly higher prediction performance than models that do not account for the task semantics, and localize the effect of task semantics largely to time-windows in $475 - 650$ms after the stimulus presentation.
- **Chapter 6:** We present a new method to interpret NLP representations, and find that the middle layers of transformers are better at predicting brain activity than other layers and that Transformer-XL's performance doesn't degrade as context is increased, unlike other popular tested models'. We find that replacing the pretrained attention with uniform attention in early layers of BERT leads to better prediction of brain activity.
- **Chapter 6:** We show that when BERT is altered to better align with brain recordings (by replacing the pretrained attention with uniform attention in early layers), it is also able to perform better at NLP tasks that probe its syntactic understanding (Marvin et al., 2018). These results are the first evidence that fMRI recordings of people reading can be used to improve a neural network NLP model.

1.3 Additional Work

This dissertation includes works in the main line of research on establishing a data-driven connection between language processing in the brain and language processing in machines, and showcasing its benefit for both disciplines. The author has contributed to other works during the PhD that relate to this research direction to various extents. These works are summarized below and interested readers are encouraged to consult the full manuscripts for more details.

Incorporating brain imaging in the training of NLP models. In Schwartz et al., 2019, we present an approach that directly introduces brain-relevant language bias in an NLP model by explicitly training an NLP model to predict language-induced brain recordings. To achieve this, we fine-tune all parameters of a state-of-the-art NLP model, which was already pre-trained in a self-supervised fashion on a large amount of text, to predict two different types of brain imaging measurements—fMRI and Magnetoencephalography (MEG)—of people reading the same chapter of a popular book. We included two different imaging modalities to discourage the encoded bias' dependence on a specific imaging modality. In fact, we show that for some participants, when a model is trained to predict MEG data, the resulting changes to the language-encoding that the model uses benefit subsequent training on fMRI data compared to starting with a language model trained only on text. This suggests that the changes to the language representations induced by the MEG data are not entirely imaging modality-specific, and that indeed the model is learning the relationship between language and brain activity as opposed to the relationship between language and a brain activity recording modality. We show that the resulting model leads to improved prediction of previously unseen brain recordings, specifically in regions that are known to support language processing. This study demonstrates the feasibility of directly biasing language models to learn relationships between text and brain activity.
Dan Schwartz, Mariya Toneva, and Leila Wehbe. "Inducing brain-relevant bias in natural language processing models". In: *Advances in Neural Information Processing Systems*. 2019, pp. 14123–14133

Structurally-guided attention mechanisms. In Abdou et al., 2021, we propose an approach that provides more control over the language model representations by injecting structural bias from specific syntacto-semantic formalisms via a structurally-guided attention mechanism, and then test how this injected bias affects the alignment of the model representations with brain recordings. We apply the proposed approach to three representative syntacto-semantic formalisms and two fMRI datasets. We find that 1) biasing the model representations towards two of the three linguistic formalisms resulted in improved alignment with brain recordings corresponding to human processing of sentences and content words, and 2) the match between the textual domain of the fine-tuning data and that of the brain recording stimuli plays an important role, despite having been previously overlooked. Our proposed approach enables the evaluation of more targeted hypotheses about the relationship between NLP models and brain activity, and opens up new opportunities for cross-pollination between NLP, computational neuroscience, and linguistics.
Mostafa Abdou, Ana Valeria Gonzalez, Mariya Toneva, Daniel Hershcovich, and Anders Søgaard.

"Does injecting linguistic structure into language models lead to better alignment with brain recordings?" In: *arXiv preprint arXiv:2101.12608* (2021)

Characterizing example forgetting in neural networks. In Toneva, Sordoni, et al., 2019, we investigate the learning dynamics of neural networks as they train on single classification tasks. Our goal is to understand whether a phenomenon related to catastrophic forgetting occurs when data does not undergo a clear distributional shift. We define a "forgetting event" to have occurred when an individual training example transitions from being classified correctly to incorrectly over the course of learning. Across several benchmark data sets, we find that: 1) certain examples are forgotten with high frequency, and some not at all; 2) a data set's (un)forgettable examples generalize across neural architectures; and 3) based on forgetting dynamics, a significant fraction of examples can be omitted from the training data set while still maintaining state-of-the-art generalization performance.
Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. "An Empirical Study of Example Forgetting during Deep Neural Network Learning". In: *International Conference on Learning Representations (ICLR)*. 2019

Fine-grained models of neural scene understanding. In Aminoff et al., 2015, we leverage well-specified computer vision systems to investigate how the neural responses of people viewing natural images relate to different attributes in the images. We find that: 1) vision systems that rely on mid- and high-level scene attributes show the highest correlations with the patterns of neural activity within the three scene-selective regions (i.e. the parahippocampal/lingual region (PPA), the retrosplenial complex (RSC), and the occipital place area (TOS)); 2) The best performing vision system accounts for neural data better than human judgments of scene similarity. One computer vision system—NEIL ("Never-Ending-Image-Learner"), which incorporates visual features learned as statistical regularities across scenes—shows significant correlations with neural activity in all three scene-selective regions and was one of the two models best able to account for variance in the PPA and TOS. These results are a first step towards a more fine-grained theory of neural scene understanding, including developing a clearer picture of the division of labor among the components of the functional scene-selective brain network.
Elissa M Aminoff, Mariya Toneva, Abhinav Shrivastava, Xinlei Chen, Ishan Misra, Abhinav Gupta, and Michael J Tarr. "Applying artificial vision models to human scene understanding". In: *Frontiers in computational neuroscience* 9 (2015), p. 8

Chapter 2

Background and Related Work

Understanding the processes that are involved in language comprehension has interested many philosophers, linguists, psycholinguists, neurolinguists, and computer scientists. In this dissertation, we aim to bridge the empirical methodologies for understanding language comprehension in the brain with the computational methodologies that are designed to process language. This chapter first summarizes the relevant neurolinguistic findings about language in the brain, and discusses the classical and modern computational methods that are designed to process language. We lastly briefly review prior work that relates neurolinguistics with computational models of language, and end by describing the brain recordings datasets that we use in this dissertation to bridge language in the brain with language in machines.

2.1 Language in the Brain

The invention of non-invasive imaging modalities which can sample large-scale brain activity (as opposed to activity from only a few neurons at a time) has enabled neuroscientific studies of cognitive functions, such as language, in healthy individuals. In this section, we first give details about the most popular non-invasive imaging modalities and then summarize the previous findings about language in the brain that have been enabled by these brain imaging modalities.

2.1.1 Sampling Language in the Brain via Brain Imaging

The most common non-invasive brain imaging modalities are Electroencephalography (EEG), Magnetoencephalography (MEG), and functional Magnetic Resonance Imaging (fMRI). Each modality offers different advantages and disadvantages when recording brain activity, which we briefly summarize below. In this dissertation, we use brain recordings from two of these brain imaging modalities that have complimentary strengths–fMRI and MEG.

Functional Magnetic Resonance Imaging (fMRI). fMRI measures the change in oxygen level in the blood that is a consequence to neural activity. When neurons become active, they consume oxygen, which causes a change in the relative levels of oxygenated and deoxygenated blood in the

corresponding brain area. This change can be detected by a Magnetic Resonance Imaging device because oxygenated and deoxygenated blood have different magnetic properties which leads to a variation in magnetic signal. This change is called the hemodynamic response and the ratio between oxygenated and deoxygenated blood is termed BOLD response (i.e. blood-oxygen-level-dependent), and is the basis for most fMRI. Once neurons in a brain area are active, it takes about 12 seconds for the BOLD response to return to its pre-activity baseline. The BOLD response is typically sampled every $1-2$ seconds. The spatial resolution of the fMRI image depends largely on the strength of the MRI magnet. Using a typical MRI with a 3T magnet results in a sample of the BOLD response in every $1-2\text{mm} \times 1-2\text{mm} \times 1-2\text{mm}$ volume pixel in the brain. A major limitation of BOLD-based fMRI is that its measurements correspond to bloodflow and not actual neuronal activity.

Electroencephalography (EEG) and Magnetoencephalography (MEG). EEG and MEG respectively measure the electric and magnetic fields that are generated by a large number of neurons firing in synchrony. EEG measures the electric field using electrodes positioned directly on the participant's scalp, and MEG measures the magnetic field using sensors that are typically housed in a helmet, in which the participant places her head. Both imaging modalities have high temporal resolutions (up to 1KHz) and measure neural activity more directly than fMRI, which measures blood flow. However, the spatial resolution of both EEG and MEG is poorer than that of fMRI, with localizing the source of the EEG signals being the most difficult because the electric fields are distorted when passing through the skull. Magnetic fields do not suffer from this distortion. However, MEG is primarily sensitive to activity in the sulci of the brain (because of the alignment of the magnetic field), and is a lot less sensitive to activity in the gyri.

2.1.2 Individual Word Processing

Using brain imaging techniques with high temporal resolution, researchers have begun to understand the stages of processing during language comprehension. In Figure 2.1, we summarize these stages of processing. The first stage occurs when we first read or hear a word, as the visual or auditory input reaches the corresponding visual or auditory cortices in 100ms. In the second stage, this input is then processed as strings of letters or phonemes around 150ms by the visual word form area in the case of reading or the posterior superior temporal gyrus and sulcus (pSTG and pSTS) in the case of hearing, which are more active during this time for language stimuli than other visual or auditory stimuli (Salmelin, 2007; Friederici, 2011). Comparing the brain responses to real and made-up words reveals that the meaning of a word is processed between $200-600$ms after the word is first presented, and this processing appears to be supported by the temporal cortex (Salmelin, 2007; Friederici, 2011). The remaining stages of processing are most relevant for comprehending multi-word language, and will be discussed in Chapter 2.1.3.

Word meaning. Theories of language processing consider the temporal lobe as central to the general retrieval and processing of words. However, these theories do not specify how the meaning of specific words or word categories (e.g. tools, animals, etc.) is represented in the brain. One

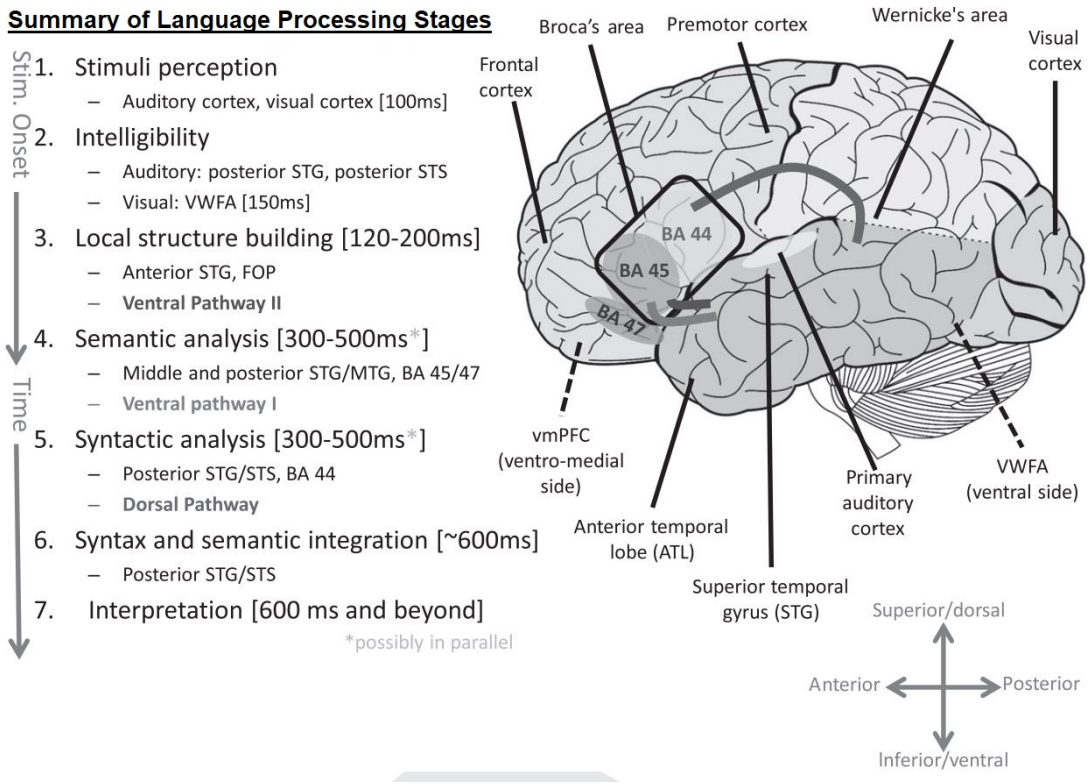


Figure 2.1: Summary of the stages of language processing in the brain. Adapted from Friederici, 2011; Murphy et al., 2018.

hypothesis for how concrete concepts (e.g. ”dog”) are represented in the brain that has significant empirical support from neuroscience studies is the Grounded Cognition Model (also known as the Embodied Cognition Model or the Simulation Model) (Kemmerer, 2014a). This hypothesis posits that concrete concepts are represented via the perceptual experiences that are associated with this concept (Barsalou, 1999; Barsalou, 2008; Pecher et al., 2005). Research guided by this hypothesis suggests that the semantic attributes of concrete concepts are stored in a distributed but organized manner across cortex, in such a way that a specific semantic attribute (e.g. auditory feature) is stored in the same part of cortex that underlies the high-level sensory perception that is related to this attribute (e.g. auditory perception). In addition to audition (Kiefer et al., 2008), there is empirical evidence for such organization of semantic attributes related to color (Simmons et al., 2007), shape (Chao et al., 1999), motion (Damasio et al., 1996), olfaction and taste (Goldberg, Perfetti, et al., 2006a; Goldberg, Perfetti, et al., 2006b). This hypothesis has further support from a computational modeling approach that predicted fMRI recordings as a function of semantic properties of a word (Mitchell, Shinkareva, et al., 2008). In this work, the authors found correspondences between a semantic property and the function of the cortical regions in which the semantic property predicted the fMRI recordings (e.g. the semantic property exemplified by the verb ”push” significantly predicted the activity in motor cortex).

These different semantic attributes are thought to be integrated together in a higher-order rep-

resentation by the bilateral anterior temporal lobes (ATL) (Visser et al., 2010). The ATL is thought to organize the semantic attributes in a way that enables distinguishing between objects that are within the scope of a particular concept and those that are out of scope. This representation can then be used by other brain areas. Support for the ATL as a hub that integrates semantic attributes comes from several sources: 1) clinical studies of patients with progressive object concept dissolution that is tightly coupled with progressive ATL atrophy (Bright et al., 2008) and 2) studies using an invasive imaging technique called rTMS show that temporary disruption in the ATLs of healthy participants reduces their ability to process object concepts (Pobric et al., 2007; Lambon Ralph et al., 2009).

2.1.3 Multi-word Composition

Understanding a multi-word phrase requires additional processing stages beyond the ones that support the processing of individual words. Specifically, theories from linguistics and cognitive psychology posit that words can be combined according to several different sets of rules to form composed meaning. Here we summarize the different types of hypothesized composition and the existing support for these processes in the human brain.

Syntactic composition. One such set of rules is represented by syntax. For example, an adjective followed by a noun can always be combined to form a noun phrase. However, the resulting syntactically composed meaning may not always be semantically valid (e.g. colorless green ideas sleep furiously (Chomsky, 2002)).

One marker for syntactic and possibly logico-semantic composition (see below) in the brain is thought to be the P600. P600 is a prototypical brain response measured via electrophysiological brain recordings (i.e. EEG, MEG). It indicates a positive deflection (relative to a reference baseline) in the measured signal which occurs between 500ms and 800ms after stimulus onset, with its maximum magnitude around 600ms. The P600 is thought to be due to syntactic violations (Kemmerer, 2014b; Coulson et al., 1998) as well as thematic role violations (Kutas, Van Petten, et al., 2006; Kuperberg, 2007), and is thought to be generated by the bilateral temporal lobes in locations that are posterior to those thought to generate the N400, described below (Service et al., 2007).

Logico-semantic composition. Another set of rules is represented by logico-semantic composition, which composes predicate-argument structures (Pylkkänen, 2020). The predicate-argument structure does not directly follow from the syntactic rules, which necessitates the separate logico-semantic composition (Partee, 2002). Pylkkänen, 2020 illustrates this disparity by contrasting the following two phrases with parallel syntactic structure: ‘she liked my eye color’ and ‘she guessed my eye color’ with the same phrases but with ’eye color’ replaced by ’eyes’. ‘She liked my eyes‘ sounds just fine, but ‘she guessed my eyes‘ does not. It turns out that this is because ‘guess‘ and ‘like‘ have different semantic restrictions (Nathan, 2006), because ‘guess‘ specifically requires a question as its argument and nouns that describe relations, such as ‘color‘, can be converted into a question, whereas nouns that describe entities, such as ‘eyes‘ cannot.

Similarly to the P600, the N400 is a prototypical brain response. It occurs between 200ms and 600ms after stimulus onset (Kutas and Federmeier, 2011). The N400 is thought to be related to the semantic effort that is necessary to integrate a word in a specific context (Kutas, Van Petten, et al., 2006), and is thought to be generated by the temporal lobes, bilaterally. The superior temporal gyrus (STG) and middle temporal gyrus (MTG) have been specifically implicated in the N400 (Kutas and Federmeier, 2011; Van Petten et al., 2006) and the N400 is thought to not be generally affected by syntax (Allen et al., 2003).

Conceptual composition. Neither syntactic nor logico-semantic composition account for the actual conceptual content of the words that are being composed. This makes room for a third type of composition – conceptual composition (Pylkkänen, 2020). This conceptual composition has been investigated in cognitive psychology with a focus on adjective-noun and noun-noun composition (Hampton, 1997; Murphy, 1990; Smith and Osherson, 1984).

The left ATL (LATL) is thought to be a site of conceptual composition, and its activity cannot be explained purely by syntactic and logico-semantic composition (Pylkkänen, 2020). For example, in a series of articles, Liina Pylkkänen and colleagues found that the LATL activity was not modulated every time an adjective and a noun combined, but instead depended on the actual concepts that were described by the words and the specificity of their meaning (Westerlund et al., 2014; Zhang et al., 2015; Ziegler et al., 2016). The hypothesis behind this conceptual composition is that the LATL effect is driven by the way that the integration of the first word affects the feature space of the second word (Pylkkänen, 2020). The general idea is that if the second word is general, its feature space is sparse, and the first word has the ability to contribute a significant proportion of features to the final composed meaning. Likewise, a very specific first word also has the ability to contribute a large proportion of features. From this hypothesis, it follows that more specific first words and more general second words would elicit large conceptual composition effects in the LATL.

2.2 Language in Machines

2.2.1 Distributional Semantic Models

Several researchers have investigated ways to specifically encode compositional structure into computational models of language. Mitchell and Lapata, 2010 constructed vector representations for adjective-noun, noun-noun, and verb-object pairs using different functions of the individual corresponding word vector representations. Baroni and Zamparelli, 2010 model adjective-noun phrases as a product between a vector that corresponds to the co-occurrence of the noun in a large corpus and a matrix that is estimated and is meant to represent the adjective. This framework was further generalized to other type-theoretic objects by Baroni, Bernardi, Zamparelli, et al., 2014, and other closely related approaches to compositional distributional semantics were proposed by Grefenstette et al., 2011 and Clark, 2013 (see Lenci, 2018 for a thorough review). In addition, Baroni, Bernardi, Do, et al., 2012 showed that adjective-noun phrase representations based on

co-occurrence statistics are useful for detecting accurate entailment relationships at the individual word-level.

2.2.2 Natural Language Processing Systems

Neural networks have undergone a revolution in the last 10 years, which is driven by the availability of larger training datasets, computational power and better optimization methods. In the domain of language, all of these three developments have allowed multi-layer neural networks to learn to extract meaning from sequences of words and to perform a wide variety of sophisticated linguistic tasks. Specifically, one of the most important developments that enabled the success of modern NLP systems is the ability to learn the statistics of language using a very simple objective, called *language modeling*. Under this objective, the system learns to predict what word will appear next, given the previous context (Graves, 2012; Mikolov et al., 2012). Language modeling appears to be a simple objective, but in modern NLP, language modeling is a very powerful tool that is used to teach a network a general-purpose understanding of language statistics, in what is called a pretraining stage. This phase is called *pre*training because it is often followed by a second stage that trains the network to perform a specific task (Devlin et al., 2018; Radford et al., 2019). In this second "fine-tuning" stage, the parameters of the NLP system are adjusted such that the system now performs the specified task accurately. In this dissertation, we make use of publicly available NLP systems that have been pretrained on a large amounts of text. We specifically investigate two such systems–a recurrence-based system called ELMo, and a transformer-based system called BERT. We provide more details about each system below.

Embeddings from Language Models (ELMo)

ELMo is a recurrence-based NLP system that incorporates multiple layers of long-short-term memory units (LSTMs) (Peters et al., 2018). For a word token t, an LSTM generates the corresponding hidden representation h_t^l in layer l using the following update equations:

$$\begin{aligned}
\tilde{c} &= \tanh(w_c[h_{t-1}^l; h_t^{l-1}] + b_c), \\
c_t &= f_t \times c_{t-1} + i_t \times \tilde{c}_t, \\
h_t^l &= o_t \times \tanh(c_t),
\end{aligned}$$

where b_c and w_c represent the learned bias and weight, and f_t, o_t, and i_t represent the forget, output, and input gates. The states of the gates are computed according to the following equations:

$$\begin{aligned}
f_t &= \sigma(w_f[h_{t-1}^l; h_t^{l-1}] + b_f), \\
i_t &= \sigma(w_i[h_{t-1}^l; h_t^{l-1}] + b_i), \\
o_t &= \sigma(w_o[h_{t-1}^l; h_t^{l-1}] + b_o),
\end{aligned}$$

where $\sigma(x)$ represents the sigmoid function and b_x and w_x represent the learned bias and weight of the corresponding gate. The learned parameters are trained to predict the identity of a word given a series of preceding words, in a large text corpus.

At each layer, for each word token ELMo combines the internal representations of two independent LSTMs–a forward LSTM (containing information from previous words) and a backward LSTM (containing information from future words). In this dissertation, we use a pretrained version of ELMo with 2 hidden LSTM layers provided by Gardner et al. (2018). The hidden state of each independent LSTM (the forward and the backward) in this pretrained version has 512 dimensions, and the system has a total of 13.6 million parameters. This system was trained on the One Billion Word Benchmark (Chelba et al., 2013), which contains approximately 800 million tokens of news crawl data from WMT 2011[1].

Bidirectional Encoder Representations from Transformers (BERT)

BERT is an NLP system that incorporates multiple layers of transformer units (Devlin et al., 2018). The central innovation in the transformer, which was originally introduced by Vaswani et al., 2017, is the multi-head attention mechanism. To understand the multi-head attention mechanism, let us first describe attention. Informally, attention can be seen as a content-based weighted combination of a sequence of input vectors. More formally,

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V,$$

where $Q, K \in \mathbb{R}^{n \times d_k}$ are the queries and keys for the corresponding input sequences of n word tokens, and $V \in \mathbb{R}^{n \times d_v}$ are the values for the corresponding word tokens. Now, the multi-head attention mechanism can be defined as follows:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(h_1, ..., h_h)W^O, \\ \text{where } h_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \end{aligned}$$

where W_i^Q, W_i^K, W_i^V are the parameter matrices for attention head h_i and W^O is the final output transformation. All learned parameters are trained to predict the identity of a masked word in a surrounding context of words, in a large text corpus (e.g. "Mary [MASK] the apple"). Additionally, two special tokens are added to the input sequence to signify the beginning and the end of the sequence (e.g. "[CLS] Mary [MASK] the apple [SEP]"). The final hidden state for the [CLS] token (i.e. the state at the last hidden layer) is used as the aggregate sequence representation for classification tasks, and the [SEP] token is used to separate multiple input sequences.

In this dissertation, we use a pretrained version of BERT provided by Hugging Face [2]. We investigate the base BERT system, which has 12 layers of transformers, 12 attention heads per layer, and 768 units in each hidden state representation. This system has a total of 110 million parameters. The system was trained using long contiguous sequences of words extracted from the BooksCorpus (Zhu, Kiros, et al., 2015) (800 million words) and English Wikipedia (2500 million words).

[1] http://statmt.org/wmt11/

[2] https://github.com/huggingface/pytorch-pretrained-BERT/

2.2.3 Linguistic Properties Captured by NLP Systems

State-of-the-art NLP systems learn to extract meaning from sequences of words to perform a wide variety of sophisticated linguistic tasks. They can be thought of, therefore, as one implementation of a language system that is likely very different from that of the brain, but may nonetheless offer insights into the nature of the linguistic tasks and the computational processes that are sufficient (or insufficient) to solve these tasks (McCloskey, 1991; Baroni, 2020). Given this perspective, here we summarize what is known about how neural network models for NLP learn to compose words.

Syntactic structure encoded by language models. Gulordava et al., 2018 show that a recurrent neural network (RNN) trained only with a language modeling objective predicts grammatical over ungrammatical nonsensical sequences of words with high accuracy. Specifically, the authors tested whether the RNN could predict the verb that agrees with the subject's accurate number over the alternate number. For example, "colorless green ideas *sleep* furiously" is an example of a grammatical nonsensical word sequence and "colorless green ideas *sleeps* furiously" is an example of an ungrammatical sequence. The tested word sequences were selected to be nonsensical so that the model could not depend on semantic or lexical cues to perform the prediction task. Lakretz et al., 2019 conducted further ablation and connectivity studies of the model from Gulordava et al., 2018 to better understand how the RNN computes the subject-verb agreement. The authors found a small set of units which were necessary for the agreement computation, as the performance approached chance level once these units were fixed to 0. Furthermore, these units were strongly connected to a subnetwork that was independently sensitive to hierarchical syntactic constituency. Taken together, these findings suggest that the network has developed a mechanism to process syntax.

Semantic role labeling (SRL) performance as an indicator of logico-semantic composition. The NLP task of semantic role labeling (SRL) aims to determine, given a sequence of text, "who did what to whom", "when", and "where". As such, this task can be seen as an indicator of whether a specific NLP model is able to encode the predicate-argument structure present in the text, and therefore perform logico-semantic composition. Both recurrence-based (Peters et al., 2018) and attention-based (Shi et al., 2019) neural network models for NLP achieve significant performance (F1 in the $80 - 90$ range) at the SRL task.

2.3 Relating Language in the Brain to Language in Machines

2.3.1 Prior Work

A few previous works have used neural network representations as a source of feature spaces to model brain activity. Wehbe, Vaswani, et al. (2014) aligned MEG brain activity with a Recurrent Neural Network (RNN), trained on an online archive of Harry Potter Fan Fiction. The authors aligned brain activity with the context vector and the word embedding, allowing them to trace sentence comprehension at a word-by-word level. Jain et al., 2018 aligned layers from a Long

Short-Term Memory (LSTM) model to fMRI recordings of subjects listening to stories to differentiate between the amount of context maintained by each brain region. Schrimpf et al., 2020 investigate the alignment of fMRI and ECoG recordings of people reading and listening to language to representations obtained from more recent NLP systems. The authors report that some of the NLP system representations can predict the brain recordings up to an estimate of the noise ceiling. Similarly, Caucheteux and King, 2020 find that the representations from a large number of neural networks align well with MEG recordings of people reading, specifically when the representations are obtained from the middle layers of these networks. This finding replicates the results of Jain et al., 2018 and (Toneva and Wehbe, 2019) that were obtained using fMRI data. Goldstein et al., 2021 align representations from a neural network and ECoG recordings of people listening to stories and find that ECoG electrodes can predict the neural network representation of upcoming words in the narrative.

Other approaches rely on computing surprisal or cognitive load metrics using neural networks to identify processing effort in the brain, instead of aligning entire representations (Frank et al., 2015; Hale et al., 2018). There is little prior work that evaluates or improves NLP models through brain recordings. Søgaard (2016) proposes to evaluate whether a word embedding contains cognition-relevant semantics by measuring how well they predict eye tracking data and fMRI recordings. Fyshe, Talukdar, et al. (2014) build a non-negative sparse embedding for individual words by constraining the embedding to also predict brain activity well and show that the new embeddings better align with behavioral measures of semantics.

2.3.2 Brain Recordings Datasets Used in Dissertation

To relate language in the brain to language in machines in this dissertation, we use 2 fMRI and 2 MEG datasets because of their complementary strengths. One of the fMRI and MEG datasets share the same experimental paradigm, which enables us to more directly compare results from these datasets. All datasets were collected from healthy individuals. Additionally, three of the four datasets were recorded under a naturalistic experimental paradigm (e.g. reading a chapter of a popular book word-by-word and watching a popular movie). Both fMRI datasets are publicly available. We will make one of the MEG datasets publicly available with the publication of Toneva, Mitchell, et al., 2020. The remaining MEG dataset cannot be made publicly available due to IRB restrictions. We provide more details about each dataset below.

Harry Potter (fMRI, MEG)

In Chapter 4 and Chapter 6, we use fMRI and MEG brain recordings from participants reading chapter 9 of *Harry Potter and the Sorcerer's Stone* (Rowling, 2012) word-by-word. The book chapter contains 5176 words. Words were presented one at a time at a rate of 0.5s each. The experimental paradigm is exactly the same in the two imaging modalities. Both datasets were recorded from participants who had read *Harry Potter and the Sorcerer's Stone* and were native English speakers. The modality-specific details and preprocessing are provided below.

fMRI data and preprocessing We use fMRI data of 9 participants, collected and made available online by Wehbe, Murphy, et al. (2014). The fMRI data was acquired at a rate of 2s per image, i.e. the repetition time (TR) is 2s. The images were comprised of $3 \times 3 \times 3mm$ voxels. The data for each participant was slice-time and motion corrected using SPM8 (Kay et al., 2008), then detrended and smoothed with a 3mm full-width-half-max kernel. The brain surface of each participant was reconstructed using Freesurfer (Fischl, 2012), and a grey matter mask was obtained. The Pycortex software (Gao, Huth, et al., 2015) was used to handle and plot the data. For each participant, $25000 - 31000$ cortical voxels were kept.

MEG data and preprocessing We use MEG data of 8 participants, collected by the authors of (Wehbe, Vaswani, et al., 2014) and shared upon our request. This data was recorded at 306 sensors organized in 102 locations around the head. MEG records the change in magnetic field due to neuronal activity and the data we used was sampled at 1kHz, then preprocessed using the Signal Space Separation method (SSS) (Taulu, Kajola, et al., 2004) and its temporal extension (tSSS) (Taulu and Simola, 2006). The signal in every sensor was downsampled into 25ms non-overlapping time bins. For each of the 5176 words in the book chapter, we therefore obtained a recording for 306 sensors at 20 time points after word onset (since each word was presented for 500ms).

Courtois NeuroMod (fMRI)

In Chapter 3, we use fMRI brain recordings from 6 participants who watch the movie Hidden Figures. This dataset is provided by the Courtois NeuroMod group (data release `cneuromod-2020`). Three participants are native French speakers and three are native English speakers. All participants are fluent in English and report regularly watching movies in English.

The length of recording for each participant is approximately 120 minutes. The fMRI sampling rate (TR) was 1.49 seconds. The images were comprised of $2 \times 2 \times 2mm$ voxels. Results included in this dissertation come from a standard preprocessing pipeline performed using fMRIPrep 20.1.0 (Esteban, Markiewicz, et al., 2018; Esteban, Blair, et al., 2018). This data is available by request at `https://docs.cneuromod.ca/en/latest/ACCESS.html#downloading-the-dataset`.

We include results from this dataset as a second naturalistic fMRI dataset, which is collected from a very different experimental paradigm (i.e. auditory language presentation vs. the visual language presentation in Harry Potter) and from a very different participant population (half were non-native English speakers).

20questions (MEG)

In Chapter 5, we aim to study the effect of a task on the brain representation of a stimulus, when the stimulus is shown while performing the task. To this end, we use MEG brain recordings from 6 participants who were asked to perform a question-answering task. The data was collected by the authors of Sudre et al. (2012a) and provided upon our request. The dataset originally contained

data from 9 participants. We exclude data from 3 of these 9 participants in our analyses because of missing trials.

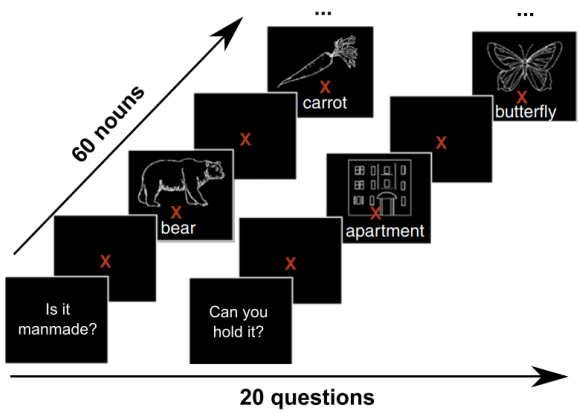


Figure 2.2: Experimental paradigm recreated from Sudre et al. (2012a). Subjects are shown a question, followed by 60 concrete nouns along with their line drawings in random order.

Figure 2.2 illustrates the experimental paradigm. Participants were first presented with a question (e.g., *"Is it manmade?"*), followed by 60 concrete nouns, along with their line drawings, in a random order. Each stimulus was presented until the subject pressed a button to respond *"yes"* or *"no"* to the initial question. Once all 60 stimuli are presented, a new question is shown for a total of 20 questions. Thus we have a total of 60 stimuli $\times$ 20 questions $=$ 1200 examples.

Preprocessing. The data were first preprocessed using the Signal Space Separation method (SSS) (Taulu and Simola, 2006) in order to isolate the signal components that originate inside of the sensor array. This method was followed by its temporal extension (tSSS) to align the measurements of the head position before each block to a common space. The MEG signal was then filtered using a low-pass filter at 150Hz and notch filters at 60Hz and 120Hz to remove contributions from electrical line noise and other very high frequency noise. Next, the Signal Space Projection method was applied to remove eye blinks, residual movement, and other artifacts.

Chapter 3

Methods

This chapter contains details about the method we use to extract representations of words and phrases of various length from natural language processing systems, and the existing and contributed methods for how to relate these computational representations of language to brain recordings of people comprehending language.

The last part of the chapter (Chapter 3.4) is based on work done in collaboration with Anand Bollu, Jennifer Williams, Christoph Dann, and Leila Wehbe and is under review. In this work, we present an in-depth analysis of the most frequently used method of relating computational representations of language to brain recordings–the linear encoding model. Using simulated data with known ground truth relationships between brain measurements, an observed stimulus, and a stimulus representation, we first highlight limitations of existing computational approaches for inferring these relationships. We propose two new metrics and an analysis framework that, when used together, enables scientific inferences that can disambiguate different possible configurations for the underlying relationships. We next test all approaches in two fMRI datasets obtained using naturalistic stimuli, and observe that the proposed methods allow us to infer a more specific relationship among brain regions with respect to the stimulus that is strikingly consistent between the two datasets. In this chapter, we include the results that were obtained for one of the datasets and include the results for the other dataset in Appendix A. Overall we present evidence that our proposed framework is a promising new tool for computational neuroscientists who are interested in mapping information processing in the brain.

3.1 Introduction

A key building block of this thesis is the ability to measure the correspondence between brain recordings of people comprehending language and computational representations of the same linguistic input that were obtained from NLP systems. One of the computational tools for measuring this correspondence is the linear encoding model, which predicts a brain measurement as a linear combination of a numerical representation of the experimental stimulus (e.g. a word embedding obtained from an NLP system). If an encoding model that is trained as a function of a specific stimulus representation X is able to predict brain recordings significantly better than chance, then

the brain recording is assumed to contain information about X.

In this chapter, we first describe how we obtain stimulus representations from NLP systems. We next discuss the existing techniques for training and evaluating linear encoding models that are used throughout this thesis. We follow by outlining the limitations of the possible scientific inferences from encoding models trained using complex stimulus representations X (e.g. X represents multiple aspects of the stimulus, for instance a word's part-of-speech and its semantic role). We lastly propose two new computational techniques that can improve scientific inferences for encoding models of complex stimuli and stimuli representations.

3.2 Obtaining stimulus representations from NLP systems

We obtain a stimulus representation $x_t \in \mathbb{R}^d$ for word w_t from a pretrained NLP system (e.g. ELMo, BERT. See Chapter 2.2.2 for details about all NLP systems considered in multiple chapters of this thesis.) by passing word w_t through the pretrained system and obtaining the token-level embeddings (i.e. from layer 0) for w_t. If word w_t contains multiple tokens, we average the corresponding token-level embeddings and use this average as the final word representation. We obtain a contextualized stimulus representation for word w_t by passing the most recent n words $(w_{t-n+1}, .., w_t)$ through the pretrained NLP system and obtaining the embeddings from intermediate layer ℓ. For systems such as ELMo, which concatenate a forward and a backward LSTM (see Ch. 2.2.2), we extract context embeddings only from the forward LSTM in order to more closely match the participants, who have not seen the future words. If word w_t contains multiple tokens, we average the corresponding layer ℓ embeddings and use this average as the final contextualized representation for word w_t. The number of words n in the context and the intermediate layer ℓ can vary and are individually specified for each work in the different chapters.

3.3 Encoding models

3.3.1 General setting

An encoding model estimates the function $g(\cdot)$ for a specific stimulus representation X and brain source Y_i, such that:

$$Y_i = g(X) + \epsilon \tag{3.1}$$

where $Y_i \in \mathbb{R}$ corresponds to the observation at a single brain source i (e.g. fMRI voxel, EEG/MEG sensor-timepoint, electrode), $X \in \mathbb{R}^d$ to the d-dimensional numerical representation of the corresponding stimulus (e.g. a word embedding obtained from inputting a word stimulus in a language model), and $\epsilon \in \mathbb{R}$ to a noise term that is independent from the stimulus representations.

Most commonly, the function $g(\cdot)$ is parameterized as a linear function (i.e. $g(X) = \langle X, \theta \rangle$, where $\theta \in \mathbb{R}^d$) and is estimated using a set of training observations for each brain source in a cross-validated fashion (i.e. for training data pairs $(x_1, y_{i,1}), \ldots, (x_m, y_{i,m})$). For a set of predictions $\widehat{Y}_i$

on heldout data, the encoding model performance is defined as follows:

$$\texttt{encoding model performance}(\widehat{Y}_i, Y_i) = \text{metric}(\widehat{Y}_i, Y_i) \ ,$$

where $\text{metric}(\cdot)$ signifies one of the evaluation metrics that we describe in detail in Chapter 3.3.4, such as Pearson correlation.

In this thesis, we follow this general setting in training and evaluating encoding models. Below we provide more details about the specific encoding model training and evaluation that is necessary for the two types of brain imaging recordings that we consider.

3.3.2 Encoding models for fMRI

Ridge regularization is used to estimate the parameters of a linear model that predicts the brain activity y_i in every fMRI voxel i as a linear combination of a particular NLP embedding x. For each output dimension (voxel), the ridge regularization parameter is chosen independently by nested cross-validation. We use ridge regression because of its computational efficiency and because of the results of Wehbe, Ramdas, et al. (2015) showing that for fMRI data, as long as proper regularization is used and the regularization parameter is chosen by cross-validation for each voxel independently, different regularization techniques lead to similar results. Indeed, ridge regression is a common regularization technique used for building predictive fMRI models (Mitchell, Shinkareva, et al., 2008; Nishimoto et al., 2011a; Wehbe, Murphy, et al., 2014; Huth, Heer, et al., 2016).

For every voxel i, a model is fit to predict the signals $y_i = [y_{i,1}, y_{i,2}, \ldots, y_{i,m}]$, where m is the number of time points, as a function of the NLP embedding. The words presented to the participants are first grouped by the TR interval in which they were presented. Then, the NLP embedding of the words in every group are averaged to form a sequence of features $x = [x_1, x_2, \ldots, x_m]$ which are aligned with the brain signals. The models are trained to predict the signal at time t, y_t, using the concatenated vector z_t formed of $[x_{t-1}, \ldots, x_{t-p}]$, where p is dependent on the length of the TR. The features of the words presented in the previous volumes are included in order to account for the lag in the hemodynamic response that fMRI records. Indeed, the response measured by fMRI is an indirect consequence of brain activity that peaks about 6 seconds after stimulus onset, and the solution of expressing brain activity as a function of the features of the preceding time points is a common solution for building predictive models (Nishimoto et al., 2011a; Wehbe, Murphy, et al., 2014; Huth, Heer, et al., 2016). For the Harry Potter fMRI data used in this thesis, we set $p = 4$ because the TR length is 2 seconds, and for the Courtois NeuroMod fMRI data we set $p = 6$ because the TR length is 1.49 seconds.

For each given participant and each NLP embedding, we perform a cross-validation procedure to estimate how predictive that NLP embedding is of the brain recordings in each voxel i. For each fold:

- The fMRI data and feature matrix $Z = z_1, z_2, \ldots z_n$, where $z_t = [x_{t-1}, \ldots, x_{t-p}]$ as described above, are split into corresponding train and validation matrices. These matrices are individually normalized (mean of 0 and standard deviation of 1 for each voxel across time), resulting in train matrices Y_i^R and Z^R and validation matrices Y_i^V and Z^V.

- Using the train fold, a model θ_i is estimated as:

$$\arg\min_{\theta_i} \|y_i^R - Z^R\theta_i\|_2^2 + \lambda_i\|\theta_i\|_2^2.$$

  A ten-fold nested cross-validation procedure is first used to identify the best λ_i for every voxel i that minimizes the nested cross-validation error. θ_i is then estimated using λ_i on the entire training fold.

- The predictions for each voxel on the validation fold are obtained as $\widehat{Y}_i^V = Z^V\theta_i$.

The above steps are repeated for each of the cross-validation folds and average evaluation metric value is obtained for each voxel i, NLP embedding, and participant.

3.3.3 Encoding models for MEG

MEG data is sampled faster than the rate of word presentation, so for each word, we have several time points recorded at 306 sensors (20 time points for Harry Potter, and 32 time points for 20questions). Ridge regularization is similarly used to estimate the parameters of a linear model that predicts the brain activity $y_{(i,\tau)}$ in every MEG sensor i at time τ after word onset. For each output dimension (sensor/time tuple (i, τ)), the Ridge regularization parameter is chosen independently by nested cross-validation.

For every tuple (i, τ), a model is fit to predict the signals $y_{(i,\tau)} = [y_{(i,\tau),1}, y_{(i,\tau),2}, \ldots, y_{(i,\tau),m}]$, where m is the number of words in the story, as a function of NLP embeddings. We use as input the word vector x without the delays we used in fMRI because the MEG recordings capture instantaneous consequences of brain activity (change in the magnetic field). The models are trained to predict the signal at word t, $y_{(i,\tau),t}$, using the vector x_t.

For each participant and NLP embedding, we perform a cross-validation procedure to estimate how predictive that NLP embedding is of brain activity in each sensor-timepoint i. For each fold:

- The MEG data $Y_{(i,\tau)}$ and feature matrix $X = x_1, x_2, \ldots x_n$ are split into corresponding train and validation matrices and these matrices are individually normalized (to get a mean of 0 and standard deviation of 1 for each sensor-timepoint across words), ending with train matrices $Y_{(i,\tau)}^R$ and X^R and validation matrices $Y_{(i,\tau)}^V$ and Z^V.

- Using the train fold, a model $\theta_{(i,\tau)}$ is estimated as:

$$\arg\min_{\theta_{(i,\tau)}} \|y_{(i,\tau)}^R - X^R\theta_{(i,\tau)}\|_2^2 + \lambda_{(i,\tau)}\|\theta_{(i,\tau)}\|_2^2.$$

  A ten-fold nested cross-validation procedure is first used to identify the best $\lambda_{(i,\tau)}$ for every sensor-timepoint tuple (i, τ) that minimizes the nested cross-validation error. $\theta_{(i,\tau)}$ is then estimated using $\lambda_{(i,\tau)}$ on the entire training fold.

- The predictions for each sensor-timepoint tuple (i, τ) on the validation fold are obtained as $\widehat{Y}_{(i,\tau)}^V = X^V\theta_{(i,\tau)}$.

The above steps are repeated for each of the cross-validation folds and an average evaluation metric value is obtained for each sensor-timepoint tuple (i, τ), each NLP embedding, and each participant.

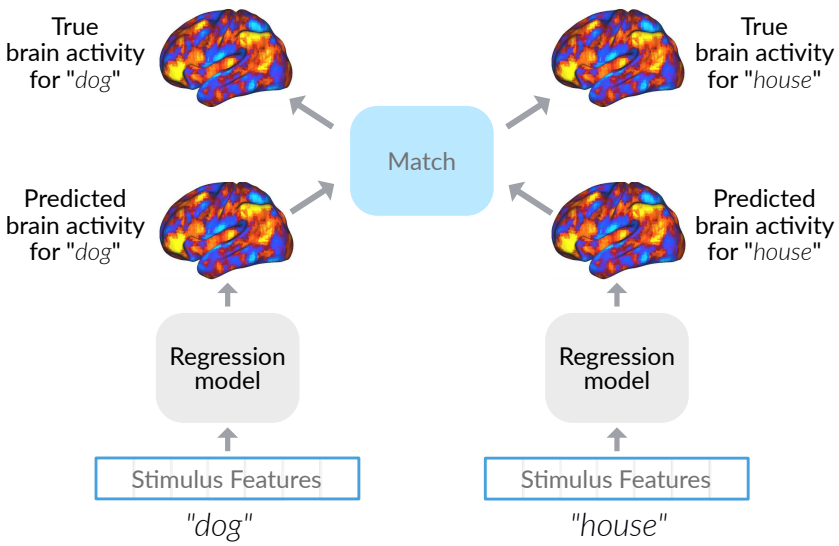


Figure 3.1: 2v2 accuracy evaluation metric. The predictions for two repetitions, $\hat{b}_1$ and $\hat{b}_2$, are being matched to their corresponding true brain activities, b_1 and b_2. The match is performed based on the distances between each of the predictions, to each of the true brain activities: $score_1 = dist(\hat{b}_1, b_1) + dist(\hat{b}_2, b_2)$ and $score_2 = dist(\hat{b}_1, b_2) + dist(\hat{b}_2, b_1)$. If $score_1$ ¡ $score_2$, we match $\hat{b}_1 \leftrightarrow b_1$ and $\hat{b}_2 \leftrightarrow b_2$ (correct match, obtaining an accuracy of 1.0), otherwise we match $\hat{b}_1 \leftrightarrow b_2$ and $\hat{b}_2 \leftrightarrow b_1$ (wrong match, accuracy 0.0). The distance used in our experiments is cosine distance.

3.3.4 Evaluation metrics

Pearson correlation. One common evaluation metric for linear encoding models is Pearson correlation, which is computed between the held-out brain recordings and the corresponding predictions in the cross-validation setting. We compute one correlation value for each of the cross-validation folds and report the average value as the final encoding model performance.

2v2 accuracy. Another evaluation metric is the 2v2 accuracy, which was introduced in Mitchell, Shinkareva, et al. (2008) and is illustrated in Figure 3.1. This metric measures whether a pair of brain predictions are closer in distance (e.g. cosine distance) to their corresponding true brain recording or to the foil brain recording in the pair. If the predictions are closer to their true recordings than to the foil recordings, the 2v2 accuracy is 1. Otherwise, the 2v2 accuracy is 0. Under this metric, the theoretical chance performance is 0.5. In this thesis, we use cosine distance as the distance metric for calculating 2v2 accuracy.

20v20 accuracy. The last evaluation metric is 20v20 accuracy, which is a generalization of the 2v2 metric described in Section 3.3.4. This metric was proposed by Wehbe, Vaswani, et al. (2014) as a way to boost the signal-to-noise ratio in estimating encoding models for single-trial data. The Harry Potter data is entirely single-trial (i.e. the participants read the book chapter only once and so each word appears only once in its context), and so Wehbe, Murphy, et al. (2014) and

Wehbe, Vaswani, et al. (2014) have observed that increasing the number of brain recordings that are classified together from 2 to 20 improves the classification accuracy for this dataset. We follow these works in using this evaluation metric in some of the work presented in this thesis.

3.4 Improved scientific inference for encoding models of complex stimuli

Linear encoding models have revealed important properties of information processing in the brain, such as that the representation of concepts is distributed across cortex but consistent across people (Mitchell, Shinkareva, et al., 2008; Kay et al., 2008; Nishimoto et al., 2011b; Huth, Nishimoto, et al., 2012; Wehbe, Murphy, et al., 2014; Huth, Heer, et al., 2016). In these cases, researchers used interpretable representations of the stimuli to link specific information with the location where this information is processed in the brain (i.e. a concrete noun stimulus was represented as its co-occurrence with a set of verbs in a large text corpora (Mitchell, Shinkareva, et al., 2008)).

More recently, researchers have begun using representations of stimuli extracted from pre-trained neural networks. Through training on large amounts of data, neural networks learn to encode different types of information about these inputs (i.e. stimuli) in order to perform a specific task (e.g. language modeling, object classification, etc.). For example, language models have been shown to contain information about part of speech (Liu et al., 2019) and semantic roles (Tenney, Das, et al., 2019). While these complex stimuli representations are able to predict brain measurements to an unprecedented extent (Yamins et al., 2014; Wehbe, Vaswani, et al., 2014; Jain et al., 2018; Toneva and Wehbe, 2019; Wang, Wehbe, et al., 2019; Schrimpf et al., 2020; Caucheteux, Gramfort, et al., 2021b; Goldstein et al., 2021; Cross et al., 2021), their complexity also makes it more difficult to make scientific inferences about what specific information is processed where in the brain. This inference becomes even more difficult in an experimental setting where participants observe naturalistic stimuli (e.g. watching movies, reading books, listening to stories), which is becoming increasingly more popular in neuroscience (Sonkusare et al., 2019; Nastase, Goldstein, et al., 2020; Hamilton et al., 2020). This naturalistic setting enables studying processing that is more easily generalizable to the everyday world, but it further complicates brain mapping because there is less control over what information the brain is actually processing about the stimulus.

As we move towards more complex stimuli representations and more complex experimental settings, it becomes important to reexamine what inferences we're able to make from our existing computational tools and adjust our toolbox accordingly. In this chapter, we first discuss the most frequently used computational tool for brain mapping–the linear encoding model–and propose two new tools that, when used together, can allow us to make stronger scientific inferences.

Motivating example. A simple conceptual example illustrated in Fig. 3.2 (A-C) motivates the need to reexamine the interpretation of encoding model performance when using complex stimuli representations. We present three different cases for the underlying relationships between two brain measurements shown in yellow and in blue, the stimulus corresponding to these measurements, and the stimulus representation that is used to train an encoding model. In the first case (Fig. 3.2A)

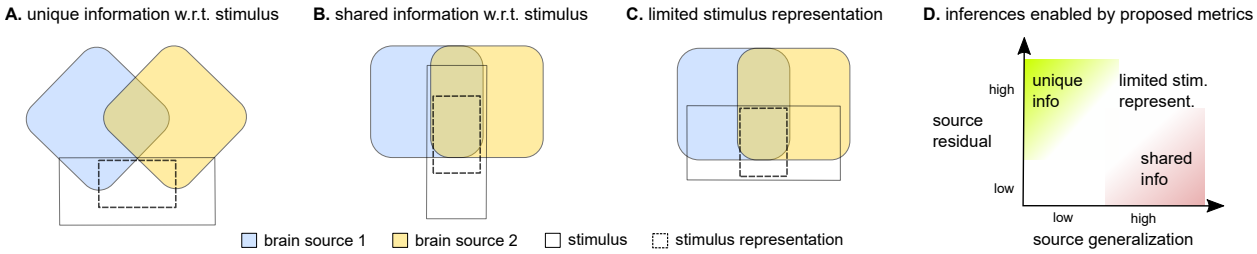


Figure 3.2: Diagrams representing different cases for the underlying shared variance between two brain measurements, the presented stimulus, and the stimulus representation (A-C), and how our proposed metrics enable us to infer these relationships (D). In contrast, in each case an encoding model will predict a similar proportion of variance in both brain measurement sources, making it difficult to disambiguate the three cases.

the two brain measurement sources process unique stimulus information. For instance, the blue source may capture the part of speech of a stimulus word and the yellow source may capture some of its semantic properties (e.g. manipulability and size). The stimulus representation also captures these multiple types of information (e.g. the stimulus representation may be derived from a language model, such as ELMo (Peters et al., 2018)), so an encoding model using this stimulus representation would predict a proportion of the variance in both sources. In this case, the encoding model performance would mislead us to think that the two brain sources process the same information about the word stimulus because the same stimulus representation predicts significant proportions of the two brain sources.

In the second and third cases (Fig. 3.2B-C) the two brain sources share information with respect to the stimulus, as indicated by the common overlap between the yellow and blue brain sources and the large stimulus rectangle. However, in Fig. 3.2B each brain source captures little unique stimulus information, while in Fig. 3.2C each source captures a lot of unique information. In both of these cases, using the same stimulus representation as input to an encoding model would predict both brain sources to a similar degree, which limits our ability to disentangle the two possible cases. This limitation is due to the limited stimulus representation in case C, which only captures aspects of the shared information. This is an issue because we can only make the claim that a set of regions is processing the same stimulus information in the case that we are able to disentangle Case B from the rest. To address these limitations, we propose two new metrics that, when used together, can disambiguate each of the three cases, as discussed more in depth in Section 3.4.2.

Real-world example of inference problem. One case study of this inference problem from the neuroscience literature is a set of findings using encoding models during naturalistic language comprehension. Several researchers have found that the activity in a wide set of bilateral regions in the temporal and prefrontal cortices deemed as the language network (Fedorenko, Hsieh, et al., 2010) can be significantly predicted as a function of various features of the presented language stimuli (Wehbe, Murphy, et al., 2014; Reddy et al., 2020; Caucheteux, Gramfort, et al., 2021a), without distinguishing between these regions in terms of the information map. Huth, Heer, et al. (2016) and Deniz et al. (2019) do show that the set of regions predicted by word meaning are tuned to different aspects of meaning, however, the set of regions in the language network are

shown to be tuned to the same aspects of meaning. A big outstanding question from these works is whether these language regions are all indeed processing the same information about the stimulus, or whether the tools we are using (e.g. stimulus representations, encoding models) and the way we are using them (e.g. by interpreting the encoding model performance) is preventing us from differentiating between them.

In the remainder of the chapter, we first highlight limitations of existing computational approaches for this scientific inference using simulated data with known ground truth relationships. We then propose two new metrics and an analysis framework that, when used together, can allow us to make stronger scientific inferences that can disambiguate all three cases. We term these metrics *source generalization* and *source residuals*. We next use our proposed tools to analyze two fMRI datasets obtained using naturalistic stimuli and show striking consistencies between the two datasets. In this chapter, we present only the results from one of these datasets–Courtois Neuromod. We present the results for the second dataset in Appendix A.

3.4.1 Related Work

Encoding models are becoming popular as datasets with complex stimuli become more common and stimulus representations from neural networks are employed as a tool to study the brain. Several works have investigated the expressivity of encoding models. Wu et al. (2006) frames brain mapping as system identification, where the neuroscientist is attempting to discover the features that different parts of the system are sensitive to. Wu et al. (2006) also describes a methodology for building encoding models and measuring the performance ceiling.

Some work has described the utility of encoding models as opposed to decoding models, in making precise inferences about representations (Naselaris et al., 2011; Weichwald et al., 2015) or have proposed ways to make decoding models more interpretable (Haufe et al., 2014). Some work has used some version of variance partitioning to differentiate between *feature spaces* (Toneva, Mitchell, et al., 2020; Caucheteux, Gramfort, et al., 2021a; Reddy et al., 2020; Heer et al., 2017; Lescroart et al., 2019). However, here we focus on encoding models with the *same feature space* at *different sources* (e.g. different voxels, different neurons, different regions, or different sensors). Some work has relied on the correlation between voxels to enforce priors on the learned models (Nunez-Elizalde et al., 2019; Wehbe, Ramdas, et al., 2015). Other than pooling information across voxels to regularize encoding models, most work in relating the information in different voxels has been in the functional connectivity literature (Van Den Heuvel et al., 2010). In that large body of work, there is not usually a relationship drawn between the stimulus and the brain activity. Instead, the activity between different sources is correlated, often during rest, and the correlation used as a metric for functional connectivity. Some work has used inter-subject correlation to identify which regions are related to a stimulus (Hasson et al., 2004; Simony et al., 2016). However, we focus on understanding the relationship between regions and the stimulus.

Because of the high temporal resolution of certain recording tools like magnetoencephalography (MEG), it is possible to compare the representations across time points. A powerful method called temporal generalization has been proposed by King et al. (2014). It allows researchers to see if the representation at a given point in time is similar to another point in time, and have been subsequently used in multiple works (Blanco-Elorrieta et al., 2017; Hebart et al., 2018; Fyshe, Su-

dre, et al., 2019; Fyshe, 2020). The generalization idea was recently adapted by Toneva, Mitchell, et al., 2020 into spatial generalization, in which the representation at different voxels are compared. This is done by using the predictions at one voxel to predict the activity in another voxel. These methods can be seen as specific instantiations of the *source generalization* metric that we describe. We further propose a new way to normalize this source generalization metric which improves its interpretability.

Another approach is to compare encoding model weights, which has typically been done after reducing the dimensionality across the brain and plotting the resulting low dimensional projections on the brain (Huth, Nishimoto, et al., 2012; Huth, Heer, et al., 2016; Deniz et al., 2019). In Cukur et al. (2013), different encoding models are estimated for the same participant under different attention conditions, and the tuning change due to attention is estimated from the weights of the different models.

3.4.2 Definitions

In this section, we define our two proposed metrics and other existing metrics that we commonly reference throughout this work. To ground these definitions, we relate each metric to an underlying setting for how a brain measurement is assumed to be generated. We introduce three such settings that are increasingly more specific about the dependence between one brain source and other sources in the same participant and other participants. Each setting introduces an additional set of assumptions, which enable us to relate the metrics to the unique and shared information in each brain source.

Single brain source, single participant

In the first setting, the activity in a single brain source that is recorded from one participant is assumed to be generated as a function of a specific representation of the presented stimulus. More concretely,

$$Y = g(X) + \epsilon \tag{3.2}$$

where $Y \in \mathbb{R}$ corresponds to the observation at a single brain source (e.g. fMRI voxel, EEG/MEG sensor-timepoint, electrode), $X \in \mathbb{R}^d$ to the d-dimensional numerical representation of the corresponding stimulus (e.g. a word embedding obtained from inputting a word stimulus in a language model), and $\epsilon \in \mathbb{R}$ to a noise term that is independent from the stimulus representations. The first setting is the one that commonly underlies encoding models. For a set of predictions $\widehat{Y}$ on heldout data, the encoding model performance is defined as follows:

$$\texttt{encoding model performance}(\widehat{Y}, Y) = \text{corr}(\widehat{Y}, Y)\,,$$

where $\text{corr}(\cdot)$ signifies Pearson correlation.

Multiple brain sources, single participant

In the second setting, we explicitly allow for a noise term reflecting shared information between two brain sources Y_1 and Y_2 that is not captured by the stimulus representation. In our modified setting,

$$Y_i = \underbrace{g_i(X)}_{\text{signal}} + \underbrace{\epsilon_i}_{\text{individual noise}} + \underbrace{\epsilon_{12}}_{\text{shared noise}} \tag{3.3}$$

for $i \in \{1, 2\}$ where $g_i(X) = \langle X, \theta_i \rangle$ are linear functions of stimulus representations $X \in \mathbb{R}^d$ with parameters $\theta_i \in \mathbb{R}^d$ and $\epsilon_1, \epsilon_2, \epsilon_{12}$ are noise terms. All noise terms are independent of each other and X and have zero mean and variances σ_1^2, σ_2^2 and σ_{12}^2, respectively.

Functional Connectivity. Functional connectivity measures the correlation of two brain sources Y_1 and Y_2 through time. In the setting above, this evaluates to

$$\text{corr}(Y_1, Y_2) = \frac{\text{cov}(Y_1, Y_2)}{\sqrt{\mathbb{V}(Y_1)\mathbb{V}(Y_2)}} = \frac{\text{cov}(g_1(X), g_2(X)) + \sigma_{12}^2}{\sqrt{\mathbb{V}(Y_1)\mathbb{V}(Y_2)}} \,,$$

where $\mathbb{V}(Y_i) = \mathbb{V}(g_i(X)) + \sigma_i^2 + \sigma_{12}^2$ is the variance of each observation. Note that $\text{corr}(Y_1, Y_2)$ does not distinguish between stimulus-related and stimulus-independent components of Y_i. That is, a correlation of Y_1 and Y_2 can be high if either $\text{cov}(g_1(X), g_2(X))$ or σ_{12}^2 is large. We support this observation using simulated brain source data in Section 3.4.3.

First proposed metric: source generalization. Disambiguating the three possible cases presented in Fig. 3.2A-C is key to conclusively inferring whether two sources process the same information about a stimulus. As outlined in Section 3.4, encoding performance is not able to disambiguate any of these three cases. To address some of these limitations, we propose our first metric *source generalization*. Intuitively, source generalization captures the amount of information shared by two sources and the stimulus representation used in an encoding model. Concretely, we define source generalization as:

$$\texttt{source generalization}(Y_1, Y_2) = \text{corr}(\widehat{Y}_1, Y_2) \,, \tag{3.4}$$

where $\widehat{Y}_1$ is a prediction on heldout data obtained by an encoding model as defined in Section 3.4.2.

Source generalization helps to disentangle the case shown in Fig. 3.2A from the cases in B and C. The source generalization will be low in case A because there is no shared information between the two brain sources and the stimulus representation, so training an encoding model on brain source 1 will not generalize to brain source 2, and vice versa. In contrast, the source generalization will be high in cases B and C because there is shared information between the two brain sources and the stimulus representation. However, source generalization is not able to disambiguate case B from case C.

Multiple brain sources, multiple participants

To disambiguate case B from case C, we consider our third setting which allows for multiple participants $P \in \{A, B\}$ and for each quantity in Equation (3.3) to differ in each subject P:

$$Y_{i,P} = \underbrace{g_{i,P}(X)}_{\text{signal of stimulus representation}} + \underbrace{h_i(Z)}_{\text{signal of } Z} + \underbrace{\epsilon_{i,P}}_{\text{individual noise}} + \underbrace{\epsilon_{12,P}}_{\text{joint noise}} . \tag{3.5}$$

We additionally allow for the presence of an additive component $h_i(Z)$ which makes explicit a dependence on features Z that are part of the stimulus, but are not captured by the stimulus representation X. For example, for a movie stimulus, X may correspond to representations of the speech in the movie that were obtained using an NLP system, and Z may represent whether or not there was a face present on the screen which is not information that is available in the NLP system representations. This enables us to express the likely common occurrence that a specific stimulus representation does not perfectly reflect all brian-relevant stimulus information.

Intersubject correlation. Intersubject correlation uses data from multiple participants who observe the same stimulus to identify regions that are correlated between participants and thus considered to be involved in processing the stimulus. More concretely,

$$\texttt{intersubject correlation}(Y_{1,A}, Y_{1,B}) = \text{corr}(Y_{1,A}, Y_{1,B}) \, , \tag{3.6}$$

where $Y_{1,A}$ and $Y_{1,B}$ are measurements of the same brain source across two participants A and B.

Second proposed metric: source residuals. While source generalization is able to disentangle case A in Fig.3.2 from cases B and C, it is unable to disentangle case B from case C. To address this limitation, we introduce the second metric–*source residuals*. Here, we build on the intuition behind intersubject correlation to estimate how much of the information that is shared between two brain sources and the stimulus is *not* shared between the two brain sources. More concretely,

$$\texttt{source residual}(Y_1, Y_2) = \text{corr}(R_{1-2,A}, R_{1-2,B}), \tag{3.7}$$

where $R_{1-2,P} = Y_{1,P} - \text{corr}(Y_{1,P}, Y_{2,P})Y_{2,P}$ is the residual of regressing $Y_{2,P}$ from $Y_{1,P}$. Source residuals disentangle case B from case A and C. Thus, we argue that using both proposed metrics together can help disentangle each case from the others.

3.4.3 Simulations

Using simulated data, we look at what encoding model performance, functional connectivity, source generalization and source residuals can tell us about the underlying relationships between two sources.

Simulating stimulus information. We generate two components that together make up all available stimulus information: $X \in \mathbb{R}^d$, which is the stimulus representation, and $Z \in \mathbb{R}^d$, which contains the remaining stimulus information not in X. A key simplifying assumption we make when generating X and Z data is that both components can be decomposed into four disjoint independent subsets of stimulus information: unique information captured by the individual brain sources (X_1, X_2, Z_1, Z_2), joint information captured by both brain sources (X_{12}, Z_{12}) and information not captured by either brain source (X_3, Z_3). Each X_i, Z_i, of length $\frac{d}{4}$, is independently sampled from a multivariate normal with mean 0 and a symmetric toeplitz covariance matrix with diagonal elements equal to 1, and X and Z are constructed by concatenating their four corresponding sub-components.

Simulating brain source data. We simulate observations at two brain sources from two distinct participants using the following data generation model (motivated by Eq. 3.5):

$$Y_{i,P} = \alpha \times \underbrace{g_{12,P}(X)}_{\text{joint signal}} + (1-\alpha) \times \underbrace{g_{i,P}(X)}_{\text{unique signal}} + \alpha \times \underbrace{N_{i,P}}_{\text{unique noise}} + (1-\alpha) \times \underbrace{N_{12,P}}_{\text{joint noise}} \tag{3.8}$$

$$\text{where } N_{i,P} = \delta \times h_{i,P}(Z) + (1-\delta) \times \epsilon_{i,P} \text{ and } N_{12,P} = \delta \times h_{12,P}(Z) + (1-\delta) \times \epsilon_{12,P}.$$

Here, each $g_{i,P}(X) = \langle \theta_{i,P}, X_i \rangle$ is a linear function of the stimulus representation that only looks at the corresponding X_i in X. In order to generate the necessary participant-specific parameters $\theta_{i,P} \in \mathbb{R}^{\frac{d}{4}}$, we first generate $\theta_i \in \mathbb{R}^{\frac{d}{4}}$ by independently sampling each of its components from a uniform distribution over $[0, 1)$. Each $\theta_{i,P}$ is then sampled from $\mathcal{N}(\theta_i, 0.25\mathbf{I})$ to allow for variation between participants. The same approach is used to generate each $h_{i,P}(Z) = \langle \phi_{i,P}, Z_i \rangle$ term. $\epsilon_1, \epsilon_2, \epsilon_{12} \in \mathbb{R}$ are terms that represent the information captured that is not related to the stimulus. Each ϵ_i is independently sampled from a standard normal distribution.

We introduce two adjustable parameters to simulate a wide range of scenarios. $\alpha \in [0, 1]$ controls how much of the stimulus representation related information captured by both brain sources is shared between them and how much of it is unique to each one. $\delta \in [0, 1]$ controls how much of the information captured by both brain sources that is unrelated to the stimulus representation is still related to the stimulus itself. It is important to note that the simulation results we present and analyze in this section are representative of an over-constrained setting. Their main purpose is to provide intuition for how the metrics discussed in Section 3.2 vary under different controlled scenarios.

Case A vs. Cases B & C. A key property that separates Case A from Cases B and C from Fig. 3.2 is the amount of unique stimulus information captured by each source that is also captured in the stimulus representation. We control this setting by keeping δ constant and varying α in Eq. 3.8. We use $\delta = 1.0$ here, but similar trends can be observed for any $\delta \in [0, 1]$. A low α simulates Case A as each brain source mostly captures unique stimulus information also present in the stimulus representation. A high α brings us closer to Cases B and C as this information is mostly shared between both brain sources for a given subject. We show the results in fig. 3.3 A. All metrics are collected and averaged across 1000 repetitions at each α. The encoding model performance, functional connectivity and source residuals remain relatively unchanged across the board. On the other hand, source generalization increases as we increase α. This suggests that looking at source

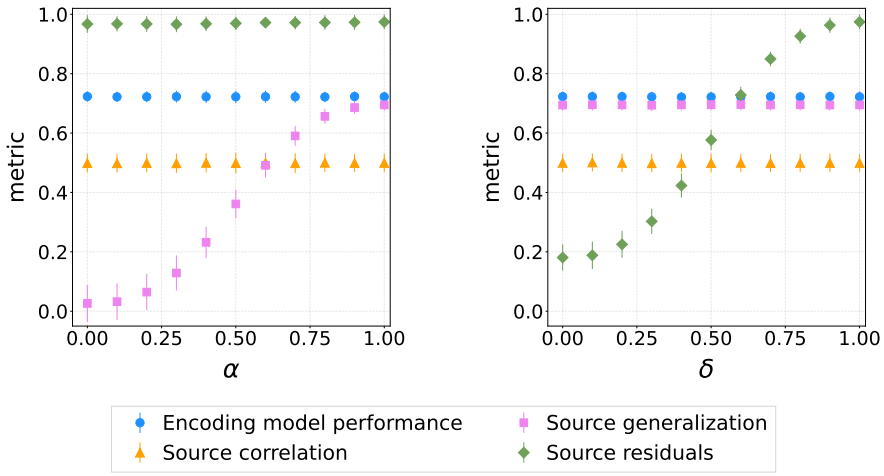


Figure 3.3: Plotting how each metric varies under simulations that separate (A) Case A from Cases B & C and (B) Case B from Case C.

generalization can allow us to recognize whether our data most resembles Case A or one of Cases B & C.

Case B vs. Case C. Cases B and C from Fig. 3.2 are similar in that the two sources share information with respect to the stimulus representation in both cases. However, they differ in the amount of unique information that each source captures about the stimulus. By setting $\alpha = 1.0$ and varying δ in Eq. 3.8, we can vary the amount of unique stimulus information that each source captures while preserving the amount of information that both sources share with respect to the stimulus representation. We show the results in fig. 3.3 B. At each value of δ, all metrics are again collected and averaged across 1000 repetitions. The encoding model performance, functional connectivity and source generalization do not vary with respect to the δ used in our simulations. However, we observe that the source residuals increase with δ. These results suggest that once we identify that we are in either Case B or Case C, looking at the source residuals can help us recognize which of these cases our data most resembles.

3.4.4 Empirical results using Courtois NeuroMod fMRI data

We compute the quantities of interest for two fMRI datasets obtained when participants viewed naturalistic stimuli. In this chapter, we present the results for one of these datasets–the Courtois NeuroMod fMRI data.

Data processing details. For each participant we downsample the fMRI data by averaging the voxel activities within the 268 functionally defined regions of interest (ROIs) from the Shen atlas per time frame (Shen et al., 2013; Finn et al., 2015), similarly to previous work (Rosenberg et al., 2015; Greene et al., 2018; Gao, Greene, et al., 2019; Doss et al., 2020). For each participant this results in a dataset of dimensions - number of TRs by 268 ROIs. These ROIs are entirely independent of our data as the Shen atlas was previously constructed from a separate group of healthy

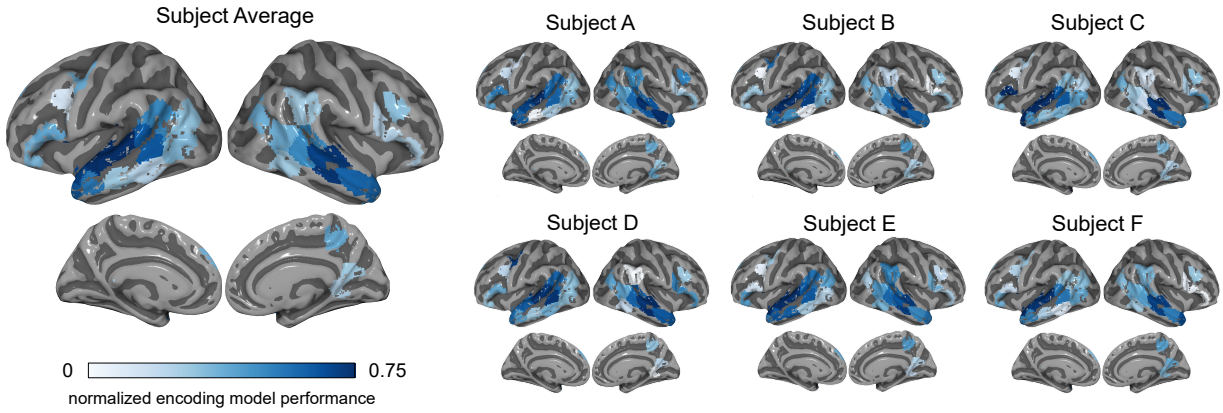


Figure 3.4: Encoding performance in 33 significantly predicted ROIs (corrected at level 0.05).

participants. The approach we propose apply to any brain regions. Because we are interested in studying naturalistic language comprehension, we chose to identify Shen atlas ROIs involved with processing language-relevant information. Regions of the brain involved with processing language-relevant information have previously been identified by (Fedorenko, Hsieh, et al., 2010; Binder et al., 2009) and are also entirely independent of our data. We consider a Shen atlas ROI to be a language ROI if $> 15\%$ of its voxels are within a region that processes language-relevant information. This procedure results in 55 Shen atlas ROIs that are language ROIs.

Stimulus representation. The approach we propose in this paper is general and can be applied to a wide variety of stimulus representations. Because we are specifically interested in studying the processing of language-relevant information, we use stimulus representations that capture the linguistic meaning of the stimuli. We follow previous work in neurolinguistics and obtain representations of the words observed by participants by feeding transcripts word-by-word into a pre-trained natural language processing model. We specifically choose ELMo (Peters et al., 2018), a bidirectional language model that incorporates multiple LSTM layers, for this purpose. Word representations obtained from the first hidden layer of ELMo, and contextualized with the previous 25 words, have been previously shown to significantly predict fMRI recordings of participants comprehending language (Toneva and Wehbe, 2019; Toneva, Mitchell, et al., 2020). We focus our analyses on representations similarly collected from the first hidden layer of ELMo when provided with chunks of 25 consecutive words, using the pretrained ELMo provided by Gardner et al. (2018).

Encoding model performance. We first investigate what we can learn about how language regions process audio-visual stimuli by interpreting encoding model performance. We estimate encoding models which predict the brain activity associated with matching video clips from an ELMo embedding of the speech in the video clip. One encoding model is estimated independently for each participant's ROI. The model is parameterized as a linear function regularized by the ridge penalty similarly to previous work (Sudre et al., 2012b; Wehbe, Vaswani, et al., 2014; Wehbe, Murphy, et al., 2014; Nishimoto et al., 2011b; Huth, Heer, et al., 2016; Toneva and Wehbe, 2019) and trained

with cross-validation. The regularization parameter is chosen by nested 10-fold cross-validation. We use 12-fold cross-validation for Courtois NeuroMod, which reflects the number of segments in which the dataset was originally collected. We evaluate the encoding model performance on heldout data for each fold.

In Figure 3.4, we present encoding model performances for the 33 language ROIs that were predicted significantly across participants in both fMRI datasets (one-sample t-test, FDR corrected for multiple comparisons across ROI at alpha level 0.05 (Benjamini et al., 1995)). The encoding model performances are normalized by the Intersubject Correlation (ISC), an estimate of the "noise ceiling" or the amount of variance in the ROI that is consistently related to the stimulus and therefore explainable (see Section 3.4.2 for a definition). We present the average normalized encoding performance across all participants in the datasets as well as for all participants individually. We observe that a set of bilateral language ROIs can be significantly predicted by the ELMo embedding for both datasets, at a group and individual participant level. These results replicate previous findings that the language ROI are well predicted by representations from ELMo (Toneva and Wehbe, 2019; Toneva, Mitchell, et al., 2020). We also observe similarly to prior work that these regions are predicted significantly by the encoding model, thereby making it difficult to distinguish between them (Huth, Heer, et al., 2016; Reddy et al., 2020; Caucheteux, Gramfort, et al., 2021a).

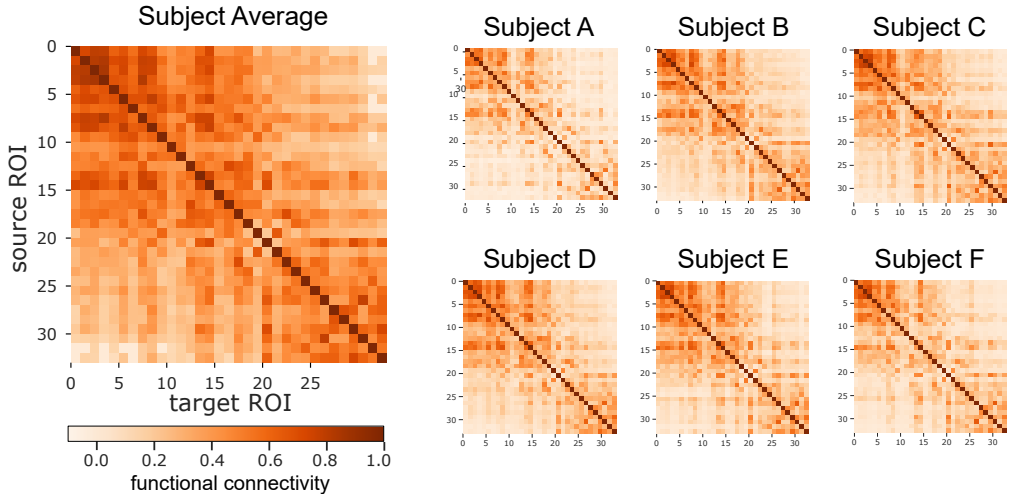


Figure 3.5: Function connectivity of the 33 language ROIs with significant encoding model performance. Non-significant values are presented as 0's (t-test, corrected for multiple comparisons across regions at alpha level 0.05)

Functional connectivity. Next we investigate whether functional connectivity can disambiguate the regions that were found to be significantly predicted by ELMo representations. We present the pairwise functional connectivity for the 33 language ROIs in both fMRI datasets in Figure 3.5. We only plot the pairwise correlation values found to be significant for each individual dataset (one-sample t-test, FDR corrected for multiple comparisons across ROI pairs at alpha level 0.05

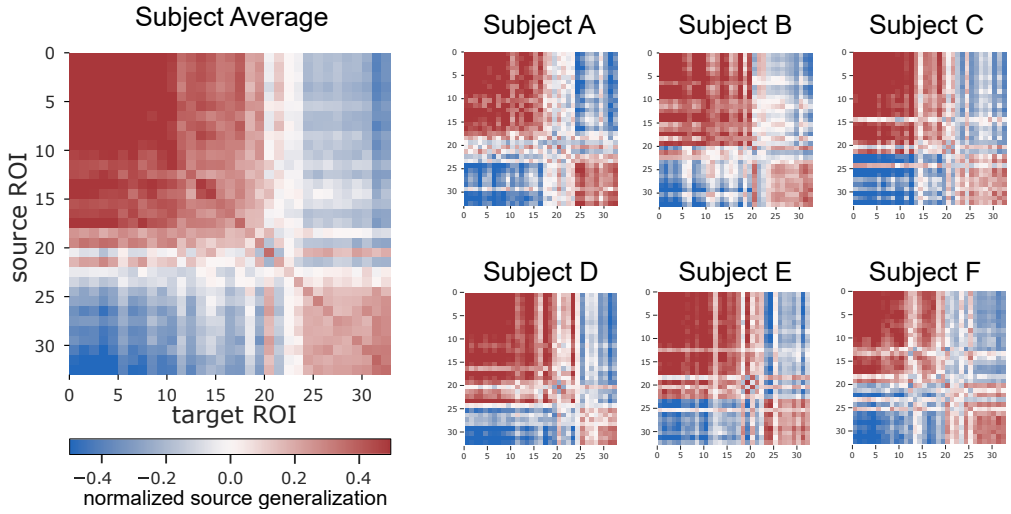


Figure 3.6: Source Generalization. ROI pairs with high norm. source generalization (red) process information captured by the stimulus representations in a similar way. Pairs with high norm. source generalization are consistent at the group and individual level in both datasets.

(Benjamini et al., 1995)). For both datasets, at the group and participant level we find that the amount of functional connectivity between language ROIs varies but the vast majority of ROI pairs have significant correlations. As observed in Figure 3.3 from the simulations, both high functional connectivity and high encoding model performance can be caused by multiple settings of the underlying shared information between the brain sources, stimulus, and stimulus representations. Since the overwhelming majority of ROI pairs have significant correlations and the individual ROI have significant encoding model performance, it is still not possible to disentangle the different cases even when combining these two metrics.

Source generalization. We further investigate the source generalization to understand if the differences in the amount of shared information between language ROIs are due to processing shared information related to the stimulus representations. We present the pairwise source generalization for the 33 language ROIs in both fMRI datasets in Figure 3.6. The source generalizations are normalized by the ISC, as an estimate of the "noise ceiling" (see Appendix for more details and suggestions for other types of normalization when investigating different scientific questions). In both datasets, at the group and participant level we find that there are differences in pairwise language ROI source generalizations. However, it is unclear if these differences are due to true differences among language ROIs in the amount of shared information related to the stimulus or to a limitation of using an ELMo embedding as our stimulus representation. Source generalization alone cannot distinguish if the relationship between two ROI is case B or C from Figure 3.2 with respect to the stimulus.

Source residuals. Next we investigate the second proposed metric, source residuals to understand the amount of unique stimulus-related information processed by language ROIs. We present

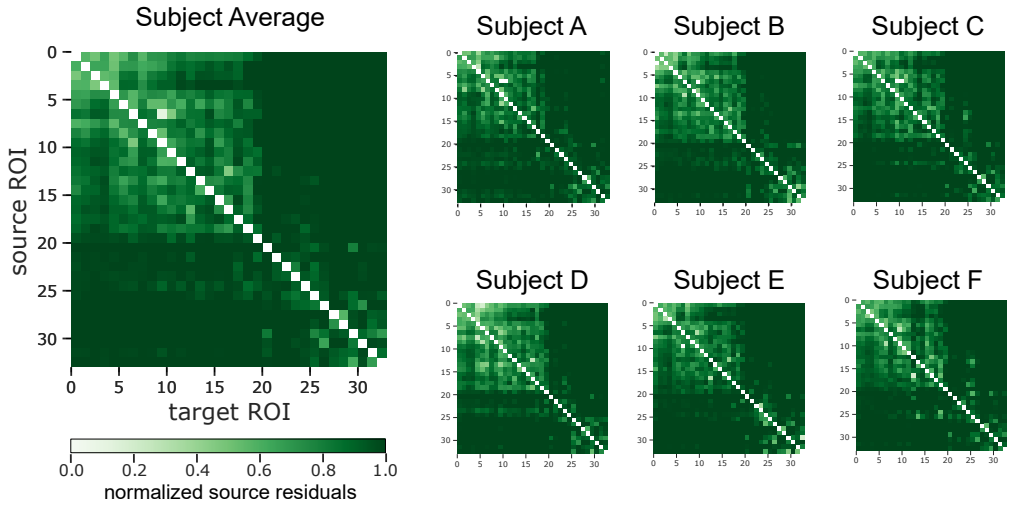


Figure 3.7: Source Residuals. ROI pairs with high norm. source residuals (dark green) are processing unique information related to the stimulus representations. These ROI pairs with high norm. source residuals are consistent at the group and individual level in both datasets.

the pairwise source residuals for the 33 language ROIs in both fMRI datasets in Figure 3.7. The source residuals are normalized by the ISC, an estimate of the "noise ceiling". In both datasets, we find differences in the source residuals at the group and participant level. The high source residuals reveal that the majority of regions process some unique information about the stimulus that cannot be fully accounted for by any of the other considered regions. However, this does not mean that the regions do not also process some shared information about the stimulus, which would have been removed during the residual computation (i.e. there will be high source residuals in both cases A and C in Fig. 3.2.

Source generalization and source residuals. We reexamine the case study presented in Section 3.4. As we are interested in answering whether language ROIs are indeed processing the same information about the stimulus or whether we cannot differentiate among the ROI due to methodological limitations, we focus on six bilateral language ROIs that have been previously shown to be difficult to disentangle (Reddy et al., 2020; Caucheteux, Gramfort, et al., 2021a). We present an example using the proposed framework to infer each of the three relationship cases (e.g. A, B, or C) within the six ROIs in Figure 3.8. These relationships are consistent across both datasets. The relationships between ROI pairs can be asymmetric (ie. ROI A could better generalize to ROI B than ROI B to ROI A), therefore we depict the relationships as a directed edges from the source ROI to the target ROI. The inferred relationship, case A, between ROI 1 and ROI 5 suggests that it is possible to differentiate the information processed between two ROI with significant encoding model performance. The inferred relationship, case B, between ROI 1 and ROI 2 support previous findings that bilateral language ROIs process shared information with respect to the stimulus. The inferred relationship, case C, between ROI 1 and ROI 3 shows that two ROI can process both shared and unique information. In this case the stimulus representation can limit our understanding

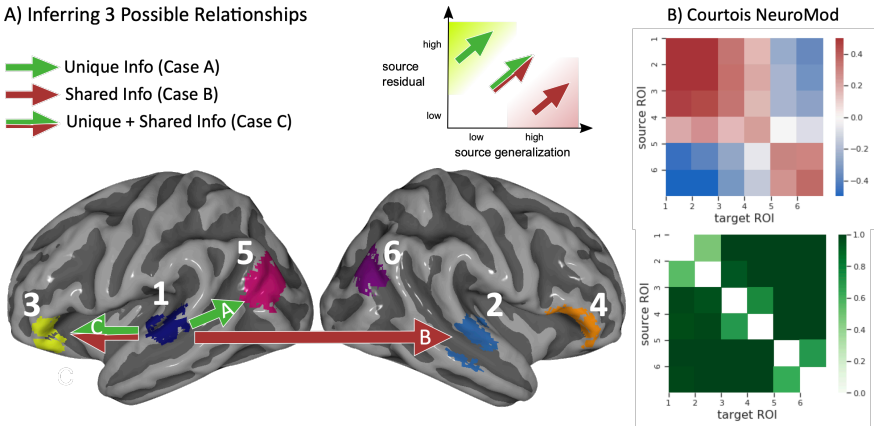


Figure 3.8: Source Generalization and Source Residuals. We use the proposed framework to infer an example of each of the three relationships between two brain sources, stimuli and stimuli representations.

of the amount of shared versus unique information processed between the two ROI.

3.4.5 Discussion

In contrast to the encoding model performance, our proposed metrics disentangle the three cases outlined in Fig.3.2 A-C. Note that in case C both source residuals and source generalization are high which indicates that the two brain sources process shared and unique information about the stimulus. We expect that a fully descriptive stimulus representation will capture this unique information that is shared between the stimulus and each individual brain source, so if both metrics are high we conclude that the stimulus representation used in the encoding model is not informative enough to disentangle the information processed in the two brain sources. This allows us to infer that we in fact need better stimulus representations rather than that the two brain sources are processing identical information about the stimulus. This interpretation suggests that Region 1 (middle superior temporal gyrus) in Fig. 3.8 and Region 3 (left inferior frontal gyrus) may be more easily distinguishable in the future using an encoding model if we have new stimulus representations that capture unique information that either region is processing about the stimuli. Our framework can be used as a test for future representations–if these future representations lead to a high encoding model performance and the source residuals between Region 1 and Region 3 continue to be high in the investigated stimulus set, but the source generalization using the new feature space decreases, then the new stimulus representation better captures some of the unique information processed by at least one of the regions.

Assumptions. Similarly to many previous works that use encoding models, we make an assumption that the fMRI recording can be modeled as a *linear* function of the stimulus representation. This linearity assumption may be limiting in the case that there are non-linear effects of the stimulus on the brain recordings that cannot be fully captured by the linear encoding model. However, non-linear encoding models may require significantly more brain recordings to train. As more brain recordings become publicly available, investigating relationships between stimuli and brain activity using non-linear encoding models may become more practical and should be studied in future work. Additionally, we assume that both the stimulus representations and the brain recordings are standardized to have mean of 0 and standard deviation of 1. We ensure that this assumption is valid by z-scoring both the stimulus representations and the brain recordings before we conduct further analyses.

Limitations and future work. One limitation of our proposed source residual metric is that the residuals contain information about the unique information in both regions that were used to compute the residuals. Isolating the residual information in an individual region may further improve our ability to disambiguate the information processed by different regions. We hope that our approach can serve as basis for such future work. Overall, our proposed framework is a promising new tool for computational neuroscientists who are interested in mapping information processing in the brain.

3.5 Takeaways

The contributions of this chapter can be summarized as follows:

- We conceptually breakdown the possible underlying relationships for the shared variance between two brain sources, the presented stimuli, and the selected stimulus representation.
- We present simulations that show limitations of commonly used methods for disambiguating these different relationships and propose two new metrics that can distinguish them.
- We showcase the use of these metrics in two fMRI datasets with naturalistic stimuli showing when we can disambiguate and when the feature set is the limitation. Our results generalize across these two datasets that capture different populations and are acquired by different labs in different countries with very different experimental setups and scanning parameters. Overall we present evidence that our proposed approach is a promising new tool for computational neuroscientists who are interested in mapping information processing in the brain.

Chapter 4

Modeling Processing of Context-Dependent Supra-Word Meaning

This chapter is based on work that is available as a preprint in:
Mariya Toneva, Tom M Mitchell, and Leila Wehbe. "Combining computational controls with natural text reveals new aspects of meaning composition". In: *bioRxiv* (2020). DOI: 10.1101/2020.09.28.316935. eprint: https://www.biorxiv.org/content/10.1101/2020.09.28.316935v2.full.pdf.

In this chapter, we provide an operational definition of a facet of composed meaning in language, that we term "supra-word meaning", by defining it as the multi-word meaning that is beyond the meaning of individual words (e.g. the composed meaning of "Mary finished the apple" includes the concept of eating which is not one of the composite words). We further propose a computational representation of supra-word meaning, using powerful recent neural network NLP models and an approach to disentangle individual-word from supra-word meaning in NLP embeddings. Using fMRI recordings, we reveal that hubs thought to process lexical-level meaning also maintain supra-word meaning, suggesting a common substrate for lexical and combinatorial semantics. However, surprisingly, we find that supra-word meaning is difficult to detect in MEG. Instead, the MEG recordings are significantly predicted by information that is unique to the individual recently-read words. The difference between the fMRI and MEG results suggests that the processing of supra-word meaning may be based on neural mechanisms that are not related to synchronized cell firing, as is the MEG signal. These results call for a more nuanced understanding of previous works that aim to study composition of sentence-level meaning using MEG as well as possibly other types of imaging modalities that rely on synchronized firing, such as EEG and ECoG. This finding also has consequences for building brain computer interfaces (BCI) to decode meaning from brain recordings. While the high temporal resolution of such devices makes them desirable, adopting them in real-world BCI may be difficult if we are not able to use them to decode composed meaning.

4.1 Introduction

Understanding language in the real-world requires us to compose the meaning of individual words in a way that makes the final composed product more meaningful than the string of isolated words. For example, we understand the statement that "Mary finished the apple" to mean that Mary finished *eating* the apple, even though "eating" is not explicitly specified (Pylkkänen, 2020). This *supra-word meaning*, or the product of meaning composition beyond the meaning of individual words, is at the core of language comprehension, and its neurobiological bases and processing mechanisms must be specified in the pursuit of a complete theory of language processing in the brain.

However, neuroscientists have instead investigated *correlates* of meaning composition that may fail to capture or isolate the supra-word meaning. In one line of research, neuroscientists follow classical neuroimaging approaches that consist of contrasting a condition of interest (e.g., semantic surprise, full sentences or erroneous sentences) with a control condition (e.g., no surprise, disconnected words or correct sentences). For example, they observe differences in brain recordings when processing an unexpected versus an expected word in a specific context (Kutas and Federmeier, 2011; Kuperberg et al., 2003; Kuperberg, 2007) or a monotonic increase in neural activity over the course of reading a sentence (Fedorenko, Scott, et al., 2016). Even though such studies have been pivotal in beginning to study the processes behind meaning composition, we argue that their findings are related to the process of integrating supra-word meaning, while missing other key components, such as the storage and maintenance of the current supra-word meaning. In a different line of work, neuroscientists build computational models of meaning through Natural Language Processing (NLP) embeddings of words and sentences (Mitchell, Shinkareva, et al., 2008; Sudre et al., 2012b; Wehbe, Murphy, et al., 2014; Wehbe, Vaswani, et al., 2014; Huth, Heer, et al., 2016; Jain et al., 2018; Toneva and Wehbe, 2019; Fyshe, Sudre, et al., 2019). Thanks to these studies, we are starting to uncover some properties of meaning representation, such as the fact that neural activity associated with single word meaning is distributed (Mitchell, Shinkareva, et al., 2008; Huth, Heer, et al., 2016). However, the neural substrates of composed meaning and the mechanism by which it is represented are still elusive and we are far from converging on a mechanistic, algorithmic understanding of meaning composition beyond individual words. One of the reasons for these limitations is the underlying correlations present in natural language, which limit the ability of researchers to make exact scientific inferences, since they lack the precise controls of traditional experiments.

In this work, we study the brain representation of supra-word meaning by using data from naturalistic reading in two neuroimaging modalities, and augmenting it with a control procedure. More formally, we define "supra-word meaning" as the composed meaning of a sequence of words that is not part of the corresponding bag-of-words, i.e., it is the new meaning formed by combining a sequence of words that is not included in the isolated meaning of those words. Types of supra-word meaning may include: 1) implied meaning (e.g. "Mary finished the apple" implies that Mary ate the apple), 2) a specific contextualized meaning of a word or phrase (e.g. "green banana" evokes the meaning of an unripe, rather than simply green-colored, banana) that can also distinguish between different senses of the same word (e.g. "play a game" versus "theater play"), and 3) the different meaning of two events that can be described with the same words but reversed semantic

roles (e.g. “John gives Mary an apple” and “Mary gives John an apple”).

We create a computational representation for this supra-word meaning, derived from recently developed natural language processing algorithms (Peters et al., 2018). We find that this representation of supra-word meaning predicts fMRI activity in the anterior and posterior temporal cortices, suggesting that these areas support the representation of composed meaning. The posterior temporal cortex is considered to be primarily a site for lexical (i.e. word-level) semantics (Hagoort, 2020; Hickok et al., 2007) so our finding that it also maintains supra-word meaning suggests a common substrate for lexical and combinatorial semantics. Furthermore, we find clusters of voxels in both the posterior and anterior temporal lobe that share a common representation of supra-word meaning, suggesting the two areas may be working together to maintain the supra-word meaning. We also find that it is very hard to detect the representation of supra-word meaning in MEG activity. MEG has been shown to reveal signatures of the *computations* involved in incorporating a word into a sentence (Halgren et al., 2002; Lyu et al., 2019), which are themselves a function of the composed meaning of the words seen so far. However, our results suggest that the sustained *representation* of the composed meaning may rely on neural mechanisms that do not lead to reliable MEG activity. This hypothesis calls for a more nuanced understanding of the body of literature on meaning composition and has important implications for the future of brain-computer interfaces.

4.2 Approach

We built on recent progress in NLP that has resulted in algorithms that can capture the meaning of words in a particular context. One such algorithm is ELMo (Peters et al., 2018), a powerful language model with a bi-directional Long Short-Term Memory (LSTM) architecture. ELMo estimates a *contextualized* embedding for a word by combining a *non-contextualized* fixed input vector for that word with the internal state of a forward LSTM (containing information from previous words) and a backward LSTM (containing information from future words). To capture information about word t, we used the input vector for word t. To capture information about the context preceding word t, we used the internal state of the forward LSTM computed at word $t-1$ (Fig. 1B). We did not include information from the backward LSTM, since it contains future words which have not yet been seen at time t.

To study supra-word meaning, the meaning that results from the composition of words should be isolated from the individual word meaning. ELMo’s context embeddings contain information about individual words (e.g., ’finished’, ’the’, and ’apple’ in the context ’finished the apple’) in addition to the implied supra-word meaning (e.g., eating) (Fig. 1C). We post-processed the context embeddings produced by ELMo to remove the contribution due to the context-independent meanings of individual words. We constructed a “residual context embedding” by removing the shared information between the context embedding and the meanings of the individual words (Fig. 1D).

Obtaining stimulus representations. We obtain two full representations for each word in the stimulus set (as opposed to residual representations, as described in the next paragraph) using the approach detailed in Chapter 3.2. Briefly, we obtain one non-contextualized word embedding and one contextualized word embedding with the context length set to 25 words. We use 25

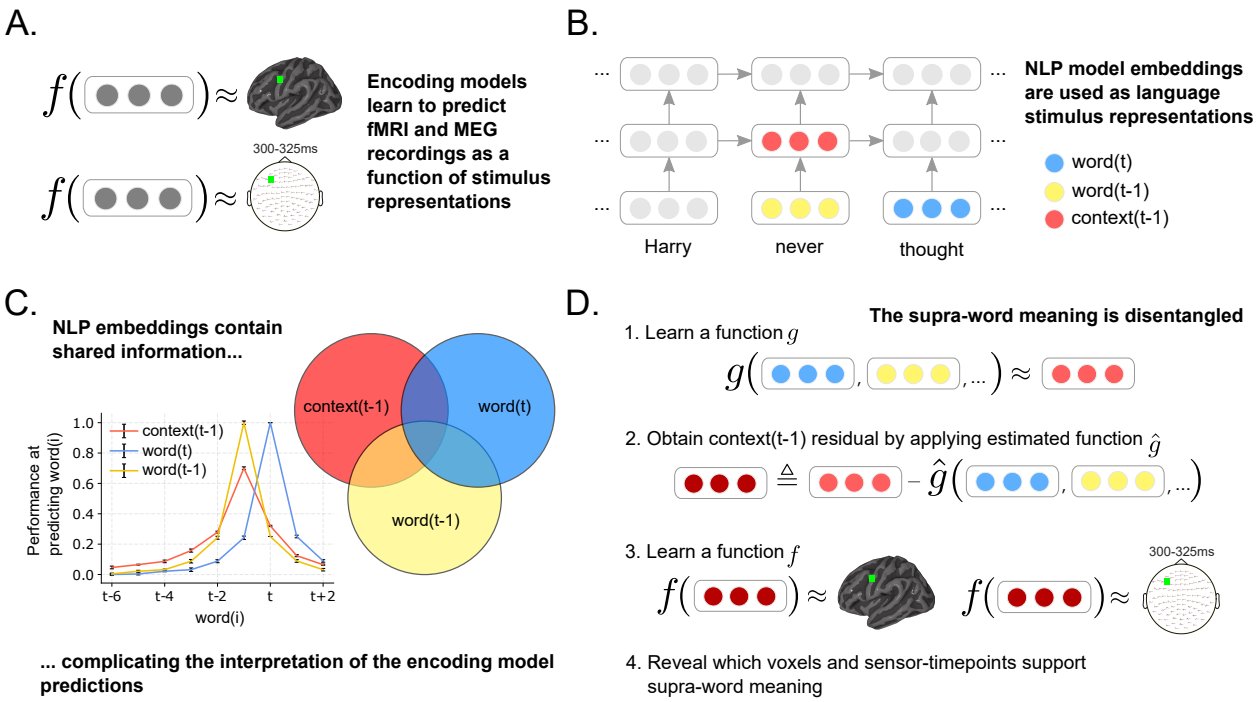


Figure 4.1: Approach. **(A)** An encoding model f learns to predict a brain recording as a function of computational representations of the text read by a participant during the experiment. A different function is learned for each voxel in fMRI and sensor-timepoint in MEG. **(B)** Computational representations of the stimulus are obtained from an NLP model that has captured language statistics from millions of documents. This model represents words using context-free embeddings (shown in yellow and blue) and context embeddings (shown in red). Context embeddings are obtained by continuously integrating each new word's context-free embedding with the most recent context embedding. **(C)** Context and word embeddings share information. The performance of the context and word embeddings at predicting the words at surrounding positions is plotted for different positions. The context embedding contains information about up to 6 past words, and word embeddings contains information about embeddings of surrounding words. To isolate the representation of supra-word meaning, it is necessary to account for this shared information. **(D)** Supra-word meaning is modeled by obtaining the residual information in the context embeddings after removing information related to the word embeddings. The supra-word meaning is used as an input to an encoding model f, revealing which fMRI voxels and MEG sensor-timepoints are modulated by supra-word meaning.

words to extract the context embedding because it has been previously shown that ELMo and other LSTMs appear to reduce the amount of information they maintain beyond $20 - 25$ words in the past (Khandelwal et al., 2018; Toneva and Wehbe, 2019).

Obtaining supra-word representations and other residual stimulus representations. We obtain three types of residual representations for each word at position t in the stimulus set: 1) residual context(t-1) embedding (i.e. supra-word meaning), 2) residual word(t-1) embedding, and 3) residual word(t) embedding. We compute all three types using the same general approach of training an encoding model described in Chapter 3.3 (i.e. ridge-regularized linear regression) but with inputs x_t and outputs y_t that change depending on the type of residual embedding. The steps to the general approach are the following, given an input x_t and output y_t:

1. Learn a linear function g that predicts each dimension of y_t as a linear combination of x_t. We follow the same steps outlined in the training of function f in the encoding model. Namely, we model g as a linear function, regularized by the ridge penalty. The model is trained via four-fold cross-validation and the regularization parameter is chosen via nested cross-validation.
2. Obtain the residual $y'_t \triangleq y_t - \hat{g}(x_t)$, using the estimate of the g function learned above. This is the final residual stimulus representation.

For the residual context(t-1) embedding, the input x_t is the concatenation of the full word embeddings for the 25 consecutive words $w_{t-24}, ..., w_t$ and the output y_t is the full context(t-1) embedding. For the residual word(t-1) embeddings, the input x_t is the concatenation of the full context(t-1) embedding and the full word embeddings for the 24 consecutive words $w_{t-24}, ..., w_t$ that exclude the full word embedding for word(t-1) and the output y_t is the full word(t-1) embedding. For the residual word(t) embeddings, the input x_t is the the concatenation of the full context(t-1) embedding and the full word embeddings for the 24 consecutive words $w_{t-24}, ..., w_{t-1}$ and the output y_t is the full word(t) embedding.

Encoding models of supra-word meaning. To investigate the neural substrates and temporal dynamics of supra-word meaning, we trained encoding models, as a function of supra-word meaning, to predict the brain recordings of nine fMRI participants and eight MEG participants as they read a chapter of a popular book in rapid serial visual presentation (see Chapter 3.3 for more details about the methods). The encoding models predict each fMRI voxel and MEG sensor-timepoint, from the text read by the participant up to that time point (Fig. 1A). The prediction performance of these models is tested by computing the Pearson correlation between the model predictions and the true held-out brain recordings.

Permutation tests. We evaluate the significance of the degree to which a single voxel or a sensor-timepoint is predicted by a specific encoding model using a standard permutation test. For encoding model predictions $\hat{Y}$ with corresponding true brain data Y, we repeat the following 1000 times:

1. permute the samples in $\hat{Y}$ in blocks of B samples, where $B = 5$TRs in fMRI (corresponding to 20 presented words) and $B = 20$ words in MEG. This block permutation is necessary in order to retain some of the auto-regressive structure in the brain recordings.

2. compute the encoding model performance between the permuted predictions and the test data: $\text{corr}(\hat{Y}_{perm}, Y)$

The p-value is computed as the proportion of the 1000 permutations when $\text{corr}(\hat{Y}_{perm}, Y) \geq \text{corr}(\hat{Y}, Y)$. The resulting p-values for all voxels/sensor-timepoints/time-windows are FDR corrected for multiple comparisons using the Benjamini-Hochberg procedure (Benjamini et al., 1995).

Chance proportions of ROI/timewindows predicted significantly. The proportion of an ROI that is significantly explained by an encoding model is defined as the proportion of all voxels in the ROI that are significantly explained by the encoding model as determined by a permutation test and a correction for multiple comparisons (as described above). However, even after a correction for multiple comparisons, we can expect that about 1% of voxels in a region will be significantly predicted by chance (this percentage depends on the alpha level of the significance test and on the type of multiple comparison correction). To test whether the predicted proportion of the ROI is significant, we contrast the proportion of the ROI/timewindow that is significantly predicted by the encoding model with a proportion of the ROI/timewindow that is significantly predicted by chance. We do this for all proportions of the same ROI/timewindow across participants, using a Wilcoxon signed-rank test. We compute the proportion of an ROI/timewindow that is significantly predicted by chance using the permutation tests described above. For each permutation k, we compute the p-value of each voxel in this permutation according to its performance with respect to the other permutations. Next for each ROI/timewindow, we compute the proportion of this ROI/timewindow with p-values< 0.01 after FDR correction, for each permutation. The final chance proportion of an ROI/time-window for a specific encoding model and participant is the average chance proportion across permutations.

Confidence intervals. We use an open-source package (Sheppard et al., 2020) to compute the 95% bias-corrected confidence intervals of the median proportions across participants. We use bias-corrected confidence intervals (Efron et al., 1994) to account for any possible bias in the sample median due to a small sample size or skewed distribution (Miller, 1988).

4.3 Results

4.3.1 Detecting regions that are predicted by supra-word meaning

To identify brain areas that represent supra-word meaning, we focus on the fMRI portion of the experiment. We find that many areas previously implicated in language-specific processing (Fedorenko, Hsieh, et al., 2010; Fedorenko and Thompson-Schill, 2014) and word semantics (Binder et al., 2009) are significantly predicted by the full context embeddings across subjects (voxel-level permutation test, Benjamini-Hochberg FDR control at 0.01 (Benjamini et al., 1995)). These areas include the bilateral posterior and anterior temporal cortices, angular gyri, inferior frontal gyri, posterior cingulate, and dorsomedial prefrontal cortex (Fig. 4.2A and Appendix Fig. B.1). A subset of these areas is also significantly predicted by residual context embeddings. To quantify these

observations, we select regions of interest (ROIs) based on the works above (Fedorenko, Hsieh, et al., 2010; Binder et al., 2009), using ROI masks that are entirely independent of our analyses and data. Full context embeddings predict a significant proportion of the voxels within each ROI across all 9 participants (Fig. 4.2B; ROI-level Wilcoxon signed-rank test, $p < 0.05$, Holm-Bonferroni correction (Holm, 1979)). In contrast, residual context embeddings predict a significant proportion of only the anterior and posterior temporal lobes. While the full context embedding is predictive of much of the fMRI recordings across the brain, the supra-word meaning is selectively predictive of two language regions - the anterior (ATL) and posterior temporal lobes (PTL).

Do the parts of the ATL and PTL that are predicted by supra-word meaning process the same information? To answer this question, we compute the source generalization metric introduced in Chapter 3.4 between all pairs of voxels in the ATL and PTL that are significantly predicted by supra-word meaning.

4.3.2 Source generalizations reveals two clusters of voxels with respect to processing supra-word meaning

We compute the source generalization (see Chapter 3.4) of voxel i to voxel j for every pair of voxels (i, j) in the ATL and PTL that are significantly predicted by supra-word meaning. We normalize the source generalization by dividing it by the encoding model performance in voxel i. The significance of the source generalization to voxel j is evaluated using a permutation test.

The source generalization matrices reveal that the PTL can be divided into two main clusters such that encoding models of supra-word meaning for voxels in one cluster can also generalize to other voxels in that cluster but not to voxels in the second cluster (Fig. 4.2C and Appendix Fig. B.2; voxel-level permutation test, Benjamini-Hochberg FDR controlled at level 0.01). Furthermore, the models of voxels within one of the PTL clusters, but not the other, significantly predict voxels in the ATL. The division of the PTL into two clusters, one of which is predictive of the ATL, can be observed within- (Fig. 4.2C, left), and across-participants (Fig. 4.2C, right). In contrast, the ATL voxels show only one cluster of voxels that are predictive both of other ATL voxels and also of PTL voxels (Appendix Fig. B.2). This pattern indicates that the organization of information in the ATL and parts of the PTL is shared and consistent across participants. To localize this shared representation, we visualize how well each ATL and PTL voxel predicts the other participants' ATLs (Fig. 4.2D and Appendix. Fig. B.3). ATL voxels are predictive of significant proportions of the ATL across participants, reinforcing the single cluster of ATL voxels observed in the source generalization matrices. Much of the left PTL predicts a significant proportion of the ATL across participants, whereas much of the right PTL does not (ROI-level Wilcoxon signed-rank test, $p < 0.05$, Holm-Bonferroni correction). The left PTL appears further subdivided, with a cluster of voxels in the posterior Superior Temporal Sulcus (pSTS) being more predictive. This suggests that the ATL and the left pSTS process a similar facet of supra-word meaning.

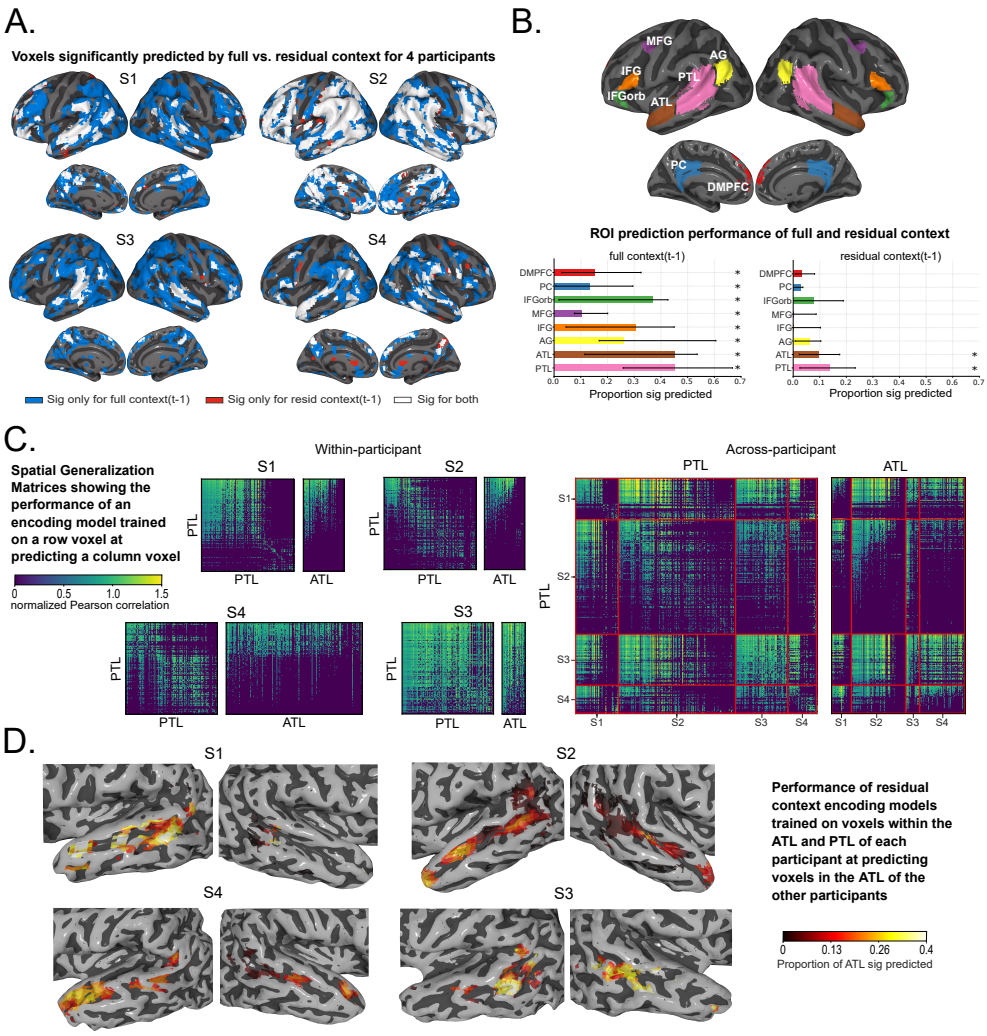


Figure 4.2: fMRI results. Visualizations for 4 of 9 participants with remainder available in Appendix Figures B.1-B.3. Voxel-level significance is FDR corrected at $\alpha = 0.01$. **(A)** Voxels significantly predicted by full-context embeddings (blue), residual-context embeddings (red), or both (white), visualized in MNI space. Most of the temporal cortex and IFG is predicted by full context embeddings, with residual context embeddings mostly predicting a subset of those areas. **(B)** ROI-level results. (Top) Language system ROIs (Fedorenko, Hsieh, et al., 2010) and two semantic ROIs (Binder et al., 2009). (Bottom) Proportion of ROI voxels significantly predicted by (Left) full context and (Right) residual context embeddings. Displayed are the median proportions across all participants and the medians' 95% confidence intervals. Full context predicts all ROIs (ROI-level Holm-Bonferroni correction, $p < 0.05$), while residual context predicts only bilateral ATL and PTL. **(C)** Source Generalization Matrices. Models trained to predict PTL voxels are used to predict PTL and ATL voxels (within-participant (Left), and across-participants (Right)). PTL cross-voxel correlations form two clusters: models that predict activity for voxels in one cluster can also predict activities of other voxels in the same cluster, but not activities for voxels in the other cluster. Across participants, only one of these clusters has voxels that predict ATL voxels. **(D)** Source generalization of supra-word encoding models trained on ATL and PTL voxels to other participants' ATL. All participants show a cluster of predictive voxels in the pSTS.

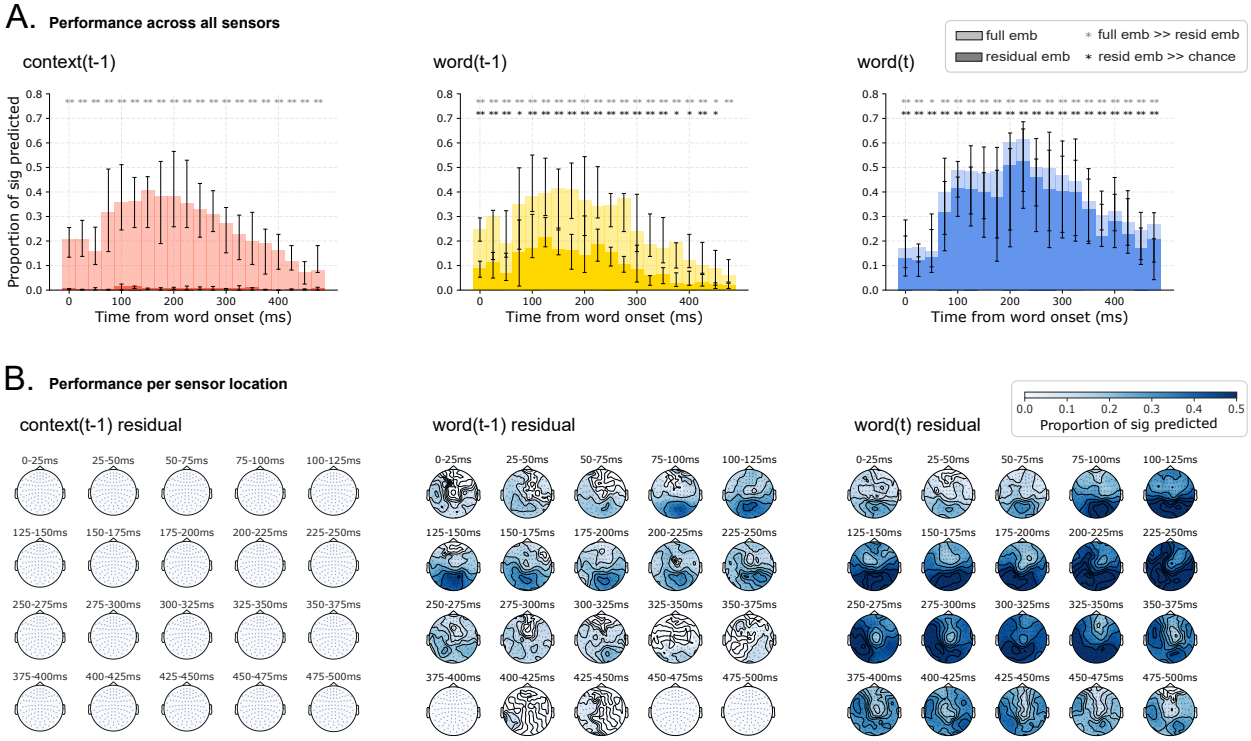


Figure 4.3: MEG prediction results at different spatial granularity. All subplots present the median across participants and errorbars signify the medians' 95% confidence intervals. **(A)** Proportion of sensors for each timepoint significantly predicted by the full and residual embeddings (visualized in lighter and darker colors respectively). Removing the shared information among the full current word, the previous word and the context embeddings results in a significant decrease in performance for all embeddings and lobes. The decrease in performance for the context embedding (left column) is the most drastic, with no timewindows being significantly different from chance for the residual context embedding. **(B)** Proportions of sensor neighborhoods significantly predicted by each residual embedding. Only the significant proportions are displayed (FDR corrected, $p < 0.05$). Context-residuals do not predict any sensor-timepoint neighborhood while both the previous and the current word residuals predict a large subset of sensor-timepoints, with performance peaks in occipital and temporal lobes.

4.3.3 The processing of supra-word meaning is invisible in MEG

To study the temporal dynamics of the emergence and representation of supra-word meaning, we turn to the MEG portion of the experiment (Fig. 3). We computed the proportion of sensors that are significantly predicted at different spatial granularity – the whole brain (Fig. 4.3A), by lobe subdivisions (Appendix Fig. B.4), and finally at each sensor neighborhood location (Fig. 4.3B; sensor-timepoint level permutation test, Benjamini-Hochberg FDR control at $\alpha = 0.01$). The full context embedding is significantly predictive of the recordings across all lobes (Fig. 4.3A, performance visualized in lighter colors; timepoint-level Wilcoxon signed-rank test, $p < 0.05$, Benjamini-Hochberg FDR correction). Surprisingly, we find that the residual context does not significantly predict any timepoint in the MEG recordings at any spatial granularity. This surprising finding leads to two conclusions. First, supra-word meaning is invisible in MEG. Second, what is instead salient in MEG recordings is information that is shared between the context and the individual words.

To understand the source of this salience, we investigated the relationship between the MEG recordings and the word embeddings for the currently-read and previously-read words. One approach to reveal this relationship is to train an encoding model as a function of the word embedding (Jain et al., 2018; Toneva and Wehbe, 2019). However, the word embedding corresponding to a word at position t is correlated with the surrounding word embeddings (Fig. 4.1C). Therefore, part of the prediction performance of the word t embedding may be due to processing related to previous words. To isolate processing that is exclusively related to an individual word, we constructed "residual word embeddings", following the approach of constructing the residual context embeddings (see Materials and Methods). We observe that the residual word embeddings for the current and previous words lead to significantly worse predictions of the MEG recordings, when compared to their corresponding full embeddings (Fig. 4.3A, middle and right panels; timepoint-level Wilcoxon signed-rank test, $p < 0.05$, Benjamini-Hochberg FDR correction). This indicates that a significant proportion of the activity predicted by the current and previous word embeddings is due to the shared information with surrounding word embeddings. Nonetheless, we find that the residual current word embedding is still significantly predictive of brain activity everywhere the full embeddings was predictive. This indicates that properties unique to the current word are well predictive of MEG recordings at all spatial granularity. The residual previous word embedding predicts fewer time windows significantly, particularly 350-500ms post word t onset. This indicates that the activity in the first 350ms when a word is on the screen is predicted by properties that are unique to the previous word. Taken together, these results suggest that the properties of recent words are the elements that are predictive of MEG recordings, and that MEG recordings do not reflect the supra-word meaning beyond these recent words.

Lastly, we directly compared how well each imaging modality can be predicted by each meaning embedding (Fig. 4.4). Residual embeddings predict fMRI and MEG with significantly different accuracy (Fig. 4.4A), with fMRI being significantly better predicted than MEG by the residual context, and MEG being significantly better predicted by the residual of the previous and current words (Wilcoxon rank-sum test, $p < 0.05$, Holm-Bonferroni correction). In contrast, the full context embeddings do not show a significant difference in predicting fMRI and MEG recordings(Fig. 4.4B). We further observe that the residual embeddings lead to an opposite pattern of prediction

in the two modalities(Fig. 4.4C). While the residual context predicts fMRI the best out of the three residual embeddings, it performs the worst out of the three at predicting MEG (Wilcoxon signed-rank test, $p < 0.05$, Holm-Bonferroni correction). In contrast, the full context and previous word embeddings do not show a significant difference in MEG prediction(Fig. 4.4D), suggesting that it is the removal of individual word information from the context embedding that leads to a significantly worse MEG prediction. These findings further suggest that fMRI and MEG reflect different aspects of language processing–while MEG recordings reflect processing related to the recent context, fMRI recordings capture the contextual meaning that is beyond the meaning of individual words.

4.4 Discussion

We enabled the investigation of emergent multi-word meaning, or *supra-word meaning*, in the brain by devising a computational representation of it that combines representations of natural text from recent neural network algorithms with a computational control that disentangles composed-from individual-word meaning. We investigated the spatial and temporal processing signatures of supra-word meaning by evaluating its ability to predict specific locations and timepoints of recorded brain activity via fMRI and MEG respectively.

We found that our devised supra-word meaning representation predicts fMRI recordings in the bilateral anterior and posterior temporal lobes (ATL and PTL). This finding supports some current hypotheses of language composition in the literature. Specifically, our results provide new evidence that the ATL processes composed meaning beyond simple concrete concepts, which supports the hypothesis that the ATL is a semantic integration hub (Visser et al., 2010; Pallier et al., 2011; Pylkkänen, 2020). Our results may also align with the hypothesis that the posterior superior temporal sulcus (pSTS, part of the PTL) is involved in building a type of supra-word meaning, by integrating information about the verb and its arguments with other syntactic information (Friederici, 2011; Frankland et al., 2015; Skeide et al., 2016). Further, our findings pose questions for the theory that posits left PTL as primarily a site of lexical (i.e. word-level) semantics, and left IFG as a hub of integrated contextual information (Hagoort, 2020). It also poses questions for the theory that combinatorial semantics are processed in the ATL while lexical semantics are processed in more posterior regions (Hickok et al., 2007). Our finding that the PTL maintains supra-word meaning indicates that the role of the PTL extends beyond word-level semantics and suggests a common substrate for lexical and combinatorial semantics. Further, we do not find evidence for supra-word meaning in left IFG, though this does not prove that left IFG does not represent supra-word meaning – the lack of significance may be due to low statistical power. Lastly, the finding that clusters of voxels in the PTL and ATL share a common representation of composed meaning suggests that the two areas may be working together to maintain the supra-word meaning.

Strikingly, we found that, even though our devised supra-meaning representation predicted a significant proportion of fMRI voxels, it did not significantly predict any sensor-timepoints in MEG. Instead, the MEG recordings were significantly predicted by information unique to both the currently-read and previously-read words. These findings suggest a difference in the underlying brain processes that fMRI and MEG capture. Indeed, while it is widely known that fMRI and MEG

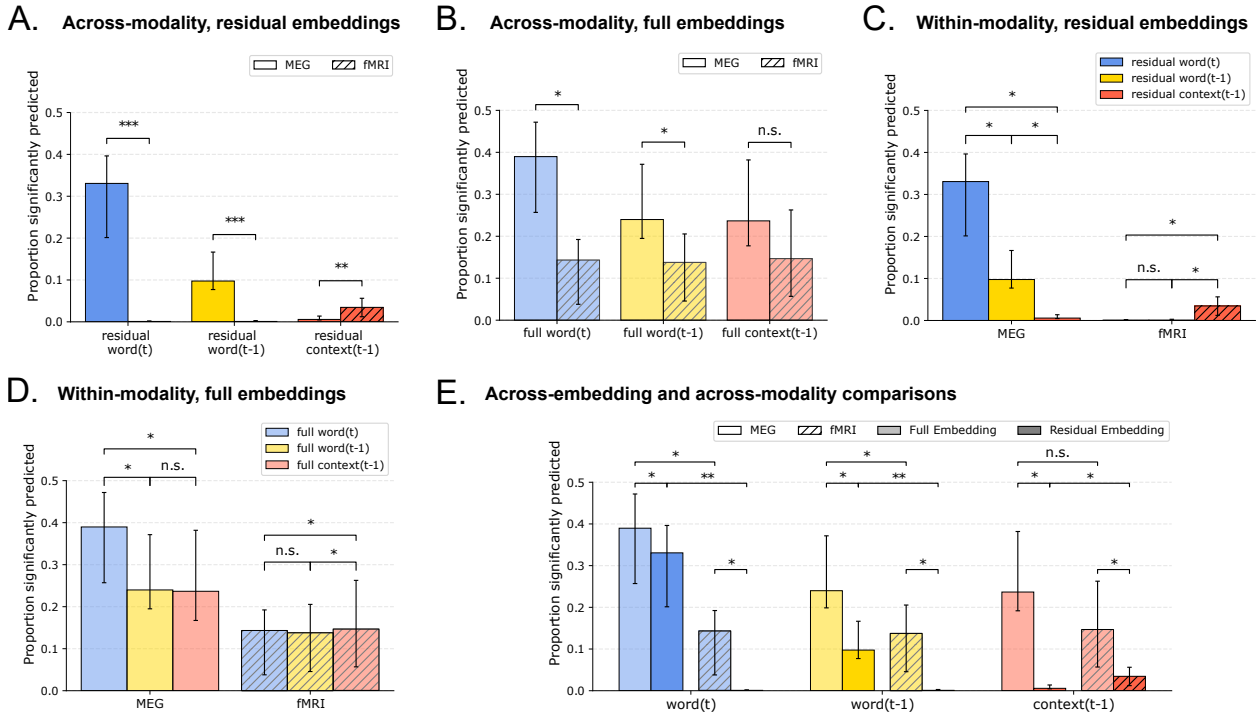


Figure 4.4: Direct comparisons of prediction performance of different meaning embeddings. Displayed are the median proportions across participants and the medians' 95% confidence intervals. Differences between modalities are tested for significance using a Wilcoxon rank-sums test. Differences within modality are tested using a Wilcoxon signed-rank test. All p-values are adjusted for multiple comparisons with the Holm-Bonferroni procedure at $\alpha = 0.05$. **(A)** Residual previous word, context, and current word embeddings predict fMRI and MEG with significant differences. **(B)** Full context embeddings do not predict fMRI and MEG with significant differences, while the full current word and previous word embeddings predict MEG significantly better than fMRI. **(C)** MEG and fMRI display a contrasting pattern of prediction by the residual embeddings. The current word residual best predicts MEG activity, significantly better than the previous word residual, which in turns predicts MEG significantly more than the context residual. In contrast, the context residuals significantly predict fMRI activity better than the previous and current word residuals. **(D)** Full previous word and context embeddings do not predict MEG significantly differently. **(E)** All full embeddings predict both fMRI and MEG significantly better than the corresponding residual embeddings.

recordings result from different physiological signals, whether they capture the same underlying brain processes is still debated (Hall et al., 2014). Our results suggest that fMRI recordings are sensitive to supra-word meaning, while MEG recordings reflect instantaneous processes related to both the current word being read and the previously-read word. A likely candidate for the instantaneous process reflected in MEG is the process of integrating the current word with the previous context. The sensitivity to the previously-read word has many possible explanations. One possible explanation is that a word might take longer to process and integrate into the composed meaning than the duration it is on the screen. Another possible explanation is that a word may constrain the processing of the word that follows it, highlighting its relevant properties and aiding with composition. The hypothesis that MEG recordings reflect the process of composition aligns well with a vast number of previous findings characterizing transient responses evoked by a stimulus that is difficult to integrate with the preceding context (Kutas and Federmeier, 2011; Kuperberg et al., 2003; Kuperberg, 2007) and results showing that MEG recordings are better fit by a model constrained by the meaning of the immediately preceding words (Lyu et al., 2019). Indeed, our results are not in disagreement with this literature – they do not show that MEG activity does not reveal word integration processes that depend on previous context. Instead, our results suggest that the representation of that previous context is not visible in MEG.

The observed difference in predicting fMRI and MEG recordings raises the hypothesis that the process of maintenance of the composed meaning does not rely on neural mechanisms that are thought to generate the MEG signal (such as synchronized current flow in pyramidal cell dendrites (Hall et al., 2014)) but on some other mechanisms that are not visible in the MEG signal or might be indistinguishable from noise (e.g. unsynchronized neural firing), but that have enough metabolic demands to generate a BOLD response. Alternate possible explanations for the lack of predictability of MEG by supra-word meaning are that the representation of supra-word meaning may be too distributed to be captured by MEG due to its poor spatial resolution. However, we observe that the supra-word meaning predicts about $5-10\%$ of all cortical fMRI voxels across participants, mostly centered in the ATL and PTL. Thus, it is unlikely that no MEG sensor-timepoint is sensitive to this signal if it is detectable in the magnetic field changes. Further, MEG is known to be sensitive to neural activity that originates in the sulci, and since we find that the voxels that are sensitive to supra-word meaning are in the sulci, this explanation is even less likely. These results call for a more nuanced understanding of previous work that aims to study composition of sentence-level meaning using MEG as well as possibly other types of imaging modalities that rely on synchronized firing, such as EEG and ECoG. Our results suggest that observed increases in activity measured by these modalities during sentence reading (Fedorenko, Scott, et al., 2016; Hultén et al., 2019) and improved fit by a model constrained by very recent context (Lyu et al., 2019) may be due to instantaneous integration processes rather than sentence-level meaning. Future work is needed to understand whether and how these imaging modalities can be used to study sentence-level meaning.

Our analysis depends on the degree to which the computational neural network we have chosen is able to represent composed meaning. Based on ELMo's competitive performance on downstream tasks (Peters et al., 2018) and ability to capture complex linguistic structure (Tenney, Xia, et al., 2019), we believe that ELMo is able to extract some aspects of composed meaning. The degree

to which this composed meaning reflects the one in the brain is an important question that we have only begun to study and needs further investigation. Secondly, our residual approach accounts only for the linear dependence between individual word embeddings and context embeddings. By construction, the internal state of the LSTM in ELMo contains non-linear dependencies on the input word vector and the previous LSTM state. It is possible however that some dimensions of the internal state of the ELMo LSTM corresponds to non-linear operations on the dimensions of the input vector alone, without a contribution from the previous internal state of the LSTM (see Materials and Methods for the LSTM equations). This non-linear transformation of the input word might not be removed by our residual procedure, and whether it aligns with processing of individual words in the brain is a question for future research.

The surprising finding that supra-word meaning is difficult to capture using MEG has implications for future neuroimaging research and applications where natural language is decoded from the brain. While high temporal imaging resolution is key to reaching a mechanistic level of understanding of language processing, our findings suggest that a modality other than MEG may be necessary to detect long-range contextual information. Further, the fact that an aspect of meaning can be predictive in one imaging modality and invisible in the other calls for caution while interpreting findings about the brain from one modality alone, as some parts of the puzzle are systematically hidden. Our results also suggest that the imaging modality may impact the ability to decode the contextualized meaning of words, which is central to brain-computer interfaces (BCI) that aim to decode attempted speech. Recent success in decoding speech from ECoG recordings (Makin et al., 2020) is promising, but needs to be evaluated carefully with more diverse and naturalistic stimuli. Using BCI to decode speech in real life is complicated by the inherent uncertainty in decoding each word and the fact that the space of all possible utterances is not constrained. It is yet to be determined if word-level information conveyed by electrophysiology will be enough to decode a person's intent, or if the lack of supra-word meaning should be compensated in other ways.

4.5 Takeaways

The contributions of this chapter can be summarized as follows:

- We introduce a new approach based on computational modeling that makes interventions in NLP systems in order to capture the meaning of the whole as separate from the meaning of the parts. This approach allows the study of complex and composed multi-word meanings in the brain in a way not previously possible.
- Cognitive neuroscience has recently taken a precipitous dive into using neural networks to model brain processes. In language alone, a fast growing number of studies have used neural networks to study single word meaning, word sequence meaning, or different levels of language processing (Wehbe, Vaswani, et al., 2014; Jain et al., 2018; Toneva and Wehbe, 2019; Abnar et al., 2019; Beinborn et al., 2019; Hollenstein et al., 2019; Gauthier et al., 2019; Caucheteux and King, 2020). While we believe in the promise of the computational modeling approach, we do also recognize that it has many challenges. The main objection to this approach is that the problem of interpreting brain processes is transformed into a problem of

interpreting a computational model–which is often a hard-to-interpret neural network. Our approach in this chapter is a proposed solution to this limitation. We harness the power of computational modeling by creating in-vitro computational controls that we use to test more specific hypotheses.

- We identify potential limitations on the type of information that is detectable in MEG. While high temporal imaging resolution is key to reaching a mechanistic level of understanding of language processing, our findings suggest that a modality other than MEG may be necessary to detect long-range contextual information. Further, the fact that an aspect of meaning can be predictive in one imaging modality and invisible in the other calls for caution while interpreting findings about the brain from one modality alone, as some parts of the puzzle are systematically hidden. Our results also suggest that the imaging modality may impact the ability to decode the contextualized meaning of words, which is central to brain-computer interfaces (BCI) that aim to decode attempted speech.

Chapter 5

Modeling Processing of Task-Dependent Meaning

This chapter is based on work published as:
Mariya Toneva, Otilia Stretcu, Barnabas Poczos, Leila Wehbe, and Tom M Mitchell. "Modeling Task Effects on Meaning Representation in the Brain via Zero-Shot MEG Prediction". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020

In this chapter, we investigate the effect of a question task on the processing of a concrete noun by predicting the millisecond-resolution MEG brain activity as a function of both the semantics of the noun and the task. This work provides the first methodology that predicts brain recordings as a function of both the observed stimulus and question task. This is important because it not only encourages neuroscientists to formulate mechanistic computational hypotheses about the effect of a question on the processing of a stimulus, but also enables the comparison of different hypotheses by evaluating how well they predict brain recordings. Secondly, all machine learning analyses are performed in a zero-shot setting, in which neither the stimulus nor the question used to evaluate the learned models is seen during training (i.e. not just as the specific stimulus-question pair but also in combination with any other question/stimulus). This work is the first to successfully apply zero-shot learning to this question. This is important because it tests the generalization of the results beyond the experimental stimuli and tasks. Using our proposed approach, we show that incorporating the task semantics (i.e., the specific question asked) significantly improved the prediction of MEG recordings, across participants. The improvement occurs $475 - 550$ms after the participants first see the word, which corresponds to what is considered to be the ending time of semantic processing for a word. These results suggest that only the end of semantic processing of a word is task-dependent. This finding may inspire new NLP training algorithms or architectures that keep some computation task-independent, in contrast to current transfer learning approaches for NLP that tune all parameters of a pretrained model when training to perform a specific task.

5.1 Introduction

One of the central goals of artificial intelligence (AI) is to build intelligent systems that understand the meaning of concepts and use it to perform tasks in the real world. Despite the great strides in learning representations, there are still many problems that could benefit from further improvements in understanding and representing *meaning*, such as symbol grounding, commonsense reasoning, and natural language understanding. While machines are limited in these areas, we do have one system that is capable of representing meaning and performing these tasks well: the human brain. Thus, looking to the brain for insights about how we represent and compose meaning may be beneficial.

Studies of meaning representation in neuroscience have revealed that the brain accesses meaning differently depending on the demands of a task (Binder et al., 2009; Gan et al., 2013; Hebart et al., 2018; Xu et al., 2018; Wang, Xu, et al., 2018). For instance, the recorded brain activity of a participant that observes the word *"cat"* differs according to whether the participant is asked to answer whether *"cat"* is an animal or a vegetable (Kiefer, 2001). The difference is shown to occur between $400 - 600$ms after *"cat"* is presented to the participant, a period when it is believed that the brain processes the semantics of the perceived word (Helenius et al., 1998), suggesting an interaction between the task and stimulus meaning. One hypothesis for the interaction that has received some experimental backing is that, in order to solve the task, the brain uses attention mechanisms to emphasize task-relevant information (Smith, Shoben, et al., 1974; Cohen et al., 1990; Kanwisher et al., 2000; Cukur et al., 2013; Nastase, Connolly, et al., 2017). However, the computational principle behind this attention mechanism is poorly understood, as it can be due to several neural properties, such as an increased response gain, sharper tuning (Brouwer et al., 2013), or a tuning shift (Cukur et al., 2013).

In this work, we propose the first computational model that implements precise hypotheses for the interaction between the semantics of tasks and that of individual concepts in the brain, and tests their ability to explain brain activity. We posit that formulating such a computational model will be a helpful step towards specifying a full account of the task-stimulus interactions. Specifically, we study how tasks interact with the semantics of concepts by building models that predict recorded brain activity of people tasked with answering questions (e.g., *"is it bigger than a microwave?"*) about concrete nouns (e.g., *"bear"*). Importantly, the proposed model is able to generalize to previously unseen tasks and stimuli, allowing us to make zero-shot predictions of brain recordings.

Using this computational framework, we show that models that predict brain recordings as a function of the task semantics significantly outperform ones that do not during time windows ($475 - 550$ms and $600 - 650$ms) which largely coincide with the end of semantic processing of a word, typically thought to last until 600ms (Helenius et al., 1998). This result suggests that only the end of semantic processing of a word becomes task-dependent. We believe that in addition to this result, neuroscientists will also be interested in the ability to computationally compare different hypotheses for the task-stimuli interactions, and we hope that our general problem formulation will benefit future research attempting to study other forms of interaction not considered in this work. Additionally, our work may be of interest to the AI community. Further understanding task effects on concept meaning in the brain may provide insights into building AI models that learn how to combine representations specific to the task with task-invariant representations of concepts,

as a step towards composing meaning that is both goal-oriented and more easily adaptable to new tasks.

5.2 Related work

Classical neuroimaging experiments that study meaning by contrasting different stimulus conditions often include a task that is related to processing the meaning of the word (such as judging the similarity of two stimuli), however these experiments do not use predictive models that systematically relate the stimulus properties to the brain recordings, and do not explicitly investigate the task effect.

A number of previous studies have used predictive models to examine the relationship between brain recordings and stimulus properties, but have also not explicitly investigated the effect of a task. In many of these studies (Mitchell, Shinkareva, et al., 2008; Fyshe, Murphy, et al., 2013; Wehbe, Vaswani, et al., 2014; Jain et al., 2018; Toneva and Wehbe, 2019), the participants performed only one task – language comprehension – and, although this complex task can arguably be broken down into simpler tasks, this question was not explicitly investigated by the authors. In contrast, Sudre et al. (2012a) explicitly tasked participants with answering yes/no questions about objects. Even though the original paradigm of Sudre et al. (2012a) results in task-dependent brain recordings, the authors average the brain recordings for the same stimulus across tasks and learn predictive models only based on the semantics of the objects. While averaging over repetitions of the same stimulus can boost the signal-to-noise ratio, it likely loses the task-dependent information in the brain recordings. Here we reanalyze the original task-dependent single-repetition data from Sudre et al. (2012a) to investigate the task-dependent brain recordings using predictive models that include representations of both the object and the question.

One previous work uses a predictive model to investigate task effects (Cukur et al., 2013), and is thus closest to ours. In this work, the authors asked participants to attend to one of two object categories in natural scene stimuli. The authors then learn two separate models, each of which is trained to predict the fMRI recordings of participants in one of the 2 tasks as a function of the stimuli representations. They then compare the learned weights of the 2 models to conclude that each task-specific model puts more emphasis on those stimulus features that are related to the task. In contrast to this work, we integrate both the task and the stimulus representations into a single zero-shot learning framework, which allows us to predict brain recordings corresponding to novel tasks and stimuli. Additionally, we predict MEG recordings which have a 2000-times finer temporal resolution than the fMRI recordings used by Cukur et al. (2013), which allows us to localize the task effect in time.

The work of Nastase, Connolly, et al. (2017) also use a computational approach to investigate task effects. These authors account for the task directly in their computational model by constructing a different representational dissimilarity space (Kriegeskorte et al., 2008) for each of two tasks, and then comparing these to the representational dissimilarity space of brain recordings. The representational spaces of the stimuli are entirely task-dependent and do not incorporate the stimulus semantics. This is a limitation because this model is not able to investigate the relationship between the brain recordings and a possible interaction between the task and stimulus. Moreover,

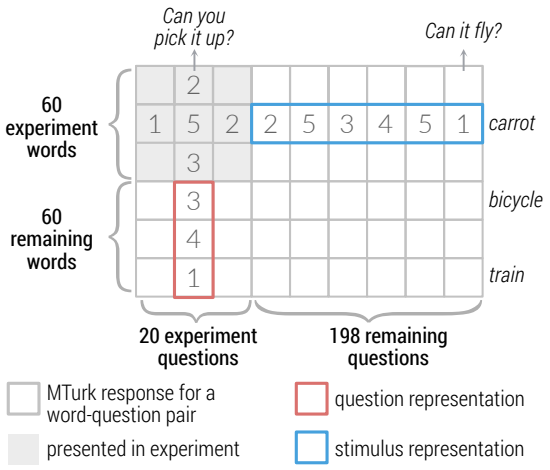


Figure 5.1: Feature representations of questions and stimuli obtained from Mechanical Turk.

similarly to Cukur et al. (2013), Nastase, Connolly, et al. (2017) predict fMRI recordings, limiting the investigation of the task effect in time.

5.3 Methodology

5.3.1 Brain data

We aim to study the effect of a task on the brain representation of a stimulus, when the stimulus is shown while performing the task. To this end, investigating the magnetoencephalograpy (MEG) dataset presented in Sudre et al. (2012a), which contains 20 different question tasks, makes for an excellent case study and was provided upon our request. The experimental paradigm and the data preprocessing are described in detail in Chapter 2.3.2.

Briefly, subjects were first presented with a question (e.g., *"Is it manmade?"*), followed by 60 concrete nouns, along with their line drawings, in a random order. Each stimulus was presented until the subject pressed a button to respond *"yes"* or *"no"* to the initial question. Once all 60 stimuli are presented, a new question is shown for a total of 20 questions. Thus we have a total of 60 stimuli $\times$ 20 questions $=$ 1200 examples.

We analyze the data from the beginning of the stimulus word presentation (i.e. 0ms) to 800ms, to avoid contributions to the brain signal from the participant's button-press (median response time across stimuli is 913ms, averaged across subjects). We further downsample the recordings in time by averaging non-overlapping 25ms windows, resulting in data of size 306 sensors $\times$ 32 time windows. We analyze data from 6 of the original 9 subjects. Data from 3 subjects were excluded because of missing trials.

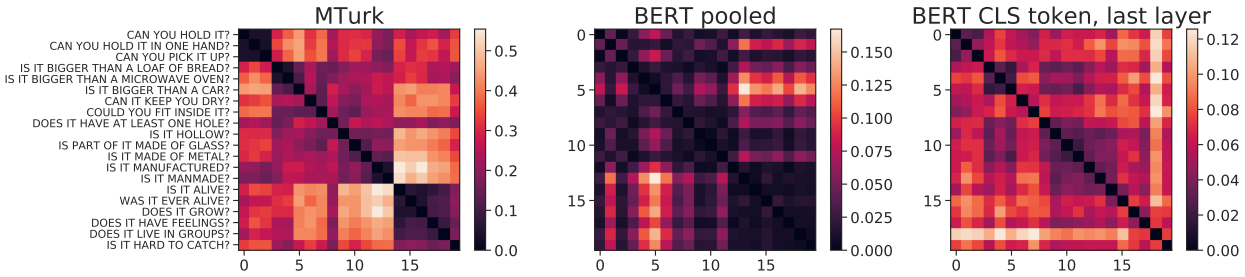


Figure 5.2: Pairwise cosine distances among the question representations from MTurk and the two types of BERT question representations. The MTurk representations appear to better cluster semantically-similar questions together.

5.3.2 Selecting representations for questions and stimuli

To study the effect of the question on the meaning representation of a word, we first need a way to represent both the semantics of the question and the word. We created two types of word and question representations: one type derived from a pretrained bidirectional model of stacked transformers (BERT) (Devlin et al., 2018), which is a popular model used for question-answering tasks, and a second type derived from Amazon Mechanical Turk (MTurk) of people answering questions about concrete nouns.

Word and question representations from BERT. BERT is a bidirectional model of stacked transformers that is trained to predict whether a given sentence follows the current sentence, in addition to predicting a number of input words that have been masked (Devlin et al., 2018). We use the base pretrained model provided by the Hugging Face Transformers library (Wolf et al., 2019). This model has 12 layers, 12 attention heads, and 768 hidden units.

We first apply WordPiece tokenization to each of our 60 word stimuli and questions. To extract the word features, we pass the tokens corresponding to each word stimuli into the BERT model and extract the corresponding token-level embeddings. We use these token-level embeddings as the BERT word representations in the following experiments. If any word contains more than 1 token, it is assigned the average of the corresponding token-level embeddings.

To extract the question features, we pass the tokenization of each question separately into BERT, with a '[CLS]' token and a '[SEP]' token appended to respectively the beginning and end of the tokenized list. This is common practice with inputting multi-word sequences into BERT. We then extracted the hidden layer activations from the CLS token at the last hidden layer, as well as the pooled output. In the pretrained model, the pooled representation of a sequence is defined as the embedding of the [CLS] token at the last layer after a transformation through a fully-connected layer and then a tanh function.

There are many different choices for where to extract the question representations from BERT – e.g. the CLS token representation from any of the 12 hidden layers, the pooled output, or a mean or max pooling across all token representations in any of the layers. To settle on which BERT representation may be best suited to predict the MEG recordings, we first visualize how similar

the representations for our 20 questions are under different BERT representations (see Figure 5.2). We observe that the MTurk representations appear to better cluster semantically-similar questions together. The representations from the CLS token at the last layer appear to cluster together sentences that share words, whereas the pooled output representations appear to lead to at least two larger clusters that correspond to questions related to animacy and size. We therefore conduct experiments using the BERT pooled output embeddings as the question representations.

Word and question representations from MTurk. The Mechanical Turk data was originally collected by Sudre et al. (2012a) and was provided at our request. Participants on MTurk were shown a set of 1000 words (e.g., *"bear", "house"*) and were requested to answer 218 questions about them (e.g., *"Is it fragile?", "Can it be washed?"*) on a scale from 1 to 5 ("definitely not" to "definitely yes"). In this dataset, 60 out of the 1000 presented words and 20 out the 218 questions corresponded to the stimuli and questions shown during the brain recording experiment. A complete list of words and questions is shown in Appendix C.

Using this dataset, we define the representation of a word as a vector containing the MTurk responses for that word to all 198 questions not in the experiment (see Figure 5.1). Moreover, we define the task (i.e. question) representation as a vector containing the MTurk responses for 60 words which are not in the experiment. Using more words did not result in improved performance on the validation set. We purposefully excluded the questions and words in the brain experiment from these representations. Note that Sudre et al., 2012a used the same word representations, but to the best of our knowledge, we are the first to represent question semantics as a collection of answers.

With permission from Sudre et al. (2012a), we provide the MTurk representations of the stimuli and questions in https://github.com/otiliastr/brain_task_effect. We further provide the MTurk human-judgments for all 1000 words, and the BERT representations.

5.3.3 Hypotheses

Next, we formulate several hypotheses of how the question integrates with the stimulus in order to give rise to a task-dependent meaning representation. First, we will introduce the notation used to define the hypotheses, as well as the concrete models described in the next section. Using the notation in Table 5.1, we propose the following hypotheses, also shown in Figure 5.3:

Hypothesis 1 (no task effect): The brain activity is not affected by the task and can mostly be explained by the stimulus. Thus, we can approximate the elicited brain activity as: $b = f_s(s)$.

Hypothesis 2 (no stimulus effect): The brain activity is not affected by the stimulus and can mostly be explained by the task. Thus, we can approximate the elicited brain activity as: $b = f_t(s)$. While this hypothesis may not predict the brain activity the best, it will allow us to localize the task effect.

Hypothesis 3 (additive): Both the stimulus and the task affect the brain activity, but their contributions are independent: $b = f_s(s) + f_t(t)$.

Hypothesis 4 (interactive): The brain activity is well explained by the stimulus, but the task

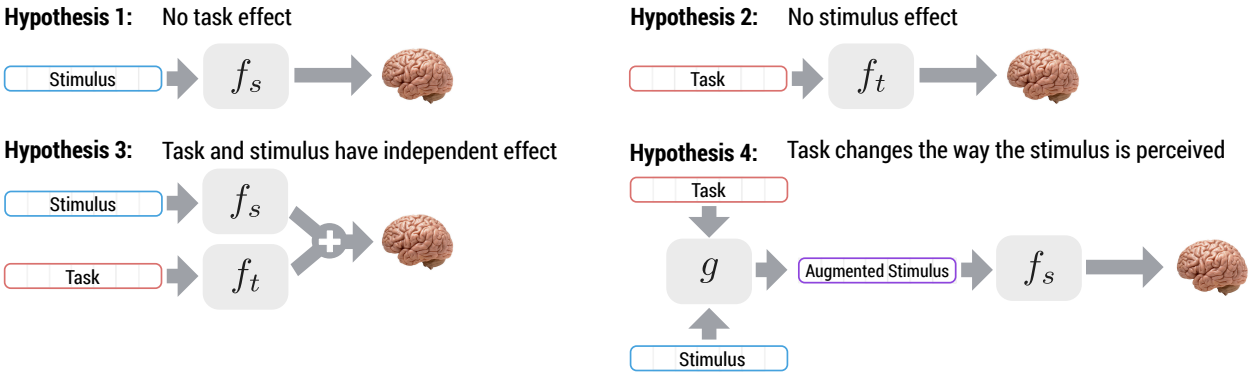


Figure 5.3: Proposed hypotheses for how the task and stimulus affect brain activity. Hypothesis 1 posits that the brain activity is not affected by the task and can mostly be explained by the stimulus. Hypothesis 2 posits that the brain activity is not affected by the stimulus and can mostly be explained by the task. Hypothesis 3 posits that both the stimulus and the task affect the brain activity, but their contributions are independent from each other. Lastly, Hypothesis 4 posits that the brain activity is affected by both the stimulus and the task and that the task augments the way the stimulus is perceived via the function g, which can be learned from data or pre-specified.

changes the way the stimulus is perceived: $b = f_s(t \otimes s)$. We can think of this as the task focusing *attention* on particular features of the stimulus that are relevant to the task (e.g., in answering the question *"Is it bigger than a car?"* for the stimulus *"dog"*, we pay more attention to the features of *"dog"* that are related to size, and ignore others such as color). This hypothesis aligns with the conclusions of Cukur et al. (2013) that a task emphasizes those semantic dimensions of the stimulus that are relevant to the task. We use the notation $\otimes$ to represent generically any type of augmentation, and in Section 5.3.4 we describe in detail the forms of attention used in our experiments.

5.3.4 Predicting brain activity under different hypotheses

We next formulate models to represent the parametric functions f_s and f_t in the proposed hypotheses and to learn the parameters that best predict the brain activity. Our notation is summarized in Table 5.1. In the rest of this paper, we refer to the hypotheses using the abbreviations H1, H2, H3 and H4.

Models

The functions f_s and f_t can be represented using any regression models that map from a feature space to the brain activity space. Prior work (Mitchell, Shinkareva, et al., 2008; Nishimoto et al., 2011a; Wehbe, Murphy, et al., 2014; Huth, Heer, et al., 2016) has shown that simple multivariate regression models such as *ridge regression* are reliable tools for predicting brain activity from stimulus features and are able to achieve good accuracy. For this reason, we will adopt the ridge regression setting for modeling f_s and f_t. In ridge regression, we model the output of a function f

as a linear combination of the input features: $\hat{y} = f(x) = xW$, where W is a parameter matrix. W is trained to minimize the loss function $\|Y - XW\|_F^2 + \lambda\|W\|_F^2$, consisting of the mean squared error of the predictions and a regularization term on the parameters to avoid overfitting. Here X and Y represent the training inputs and targets, respectively, stacked together, $\|.\|_F$ denotes the Frobenius norm, and $\lambda > 0$ is a tunable hyperparameter representing the regularization weight. In our setting, the targets of the prediction Y consist of the MEG recording of the brain activity, Y_b, described in Table 5.1. However, the inputs X depend on the hypothesis being tested, as we describe further.

Hypothesis 1: Under a *no task effect* hypothesis, we predict the brain activity as a function of the stimulus features only, $Y_b = f_s(X_s) = X_s W_s$, where $W_s \in \mathbb{R}^{F_s \times LT}$. The objective function is:

$$\min_{W_s} \|Y_b - X_s W_s\|_F^2 + \lambda\|W_s\|_F^2 \tag{5.1}$$

Hypothesis 2: Under a *no stimulus effect* hypothesis, we predict the brain activity as a function of the task features only, $Y_b = f_t(X_t) = X_t W_t$, where $W_t \in \mathbb{R}^{F_t \times LT}$. Our objective function becomes:

$$\min_{W_t} \|Y_b - X_t W_t\|_F^2 + \lambda\|W_t\|_F^2 \tag{5.2}$$

Hypothesis 3: Under an *additive effect* hypothesis, we predict the brain activity as the sum of the stimulus contribution and task contribution: $Y_b = f_s(X_s) + f_t(X_t) = X_s W_s + X_t W_t$. Note that this is equivalent to a single regression function $f(X_s, X_t) = [X_s, X_t] \cdot [W_s; W_t]$, where $[X_s, X_t] \in \mathbb{R}^{R \times (F_s + F_t)}$ is a concatenation of the stimulus and task features, and $W = [W_s; W_t] \in \mathbb{R}^{(F_s+F_t) \times LT}$ is a concatenation of their corresponding weight matrices. Thus, the objective can be written as:

$$\min_{W_s, W_t} \|Y_b - [X_s, X_t] \cdot [W_s; W_t]\|_F^2 + \lambda\|[W_s; W_t]\|_F^2 \tag{5.3}$$

Hypothesis 4: Under an *interactive effect* hypothesis, we predict the brain activity as a function of the *augmented* stimulus features. The intuition is that the task *augments* the features that are relevant. In this work, we consider an implementation of the augmentation using *soft attention* (Hermann et al., 2015), in which the task reweighs the contribution of the stimulus features. To simplify the notation in the following formulations, we will use t and s to refer to both the identity and the representation of a task and a stimulus in the experiment. Each task t is associated with a set of attention parameters $a_t \in \mathbb{R}^{F_s}$ that rescale the original stimulus features when the stimulus s is presented under question t. Thus, the augmented stimulus features under question t become $\bar{s} = a_t \otimes s$, where $\otimes$ represents element-wise multiplication. The augmented stimuli for all training examples can be stacked together in an a matrix $X_{\bar{s}}$, and used as input to a ridge regression model, similar to H1:

$$\min_{W_s} \|Y_b - X_{\bar{s}} W_s\|_F^2 + \lambda\|W_s\|_F^2 \tag{5.4}$$

Table 5.1: Notation used in defining the proposed hypotheses and models.

N_s	num. unique stimuli in experiment, 60	$\hat{b}$	predicted brain activity; $\hat{b} \in \mathbb{R}^{LT}$
N_t	num. unique tasks in experiment, 20	X_s	stimuli representations, stacked for all repetitions; $X_s \in \mathbb{R}^{R \times F_s}$
R	total num. repetitions, over all stimuli, 1200	X_t	task representations, stacked for all repetitions; $X_t \in \mathbb{R}^{R \times F_t}$
L	space dimension of the brain activity, 306	Y_b	recorded brain activity, stacked for all repetitions; $Y_b \in \mathbb{R}^{R \times LT}$
T	time dimension of the brain activity, 32	f_s	function mapping from s to $\hat{b}$; $f_s : \mathbb{R}^{F_s} \rightarrow \mathbb{R}^{LT}$
F_s	num. features in stimulus representation, 198 for MTurk; 768 for BERT	f_t	function mapping from t to $\hat{b}$; $f_t : \mathbb{R}^{F_t} \rightarrow \mathbb{R}^{LT}$
F_t	num. features in task representation, 60 for MTurk; 768 for BERT		
s	stimulus representation; $s \in \mathbb{R}^{F_s}$		
t	task representation; $t \in \mathbb{R}^{F_t}$		

The attention vectors a_t can be precomputed or learned along the regression parameters, as follows:

H4.1. Precomputed attention: The MTurk features have interpretable dimensions for both tasks and stimuli, which enables us to directly compute the hypothesized relevance of different stimuli dimensions to each task. As described in Section 5.3.2, each semantic dimension of a word corresponds to one of the $F_s = 198$ non-experimental questions (see Figure 5.1). Given this relationship, we compute the attention parameters for every stimulus presented under task t as $a_t = \texttt{softmax}([a_{t,\tilde{t}_j}])$ for $j \in \{1, \ldots, F_s\}$, where $\tilde{t} \in \mathbb{R}^{F_t}$ is a representation of a non-experimental question, and $a_{t,\tilde{t}} = \texttt{cosine\_similarity}(t, \tilde{t})$. We observe that this precomputed attention indeed emphasizes semantically-relevant word features. For example, the word features with highest attention for the question *"Is it made of metal?"* are *"Is it silver?"* and *"Is it mechanical?"*. The top 5 word features with highest attention for each question are provided in Appendix C.

H4.2. Learned attention: We learn the attention parameters together with the regression parameters with the objective of predicting the brain recordings as accurately as possible. A direct approach would be to learn a different set of attention parameters a_t for every task t. However, since our goal is to be able to make *zero-shot* predictions for tasks and stimuli never seen during training, we instead learn how to map the features of the task to an attention vector. In our experiments we did so by learning an attention matrix $A \in \mathbb{R}^{F_t \times F_s}$, such that $a_t = \sigma(tA)$, where $\sigma(.)$ represents the sigmoid function, applied element-wise. Putting all pieces together, our objective function becomes:

$$\min_{W_s, A} \|Y_b - \sigma(X_t A) X_s W_s\|_F^2 + \lambda \|W_s\|_F^2 + \lambda_A \|A\|_F^2 \tag{5.5}$$

Training and evaluation

We next train and evaluate all models. Our goal is to predict brain recordings for any new task and word (i.e. zero-shot). Thus, we train all models using *leave-k-out* cross-validation, in which we leave out all training examples that correspond to task-stimuli pairs that contain either a task or a word that will be used for testing. We choose the regularization parameters via nested cross-validation.

We evaluate the predictions from each model by using them in a classification task on the held-out data, in the *leave-k-out* setting. The classification task is whether we are able to match the brain data predictions for two heldout task-stimuli pairs to their corresponding true brain data. This task has been previously proposed for settings with low signal-to-noise ratio (Mitchell, Shinkareva, et al., 2008). The classification is repeated for each leave-k-out fold and an average classification accuracy is obtained for each sensor-timepoint. We refer to this accuracy as *2v2 accuracy*. The theoretical chance performance is 0.5. A more detailed explanation about this metric can be found in Chapter 3.3.4. Our code with all training and evaluation details is available at https://github.com/otiliastr/brain_task_effect.

Input and output normalization. When training our prediction models, both the model inputs and the targets are normalized as follows. For the input features (which could be word features, question features or both, depending on the hypothesis), we z-score each feature x_i along the sample dimension by assigning $x_i \leftarrow \frac{x_i - \text{mean}(x_i)}{\text{std}(x_i)}$, such that each feature x_i has mean 0 and standard deviation 1 across the samples. Similarly, we z-score every sensor-timepoint of the outputs across samples. However, to make sure our evaluation is correct and no information about the test data has been leaked during training, the mean and standard deviation used when z-scoring is calculated only over the training data.

Train/test splits. The next step is to train and evaluate the proposed models. For this, we need to separate our data into a train set and a test set. Since our dataset contains very few samples, as it is usually the case in neuroscience, we adopt the common *leave-k-out* cross-validation approach (Mitchell, Shinkareva, et al., 2008; Sudre et al., 2012a; Wehbe, Murphy, et al., 2014), in which the data is repeatedly split into 2 groups: one containing k repetitions for test, and one with $R - k$ repetitions used for training. Each training set is further split into 2 subgroups using a similar approach: one for training, and one for parameter validation. A model is trained on the inner training set using multiple hyperparameters, and the ones with the best average validation accuracy are selected. Using the best hyperparameters, we retrain on the train+validation data, and compute the final accuracy on the test set. A common choice for k in approaches that average the stimulus repetitions (Mitchell, Shinkareva, et al., 2008; Sudre et al., 2012a) is 2, because this allows us to compare the brain activities for two left out stimuli (as described in the next paragraph), while training on as much data as possible. However, since we want our models to perform zero-shot learning and to be able to make predictions for both words and questions that have not been seen during training, we leave out from training 2 stimuli with all their 20 repetitions under different questions, but also 2 questions with all 60 words about which this question was asked (i.e. a total of $2 \times 20 + 60 \times 2 = 160$ examples). Out of these 160 examples, we only test the model performance

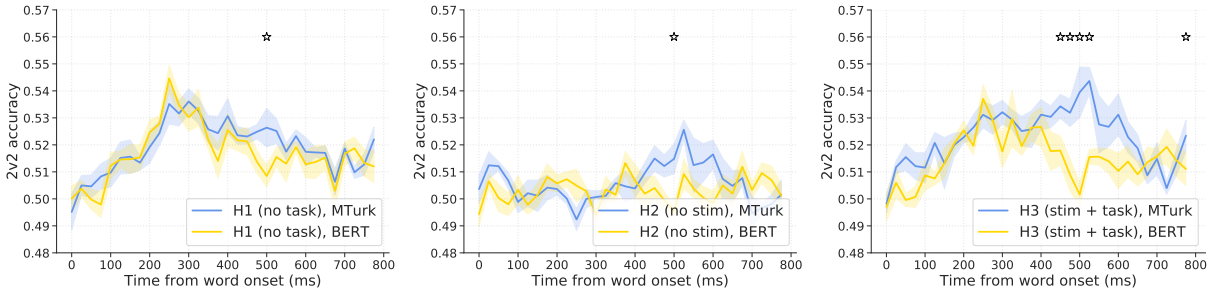


Figure 5.4: Comparisons of 2v2 accuracies of predictions computed using MTurk vs. BERT features in each hypothesis. Mean accuracy and standard error across subjects plotted. Points where the means across subjects are significantly different are marked with a $\star$ symbol.

on the word-question pairs for which neither the word nor the question appear in training. We do this type of splitting both when performing train/test splitting, and for train/validation splitting.

Optimization. While H1, H2 and H3 can be solved in closed form, we use the Cholesky decomposition approach provided in the Python `scikit-learn` package (Pedregosa et al., 2011) for computational reasons. In H4, we need to optimize the parameters of the functions g and f_s together, and thus we implemented this using the TensorFlow framework (Martín Abadi et al., 2015) and trained end-to-end using the Adam optimizer (Kingma et al., 2015) with default parameters and learning rate 0.001.

Parameters and hyperparameters. The only hyperparameters in our framework are the regularization parameters λ (for all hypotheses) and λ_A (only for H4.1). Their values were chosen from the set of values $\{10^{-5}, 10^{-4}, ..., 10^{7}\}$ using the train/validation/test splitting described above. We also allowed the model to select different λ per sensor-timepoint, but found that using the same value for outputs is more stable and leads to better validation accuracy overall. Moreover, we found conducting the parameter validation using the cross-validation setting described above on all subjects and all hypotheses to be prohibitive, and thus we performed the hyperparameter tuning for each hypothesis on a single subject, which was then excluded from testing. Regarding the number parameters, each hypothesis has a different number of parameters, depending on the size of the of the inputs and extra attention parameters, with H3 $>$ H4.1 $>$ H1 $=$ H4.2 $>$ H2. To ameliorate any effects of overfitting, we allow each hypothesis to choose its own regularization parameters.

5.4 Results and discussion

5.4.1 BERT vs. MTurk representations.

Features extracted from BERT for both the stimuli and questions perform significantly worse than the Mechanical Turk (MTurk) features in several timewindows across hypotheses (see Figure 5.4) (paired t-test, significance level 0.05, FDR controlled for multiple comparisons (Benjamini et al.,

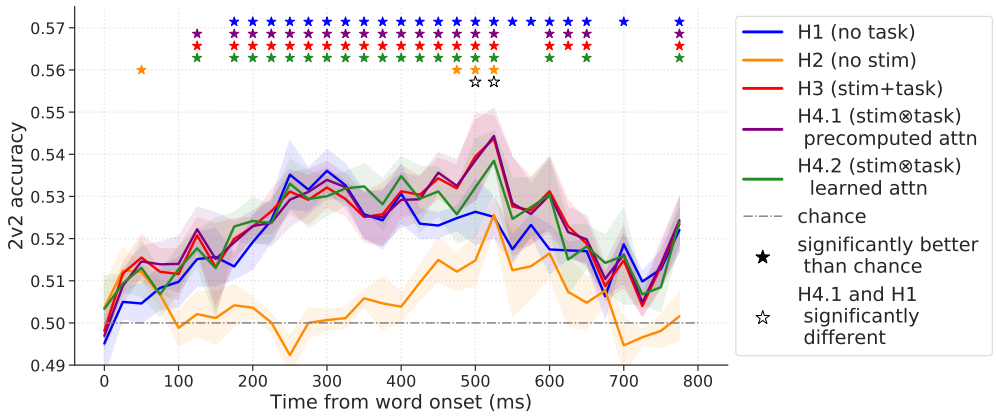


Figure 5.5: Performance of all hypotheses at predicting the MEG recordings in 25ms windows, averaged over sensors. We show the mean and std. error over subjects. The task effect is mostly localized to $475-550$ms. Hypotheses that incorporate both the stimulus and task perform similarly across time.

1995)). Some of the difference in performance may be due to the large difference in dimensionality between the BERT and MTurk features (stimulus representations: 198 for MTurk vs. 768 for BERT; question representations: 60 for MTurk vs. 768 for BERT). The BERT features have much higher dimensionality than the MTurk features, which may lead to more overfitting when using the BERT features. However, this is likely not the only cause for the difference, as the question representations from BERT appear much worse at predicting the MEG recordings in the $450-550$ms timewindow, where we show the question/task semantics contribute most to the MEG recordings (see Figure 5.5). It is likely that the pretrained BERT model is not able to compose the input words in a way that is as brain-aligned as the question representations from Mechanical Turk. It would be an interesting future direction to fine-tune BERT on a question-answering task and compare the performance of the pretrained question representations with that of the fine-tuned BERT representations. Because the MTurk representations significantly outperform the BERT representations for predicting MEG recordings, we focus on the MTurk representations in all future experiments.

5.4.2 Effect of incorporating question task semantics

Time window results. We present the 2v2 accuracy per 25ms time window of all tested hypotheses in Figure 5.5. The time points for which each accuracy significantly differs from chance are indicated with a $\star$ symbol (one-sample t-test, 0.05 significance level, FDR controlled for multiple comparisons (Benjamini et al., 1995)). We observe that the hypothesis that only considers the question task semantics (H2) performs significantly better than chance in one early time window ($50-75$ms) and much later during $475-550$ms. The remaining hypotheses also perform better than chance in the same $475-550$ms window, but we observe that during the majority of that time H3 and H4.1 perform significantly better than H1 (paired t-test, 0.05 significance level, FDR controlled for multiple comparisons).

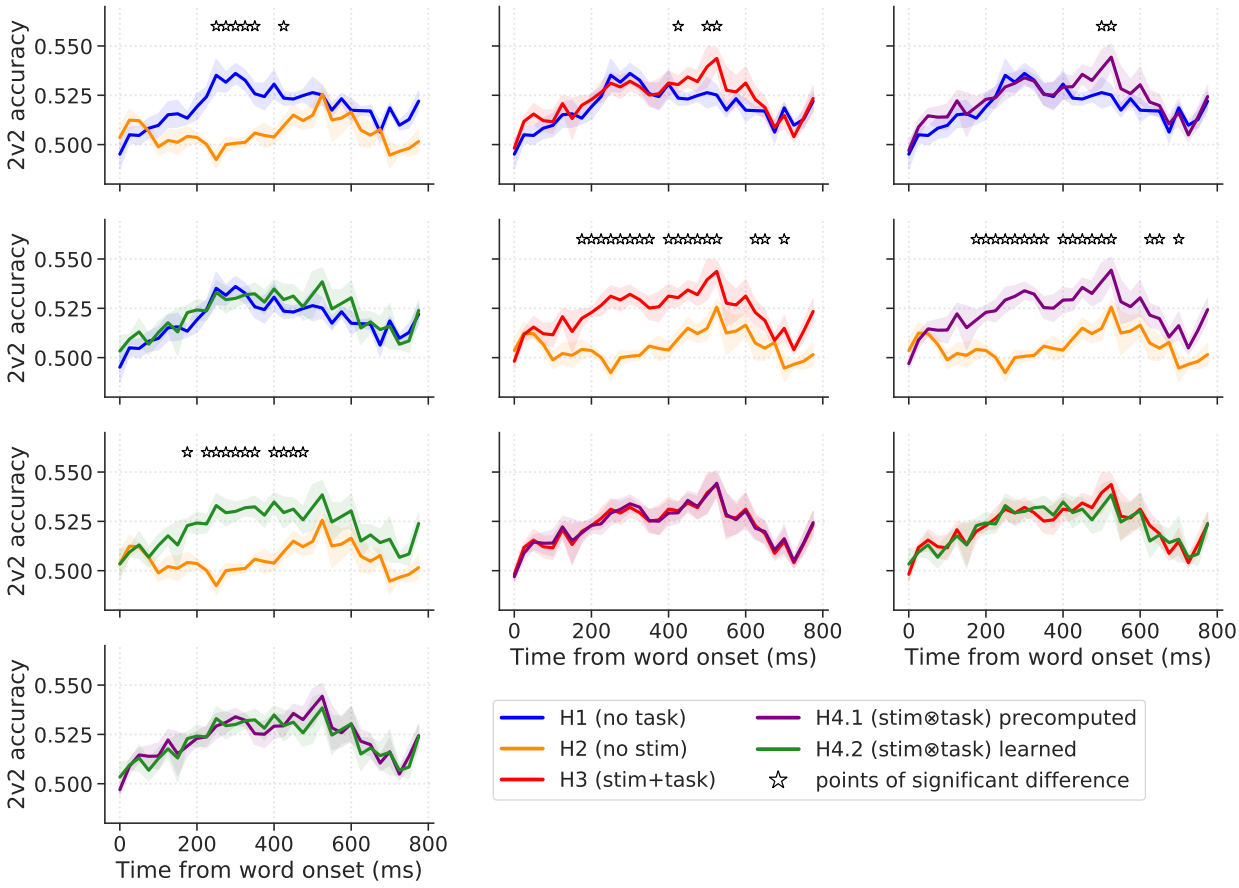


Figure 5.6: Pairwise comparisons of 2v2 accuracy performance across all tested hypotheses. All timepoints where there is significant difference between the performances of the two displayed hypotheses are marked with a star.

Additionally, we present the pairwise comparisons of 2v2 accuracy performance across all tested hypotheses in Figure 5.6. All timepoints where there is significant difference between the performances of the two displayed hypotheses are marked with a star (paired t-test, significance level 0.05, FDR controlled for multiple comparisons). The hypothesis that does not incorporate information about the stimulus (H2) performs significantly worse than all hypotheses that do during $250 - 400$ms. The hypothesis that does not incorporate information about the task (H1) performs significantly worse than two hypotheses that do (H3 and H4.1) in time windows between 450 and 550ms. There is no significant difference in the performances across timewindows of all hypotheses that are a function of both the stimulus and task (H3, H4.1, and H4.2). We conclude that incorporating the question task semantics can improve the prediction of MEG recordings. Note that all discussed times are measured relative to stimulus onset.

Sensor-timepoint results. We investigate the task effect further by comparing the contribution of the question-specific precomputed attention and the word features to the accuracy of H4.1 by

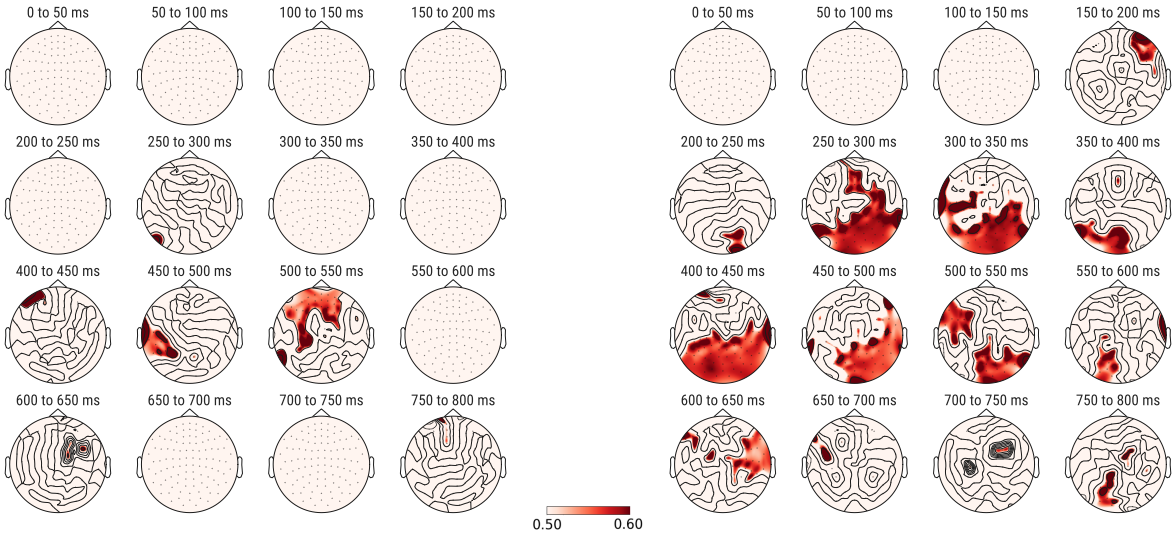


Figure 5.7: Question- and word-contribution in predicting MEG recordings. Mean 2v2 accuracy across subjects of predicting sensor-timepoints in 50ms windows using H4.1, when predicting the brain recordings for two word-question pairs that share the same word (Left) and the same question (Right). Displayed accuracies are significantly greater than 0.5. The main question contribution appears in the frontal and temporal lobes during $400 - 550$ms, whereas the word contribution is distributed across the occipital and temporal lobes during $200 - 650$ms, providing evidence that the end of word processing is task-dependent.

computing the 2v2 accuracy in two special cases: (1) when the two tested word-question pairs share the same word (i.e. (q_1, w_1) vs (q_2, w_1)), higher-than-chance accuracy is attributed to the pre-computed attention features; (2) when the two tested word-question pairs share the same question (i.e. (q_1, w_1) vs (q_1, w_2)), higher-than-chance accuracy is attributed to the word features. These results are presented per sensor-timepoint in Figure 5.7, where only higher-than-chance accuracies across participants are shown (one-sample t-test, 0.05 significance level, FDR controlled for multiple comparisons). The results are visualized using MNE-Python Gramfort et al., 2013. The main contribution of the question-specific attention appears between 400-550ms, localized to the frontal and the left temporal lobes. The contribution of the stimulus features is more distributed, both in time and space. The effect of word semantics begins at 150ms and extends until the end of the considered time, with major contributions in the occipital lobes ($200 - 600$ms) and temporal lobes ($400 - 550$ms, $600 - 650$ms). For ease of visualization, here we present results for 50ms time windows. The results for 25ms time window align with the presented effects and are provided in Appendix C.

5.4.3 Comparison of task-stimulus interaction hypotheses

We further test which of the 3 hypothesized types of task-stimulus interaction (i.e., independent in H3, precomputed attention in H4.1, or learned attention in H4.2) best explains the observed MEG recordings. We observe that there is no significant difference among these hypotheses when

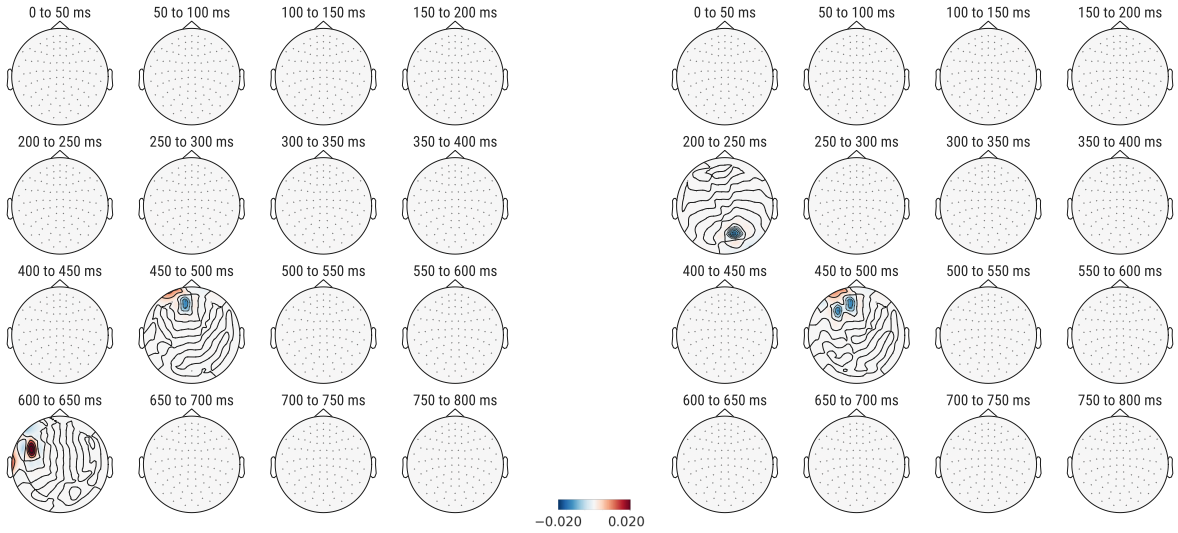


Figure 5.8: Significant differences in performance between H4.1 and H3 per sensor-timepoint (H4.1 accuracy - H3 accuracy). The red points are those where H4.1 significantly outperforms H3, and the blue are those where H3 significantly outperforms H4.1. We're only displaying the significant differences for those timepoints where H4.1 performs significantly better than chance (Left) and where H3 performs significantly better than chance (Right).

averaging over the performance in all sensors (significance shown in Supplementary Figure 5.6).

Sensor-timepoint results. We present the significant differences in performance between H4.1 and H3 per sensor-timepoint in Figure 5.8. We have plotted H4.1 accuracy - H3 accuracy (the red points are those where H4.1 significantly outperforms H3, and the blue are those where H3 significantly outperforms H4.1). We're only displaying the significant differences for those timepoints where H4.1 performs significantly better than chance (Left) and where H3 performs significantly better than chance (Right).

A group of sensors in the occipital lobes are significantly better predicted by H3 than by H4.1 at $200-250$ms (paired t-test, 0.05 significance level, FDR controlled for multiple comparisons). This is when semantic processing of a word begins, so H3 may outperform H4.1 here because H3 has an independent contribution from the word representation. Both H3 and H4.1 perform significantly better than chance during $450-500$ms, and there are different sensor groups in the frontal lobe that are significantly better predicted by each hypothesis than the other. This suggests that this time point may contain both independent and interactive contributions of the task. We lastly observe that H4.1 outperforms H3 in the left temporal lobe during $600-650$ms. This localization suggests that the word and question semantics may interact in this time period, rather than be processed independently.

Learned attention. Because there are no significant differences between the performance of H4.2 and H4.1, we compare the learned attention parameters in H4.2 to the precomputed ones in

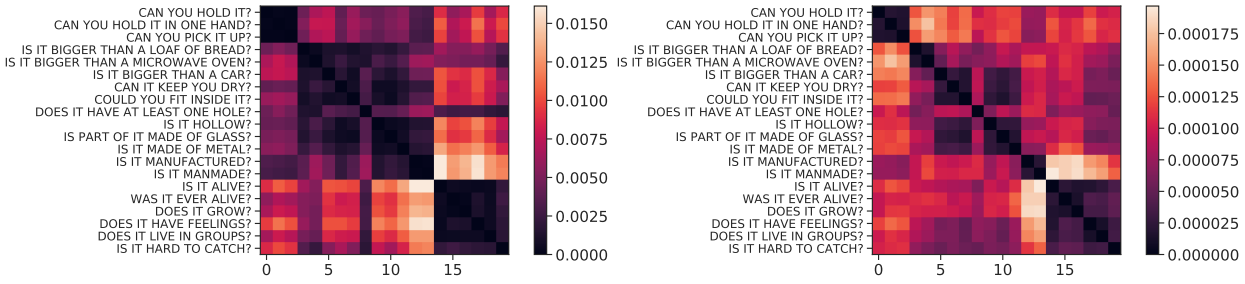


Figure 5.9: Pairwise cosine distances across the question-wise attention in H4.1 (Left) and H4.2 (Right). The question-wise attention in H4.2 is an average over participants. The Pearson correlation between these matrices (upper-triangle only) is 0.69, indicating a high degree of correspondence between the precomputed attention (Left) and the mean learned attention (Right).

H4.1. To this end, we compute the pairwise cosine distances across the precomputed question-wise attention and across the learned question-wise attention. The resulting cosine distance matrices are visualized in Figure 5.9. To quantify the similarity between the two, we compute the Pearson correlation between the upper-triangles of the two matrices, which comes out to 0.69. This indicates that the learned attention, that is randomly initialized, learns to combine the question and word features in a way that is very similar to the precomputed attention.

This suggests that either the precomputed attention is one optimal way to combine the stimulus and task in predicting the brain recordings, or that more samples are needed to learn a better combination. To further understand the effect of the sample size, we evaluated both H4.2 and H3 with varying amounts of training data. We tested H3 because it is a simpler model that we expect to learn with fewer samples. Since it would be difficult to collect more data beyond the 20 questions and 60 words, we performed this experiment by reducing the amount of data we allow the model to train on.

We trained both H4.2 and H3 with decreasing amounts of data and tested their performance. The results are shown in Figure 5.10. H3 continues to improve as we add more examples, up to the maximum (i.e. 1044 samples = 58words$\times 18$ questions). This suggests that even this simpler model may benefit from more training data. H4.2 also appears to improve with more samples, however it is less clear whether the performance peak has been reached or whether this is due to the difficulty of the optimization problem. These results are even more clear in the time interval $450 - 600$ms, where we expect the two models to perform the best according to the results in Figure 5.5.

5.4.4 Effect size

We note that the magnitudes of the presented effects (i.e. accuracies, differences between hypotheses) are limited due to the small amount of data and the underlying difficulty of analyzing single-trial MEG data. The accuracies we observe are on par with other reported single-trial MEG accuracies (Wehbe, Vaswani, et al., 2014). Other work has mitigated the low signal-to-noise ratio of single-trial MEG by averaging the recordings corresponding to different repetitions of the same stimulus (Sudre et al., 2012a) or grouping 20 examples together for a 20v20 classification

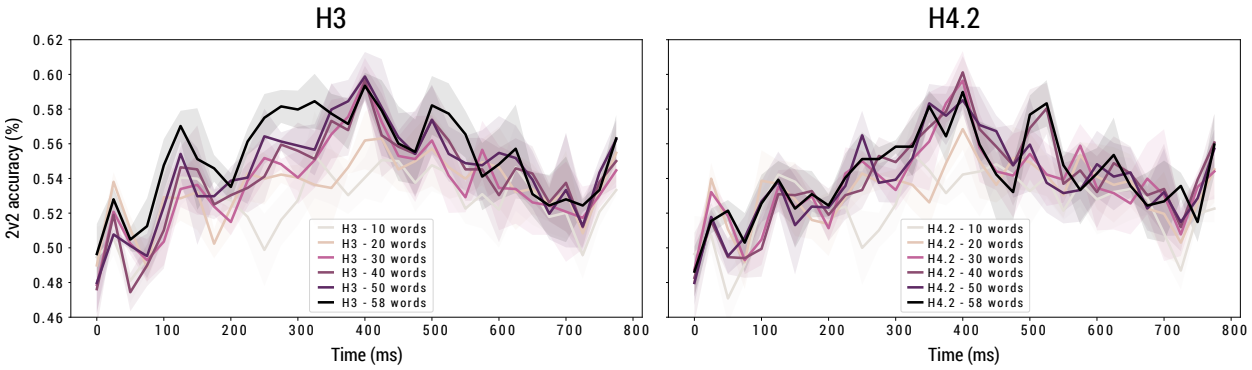


Figure 5.10: Experiments with various amounts of training data.

task (Wehbe, Vaswani, et al., 2014). Neither is applicable here because our data does not contain repetitions of the same question-stimulus pair, and our zero-shot setting would require us to hold out a large portion of our training set if we were to evaluate on 20 stimulus-question pairs.

In the absence of these options, we have taken careful precautions to validate our results (by evaluating our models on held-out data in a cross-validated fashion) and evaluated the significance of the model performances and differences between them, and corrected for multiple comparisons. We trust that the effects we have shown to be significant are indeed true, but we note that there may be effects that we are not able to reveal due to limited power and hope that neuroscientists will apply our methods in the future to larger datasets with multiple repetitions.

5.4.5 Discussion and relation to previous results

Taken together, our results point to a robust effect of the question task semantics on the brain activity during $475 - 550$ms. We also find an effect of the interaction between the question and stimulus semantics during $600 - 650$ms, localized to the temporal lobe. The temporal lobes are implicated in semantic processing (Binder et al., 2009; Hagoort, 2013; Skeide et al., 2016; Hickok et al., 2016) and specifically in maintaining relevant lexical semantic information for the purposes of integration (Hagoort, 2020). Since this effect occurs past the time when a word is thought to be processed (i.e. up to 600ms), it may be related to maintaining specific semantic dimensions that help answer the question (the median response time across participants is 913ms). In addition to being localized to the temporal lobes, the earlier question effect is also found in the frontal lobes, which are thought to support attention (Duncan, 1995; Stuss, 2006). A task effect that is related to attention is consistent with findings from (Cukur et al., 2013; Nastase, Connolly, et al., 2017). Our results expand these previous findings by characterizing the temporal dynamics of the task-stimulus interactions.

5.5 Conclusion and future work

We propose a computational framework for comparing different hypotheses about how a task affects the meaning of an observed stimulus. The hypotheses are formulated as prediction problems, where a model is trained to predict the brain recordings of a participant as a function of the task and stimulus representations. We show that incorporating the semantics of a question into the predictive model significantly improves the prediction of MEG recordings of participants answering questions about concrete nouns. The timing of the effect coincides with the end of semantic processing for a word, as well as times when the participant is deciding how to answer the question.

These results suggest that only the end of semantic processing of a word is task-dependent. This finding may inspire new NLP training algorithms or architectures that keep some computation task-independent, in contrast to current transfer learning approaches for NLP that tune all parameters of a pretrained model when training to perform a specific task (Devlin et al., 2018). Moreover, future work can extend our methods to incorporate representations of tasks and stimuli from powerful neural networks that are augmented with improved commonsense knowledge (Da et al., 2019), which would eliminate the need for human-judgment annotations. Furthermore, only one of the tested hypotheses (H4.1) is experiment-dependent, while all others can be applied to data from any neuroscience experiment, as long as task and stimulus feature representations can be obtained. Our results pose a challenge for future research to formulate new hypotheses for earlier effects on processing as a function of the task and stimuli.

5.6 Takeaways

The contributions of this chapter can be summarized as follows:

- We propose a means of representing the semantics of the question task that shows a significant relationship with the elicited brain response. We believe such an approach could be useful to future studies on question-answering in the brain.
- We provide the first methodology that can predict brain recordings as a function of *both* the observed stimulus and question task. This is important because it will not only encourage neuroscientists to formulate mechanistic computational hypotheses about the effect of a question on the processing of a stimulus, but also enable neuroscientists to test these different hypotheses against each other by evaluating how well they can align with brain recordings. While we have implemented and compared several hypotheses for this effect, and have found some to be better than others, parts of the MEG recordings remain to be explained by future hypotheses. We hope neuroscientists will build on our method to formulate and test such future hypotheses. We make our code publicly available to facilitate this.
- We perform all learning in a zero-shot setting, in which neither the stimulus nor the question used to evaluate the learned models is seen during training (i.e. not just as the specific stimulus-question pair but also in combination with any other question/stimulus). Note that this is not the case in previous work that examines task effects, and we are the first to demonstrate how zero-shot learning can be applied successfully to this question. This is important

for scientific discovery because it can test the generalization of the results beyond the experimental stimuli and tasks.

- We show that models that integrate task and stimulus representations have significantly higher prediction performance than models that do not account for the task semantics, and localize the effect of task semantics largely to time-windows in $475 - 650$ms after the stimulus presentation.

Chapter 6

Interpreting and Improving NLP Models Using Brain Recordings

This chapter is based on work published as:
Mariya Toneva and Leila Wehbe. "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)". In: *Advances in Neural Information Processing Systems*. 2019, pp. 14928–14938

In this chapter, we develop a method that uses prior neurolinguistic evidence to evaluate the presence of specific brain-relevant information in the representations of an NLP model. The method presents the same text word-by-word to a person in a neuroimaging device and an NLP model, and measures how well the network-derived representations align with the brain recordings in relevant brain regions. This work showed that we can use this method and a snapshot of brain activity, captured by functional magnetic resonance imaging, to reveal how much context is encoded in the representations derived from 4 popular pretrained NLP models. As the provided context length increased, the activity in context-processing brain regions best aligned with representations from a model that uses both recurrence and self-attention, suggesting that this model is best able to encode long-range context. We further showed that altering a state-of-the-art pretrained model to better predict fMRI recordings also significantly improved this model's ability to generalize to a different data distribution from the one it was trained on. These results are the first evidence that fMRI recordings of people reading can be used to improve the generalization performance of a neural network NLP model.

6.1 Introduction

The large success of deep neural networks in NLP is perplexing when considering that unlike most other NLP approaches, neural networks are typically not informed by explicit language rules. Yet, neural networks are constantly breaking records in various NLP tasks from machine translation to sentiment analysis. Even more interestingly, it has been shown that word embeddings and language models trained on a large generic corpus and then optimized for downstream NLP tasks produce

even better results than training the entire model only to solve this one task (Peters et al., 2018; Howard et al., 2018; Devlin et al., 2018). These models seem to capture something generic about language. What representations do these models capture of their language input?

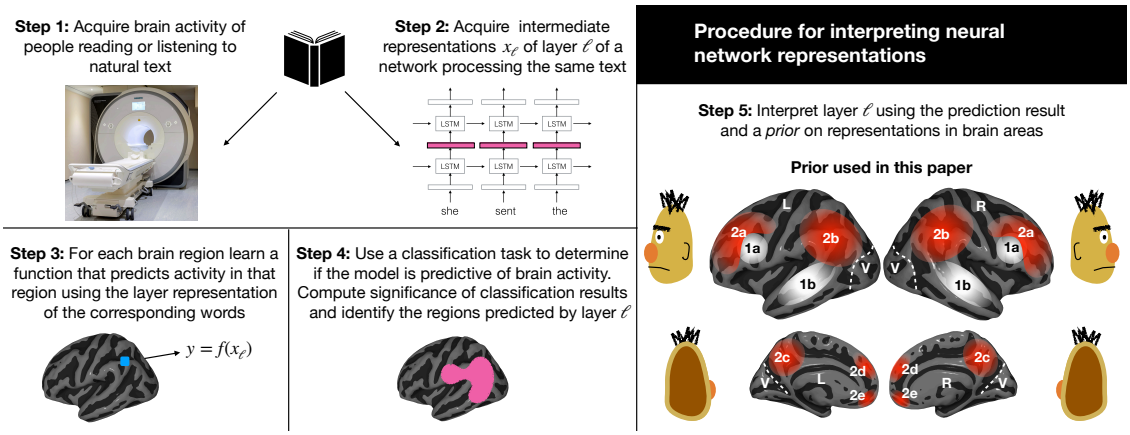


Figure 6.1: Diagram of approach and prior on brain function. The prior was constructed using the results of Lerner et al. (2011): regions in group 1 (white) process information related to isolated words and word sequences while group 2 (red) process only information related to word sequences (see Section 6.1.1). V indicates visual cortex. The drawing indicates the views of the brain with respect to the head. See Table 6.1 for names of brain areas.

Different approaches have been proposed to probe the representations in the network layers through NLP tasks designed to detect specific linguistic information (Conneau et al., 2018; Zhu, Li, et al., 2018; Linzen et al., 2016). Other approaches have attempted to offer a more theoretical assessment of how recurrent networks propagate information, or what word embeddings can represent (Peng et al., 2018; Chen et al., 2017; Weiss et al., 2018). Most of this work has been centered around understanding the properties of sequential models such as LSTMs and RNNs, with considerably less work focused on non-sequential models such as transformers.

Using specific NLP tasks, word annotations or behavioral measures to detect if a type of information is present in a network-derived representation (such as a word embedding of an LSTM or a state vector of a transformer) can be informative. However, complex and arguably more interesting aspects of language, such as high level meaning, are difficult to capture in an NLP task or in behavioral measures. We therefore propose a novel approach for interpreting neural networks that relies on the only processing system we have that does understand language: the human brain. Indeed, the brain does represent complex linguistic information while processing language, and we can use brain activity recordings as a proxy for these representations. We can then relate the brain representations with neural network representations by learning a mapping from the latter to the former. We refer to this analysis as aligning the neural network representations with brain activity.

1a	Inferior Frontal Gyrus
1b	Middle/Superior Temporal
2a	Lateral Middle/Superior Frontal
2b	Supramarginal Gyrus / Posterior Superior Temporal / Angular Gyrus
2c	Precuneus
2d	Medial Superior Frontal
2e	Medial Orbito-Frontal

Table 6.1: Name of regions of interest visualized in Figure 6.1. Regions were approximated from the results of (Lerner et al., 2011).

6.1.1 Proposed approach

We propose to look at brain activity of subjects reading naturalistic text as a source of additional information for interpreting neural networks. We use fMRI (functional Magnetic Resonance Imaging) and Magnetoencephalography (MEG) recordings of the brain activity of these subjects as they are presented text one word at a time. We present the same text to the NLP model we would like to investigate and extract representations from the intermediate layers of the network, given this text. We then learn an alignment between these extracted representations and the brain recordings corresponding to the same words to offer an evaluation of the information contained in the network representations. Evaluating neural network representations with brain activity is a departure from existing studies that go the other way, using such an alignment to instead evaluate brain representations (Wehbe, Vaswani, et al., 2014; Frank et al., 2015; Hale et al., 2018; Jain et al., 2018).

To align a layer ℓ representation with brain activity, we first learn a model that predicts the fMRI or MEG activity in every region of the brain (fig. 6.1). We determine the regions where this model is predictive of brain activity using a classification task followed by a significance test. If a layer representation can accurately predict the activity in a brain region r, then we conclude that the layer shares information with brain region r. We can thus make conclusions about the representation in layer ℓ based on our prior knowledge of region r.

Brain recordings have inherent, meaningful structure that is absent in network-derived representations. In the brain, different processes are assigned to specific locations as has been revealed by a large array of fMRI experiments. These processes have specific latencies and follow a certain order, which has been revealed by electrophysiology methods such as MEG. In contrast to the brain, a network-derived representation might encode information that is related to multiple of these processes without a specific organization. When we align that specific network representation with fMRI and MEG data, the result will be a decomposition of the representation into parts that correspond to different processes and should therefore be more interpretable. We can think of alignment with brain activity as a "demultiplexer" in which a single input (the network-derived representation) is decomposed into multiple outputs (relationship with different brain processes).

There doesn't yet exist a unique theory of how the brain processes language that researchers agree upon (Hickok et al., 2007; Friederici, 2011; Hagoort, 2003). Because we don't know which

of the existing theories are correct, we abandon the theory-based approach and adopt a fully data-driven approach. We focus on results from experiments that use naturalistic stimuli to derive our priors on the function of specific brain areas during language processing. These experiments have found that a set of regions in the temporo-parietal and frontal cortices are activated in language processing (Lerner et al., 2011; Wehbe, Murphy, et al., 2014; Huth, Heer, et al., 2016; Blank et al., 2017) and are collectively referred to as the language network (Fedorenko and Thompson-Schill, 2014). Using the results of Lerner et al. (2011) we subdivide this network into two groups of areas: group 1 is consistently activated across subjects when they listen to disconnected words or to complex fragments like sentences or paragraphs and group 2 is consistently activated only when they listen to complex fragments. We will use group 1 as our prior on brain areas that process information at the level of both short-range context (isolated words) and long-range context (multi-word composition), and group 2 as a prior on areas that process long-range context only. Fig. 6.1 shows a simple approximation of these areas on the Montreal Neurological Institute (MNI) template, and the names of the corresponding brain areas are presented in Table 6.1. Inspection of the results of Jain et al. (2018) shows they corroborate the division of language areas into group 1 and group 2. Because our prior relies on experimental results and not theories of brain function, it is data-driven.

We use this setup to investigate a series of questions about the information represented in different layers of neural network models. We explore four recent models: ELMo, a language model by Peters et al. (2018), BERT, a transformer by Devlin et al. (2018), USE (Universal Sentence Encoder), a sentence encoder by Cer et al. (2018), and T-XL (Transformer-XL), a transformer that includes a recurrence mechanism by Dai et al. (2019). We investigate multiple questions about these networks. Is word-level specific information represented only at input layers? Does this differ across recurrent models, transformers and other sentence embedding methods? How many layers do we need to represent a specific length of context? Is attention affecting long range or short range context?

Intricacies As a disclaimer, we warn the reader that one should be careful while dealing with brain activity. Say a researcher runs a task T in fMRI (e.g. counting objects on the screen) and finds it activates region R, which is shown in another experiment to also be active during process P (e.g. internal speech). It is seductive to then infer that process P is involved during task T. This "reverse inference" can lead to erroneous conclusions, as region R can be involved in more than one task (Poldrack, 2006). To avoid this trap, we only interpret alignment between network-derived representations and brain regions if (1) the function of the region is well studied and we have some confidence on its function during a task similar to ours (e.g. the primary visual cortex processing letters on the screen or group 2 processing long range context) or (2) we show a brain region has overlap in the variance explained by the network-derived layer and by a specific process, in the same experiment. We further take sound measures for reporting results: we cross-validate our models and report results on unseen test sets. Another possible fallacy is to directly compare the performance of layers from different networks and conclude that one network performs better than the other: information is likely organized differently across networks and such comparisons are misleading. Instead we only perform controlled experiments where we look at one network

and vary one parameter at a time, such as context length, layer depth or attention type.

6.2 Related work on brains and language

Most work investigating language in the brain has been done in a controlled experiment setup where two conditions are contrasted (Friederici, 2011). These conditions typically vary in complexity (simple vs. complex sentences), vary in the presence or absence of a linguistic property (sentences vs. lists of words) or vary in the presence or absence of incongruities (e.g. semantic surprisal) (Friederici, 2011). A few researchers instead use naturalistic stimulus such as stories (Brennan et al., 2010; Lerner et al., 2011; Speer et al., 2009; Wehbe, Murphy, et al., 2014; Huth, Heer, et al., 2016; Blank et al., 2017). Some use predictive models of brain activity as a function of multi-dimensional features spaces describing the different properties of the stimulus (Wehbe, Murphy, et al., 2014; Huth, Heer, et al., 2016).

A few previous works have used neural network representations as a source of feature spaces to model brain activity. Wehbe, Vaswani, et al. (2014) aligned the MEG brain activity we use here with a Recurrent Neural Network (RNN), trained on an online archive of Harry Potter Fan Fiction. The authors aligned brain activity with the context vector and the word embedding, allowing them to trace sentence comprehension at a word-by-word level. Jain et al., 2018 aligned layers from a Long Short-Term Memory (LSTM) model to fMRI recordings of subjects listening to stories to differentiate between the amount of context maintained by each brain region. Other approaches rely on computing surprisal or cognitive load metrics using neural networks to identify processing effort in the brain, instead of aligning entire representations (Frank et al., 2015; Hale et al., 2018).

There is little prior work that evaluates or improves NLP models through brain recordings. Søgaard (2016) proposes to evaluate whether a word embedding contains cognition-relevant semantics by measuring how well they predict eye tracking data and fMRI recordings. Fyshe, Talukdar, et al. (2014) build a non-negative sparse embedding for individual words by constraining the embedding to also predict brain activity well and show that the new embeddings better align with behavioral measures of semantics.

6.3 Approach

Network-derived Representations The approach we propose in this paper is general and can be applied to a wide variety of current NLP models. We present four case-studies of recent models that have very good performance on downstream tasks: ELMo, BERT, USE and T-XL.

- ELMo is a bidirectional language model that incorporates multiple layers of LSTMs. It can be used to derive contextualized embeddings by concatenating the LSTM output layers at that word with its non-contextualized embedding. We use a pretrained version of ELMo with 2 LSTM layers provided by Gardner et al. (2018).
- BERT is a bidirectional model of stacked transformers that is trained to predict whether a given sentence follows the current sentence, in addition to predicting a number of input words that have been masked (Devlin et al., 2018). Upon release, this recent model achieved

state of the art across a large array of NLP tasks, ranging from question answering to named entity recognition. We use a pretrained model provided by Hugging Face [1]. We investigate the base BERT model, which has 12 layers, 12 attention heads, and 768 hidden units.

- USE is a method of encoding sentences into an embedding (Cer et al., 2018) using a task similar to Skip-thought (Kiros et al., 2015). USE is able to produce embeddings in the same space for single words and passages of text of different lengths. We use a version of USE from tensorflow hub trained with a deep averaging network [2] that has 512 dimensions.
- T-XL incorporates segment level recurrence into a transformer with the goal of capturing longer context than either recurrent networks or usual transformers (Dai et al., 2019). We use a pretrained model provided by Hugging Face[1], with 19 layers and 1024 hidden units.

We investigate how the representations of all four networks change as we provide varying lengths of context. We compute the representations $x_{\ell,k}$ in each available intermediate layer ($\ell \in \{1, 2\}$ for ELMo; $\ell \in \{1, ..12\}$ for BERT; ℓ is the output embedding for USE; $\ell \in \{1, ..19\}$ for T-XL). We compute $x_{l,k}$ for word w_n by passing the most recent k words $(w_{n-k+1}, .., w_n)$ through the network.

fMRI and MEG data In this paper we use fMRI and MEG data which have complementary strengths. fMRI is sensitive to the change in oxygen level in the blood that is a consequence to neural activity, it has high spatial resolution (2-3mm) and low temporal resolution (multiple seconds). MEG measures the change in the magnetic field outside the skull due to neural activity, it has low spatial resolution (multiple cm) and high temporal resolution (up to 1KHz). We use fMRI data published by Wehbe, Murphy, et al. (2014). 8 subjects read chapter 9 of *Harry Potter and the Sorcerer's stone* Rowling (2012) which was presented one word at a time for a fixed duration of 0.5 seconds each, and 45 minutes of data were recorded. The fMRI sampling rate (TR) was 2 seconds. The same chapter was shown by Wehbe, Vaswani, et al. (2014) to 3 subjects in MEG with the same rate of 0.5 seconds per word. Details about the data and preprocessing can be found in Chapter 2.3.2.

Encoding models For each type of network-derived representation $x_{\ell,k}$, we estimate an encoding model that takes $x_{\ell,k}$ as input and predicts the brain recording associated with reading the same k words that were used to derive $x_{\ell,k}$. Chapter 3.3.2 provides full details about how we train these encoding models. Briefly, we estimate a function f, such that $f(x_{l,k}) = y$, where y is the brain activity recorded with either MEG or fMRI. We follow previous work (Sudre et al., 2012b; Wehbe, Murphy, et al., 2014; Wehbe, Vaswani, et al., 2014; Nishimoto et al., 2011a; Huth, Heer, et al., 2016) and model f as a linear function, regularized by the ridge penalty. The model is trained via four-fold cross-validation and the regularization parameter is chosen via nested cross-validation.

Evaluation of predictions We evaluate the predictions from each encoding model by using them in a classification task on held-out data, in the four-fold cross-validation setting. The classifica-

[1]https://github.com/huggingface/pytorch-pretrained-BERT/
[2]https://tfhub.dev/google/universal-sentence-encoder/2

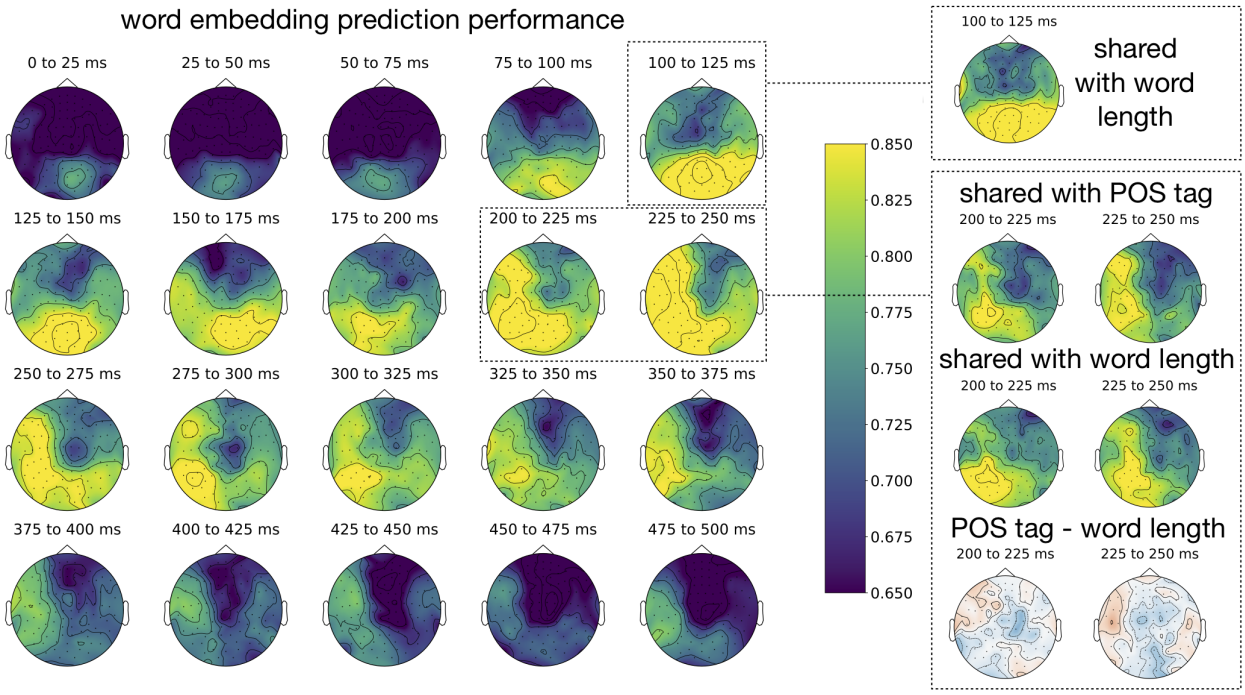


Figure 6.2: 20v20 accuracy at predicting MEG activity at each sensor location and time point using the ELMo word embedding, compared with the performance shared with word length and Part-Of-Speech (POS) tags. Around 200-250ms, the word embedding predicts a part of the activity at the top of the helmet, and this is shared mostly with the POS tags and not with word length (see bottom-right comparison). Indeed, we know from electrophysiology studies studies that POS violations incur a response around 200ms after word onset in the front of the brain (Frank et al., 2015), which aligns with our analysis. We hypothesize from these results that the word-embedding contains both word length and POS information.

tion task is to predict which of two sets of words was being read based on the respective feature representations of these words (Mitchell, Shinkareva, et al., 2008; Wehbe, Murphy, et al., 2014; Wehbe, Vaswani, et al., 2014). This task is performed between sets of 20 consecutive TRs in fMRI (accounting for the slowness of the hemodynamic response), and sets of 20 randomly sampled words in MEG. The classification is repeated a large number of times and an average classification accuracy is obtained for each voxel in fMRI and for each sensor/timepoint in MEG. We refer to this accuracy of matching the predictions of an encoding model to the correct brain recordings as "prediction accuracy". The final fMRI results are reported on the MNI template, and we use pycortex to visualize them (Gao, Huth, et al., 2015).

We use a new empirical based method to compute statistical significance that relies on the distribution of average accuracies over a subject's brain to estimate the False Discovery Proportion (FDP). The voxel accuracies belong to two distributions: either the voxel has chance accuracy or the voxel is truly predicted by the corresponding layer ℓ. Average chance accuracy for our binary balanced task is 0.5, however the accuracies due to chance performance might have a varying

distribution around 0.5. The accuracies above 0.5 are a mixture of predicted voxels and voxels with chance performance. We assume that chance performance is symmetrically distributed around 0.5, and we use the set of accuracies that are less than 0.5–which we consider to be in the chance distribution–to estimate the distribution of chance accuracies above 0.5. We want to find a set of voxels where to reject the null hypothesis such that the FDP is ≤ 0.05. For that purpose we find the smallest margin δ, $0 < \delta < 0.5$ such that:

$$\widehat{\text{FDP}} = \frac{1 + \#\{\text{voxel } s.t. \text{ accuracy} \leq 0.5 - \delta\}}{1 \vee \#\{\text{voxel } s.t. \text{ accuracy} \geq 0.5 + \delta\}} \leq q$$

where $q = 0.05$, by starting at $\delta = 0.001$ and increasing it in increments of 0.001, stopping when $\widehat{\text{FDP}} \leq 0.05$ or the limit is reached. This approach is adapted from the Barber-Candès approach which has been proposed and analyzed by Barber et al., 2015; Arias-Castro et al., 2017; Rabinovich et al., 2017, and shown to control the False Discovery Rate (FDR) at level q when δ_{final} is chosen as a threshold. We reject the null hypothesis for all voxels where the accuracy is $\geq 0.5 + \delta_{\text{final}}$.

To combine results across different subjects, we use pycortex (Gao, Huth, et al., 2015) to transform each subject to the Montreal Neurological Institute (MNI) space, the most commonly used template space in fMRI. We can then average the results of different participants. See Chapter 3.3.4 for more details about our methods.

Proof of concept Since MEG signals are faster than the rate of word presentation, they are more appropriate to study the components of word embeddings than the slow fMRI signals that cannot be attributed to individual words. We know that single word non-contextualized embeddings likely have information about the part-of-speech (POS) and the length of a word. We will show here how our approach can recover this information from brain activity as a proof-of-concept. We know from the Neuroscience literature that MEG activity can be related to the length of the current word (Sudre et al., 2012b) and its part of speech (Frank et al., 2015) at different times. We investigate whether word length and part-of-speech (POS) information is also present in the non-contextualized embedding by computing the shared performance ($A \cap B$) between the pairs of features (A and B) as $A + B - A \cup B$ as explained in the previous section.

We present the results in Figure 6.2. The current word embedding is able to predict activity as the current word is being perceived starting at the back of the sensor helmet (generally on top of the visual cortex) around 100ms. This is when we expect the visual signal to start reaching the visual cortex. Indeed, we see that the word-embedding and the word length have overlap in the activity they predict in the visual cortex at that time. Gradually, the areas predicted by the word embedding move forward in the brain towards areas known to be involved in more high level aspects of reading. Around 200-250ms, we see the word embedding predicts a part of the activity at the top of the helmet, and this is shared mostly with the POS tags and not with word length (see bottom-right comparison). Indeed, we know from electrophysiology studies studies that POS violations incur a response around 200ms after word onset in the front of the brain Frank et al., 2015, which aligns with our analysis. From these results we can hypothesize that the word-embedding contains both word length and POS information, as was expected.

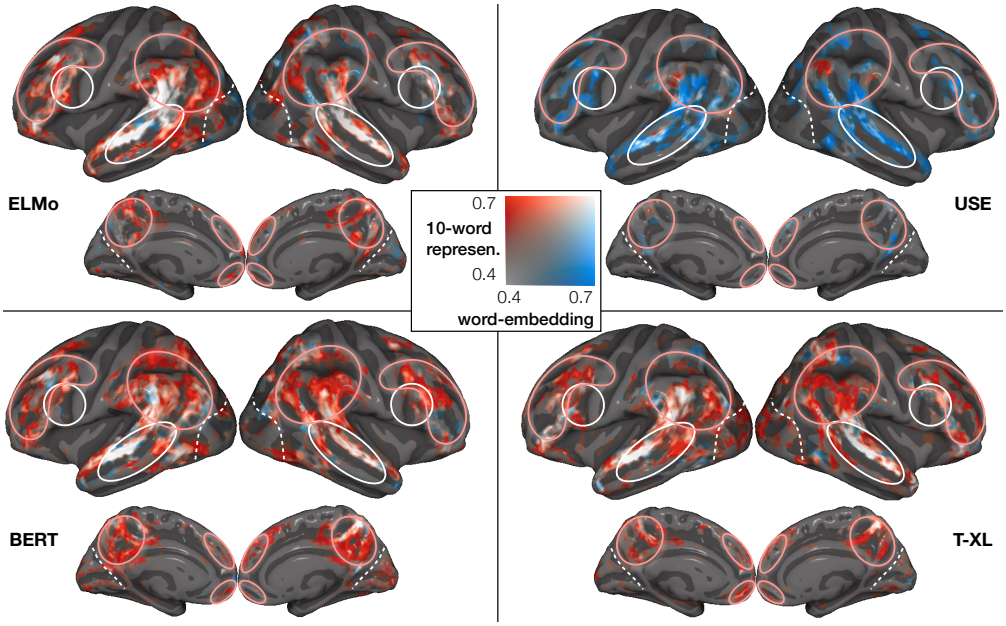


Figure 6.3: Comparison between the prediction performance of two network representations from each model: a 10-word representation corresponding to the 10 most recent words shown to the participant (Red) and a word-embedding corresponding to the last word (Blue). Areas in white are well predicted from both representations. These results align to a fair extent with our prior: group 2 areas (red outlines) are mostly predicted by the longer context representations while areas 1b (lower white outlines) are predicted by both word-embeddings and longer context representations.

6.4 Interpreting long-range contextual representations

Integrated contextual information in ELMo, BERT, and T-XL One question of interest in NLP is how successfully a model is able to integrate context into its representations. We investigate whether the four NLP models we consider are able to create an integrated representation of a text sequence by comparing the performance of encoding models trained with two kinds of representations: a token-level word-embedding corresponding to the most recent word token a participant was shown and a 10-word representation corresponding to the 10 most recent words. For each of the models with multiple layers (all but USE), this 10-word representation was derived from a middle layer in the network (layer 1 in ELMo, layer 7 in BERT, and layer 11 in T-XL). We present the qualitative comparisons across the four models in figure 6.3, where only significantly predicted voxels for each of the 8 subjects were included with the false discovery rate controlled at level 0.05 (see section 3 of supplementary materials for more details). We provide a quantitative summary of the observed differences across models for the 1b regions and group 2 regions in Figure6.4. We observe similarities in the word-embedding performances across all models, which all predict the brain activity in the left and right group 1b regions and to some extent in group 1a regions. We also observe differences in the longer context representations between USE and the rest of the models:

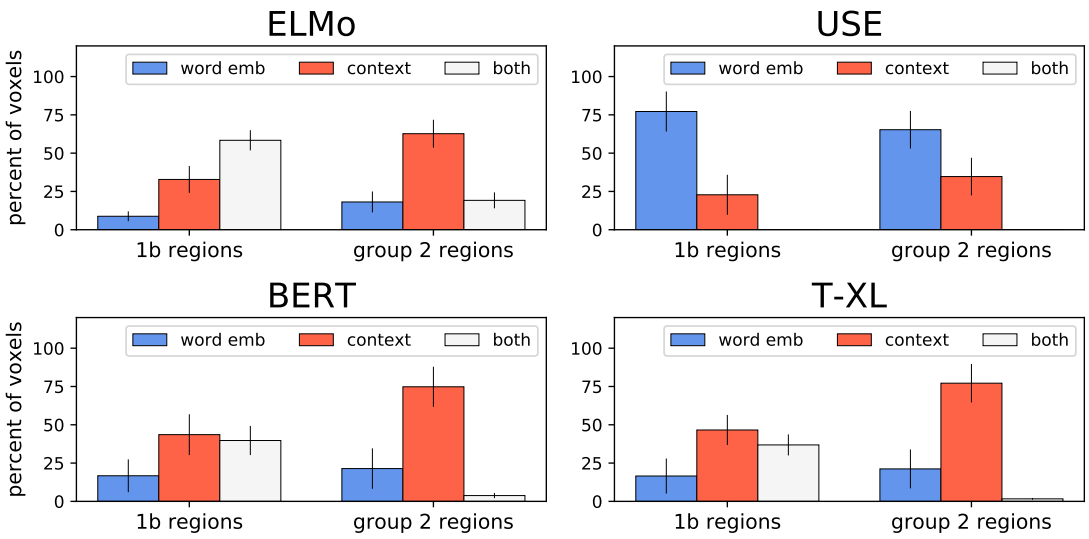


Figure 6.4: Amount of group 1b regions and group 2 regions predicted well by each network-derived representation: a 10-word representation corresponding to the 10 most recent words shown to the participant (Red) and a word-embedding corresponding to the last word (Blue). White indicates that both representations predict the specified amount of the regions well (about 0.7 threshold). We present the mean and standard error of the percentage of explained voxels within the specified regions over all participants.

- ELMo, BERT, and T-XL long context representations predict subsets of both group 1 regions and group 2 regions. Most parts that are predicted by the word-embedding are also predicted by the long context representations (almost no blue voxels). We conclude that the long context representations most probably include information about the long range context and the very recent word embeddings. These results may be due to the fact that all these models are at least partially trained to predict a word at a given position. They must encode long range information and also local information that can predict the appropriate word.
- USE long context representations predict the activity in a much smaller subset of group 2 regions. The low performance of the USE vectors might be due to the deep averaging which might be composing words in a crude manner. The low performance in predicting group 1 regions is most probably because USE computes representations at a sentence level and does not have the option of retaining recent information like the other models. USE long context representations therefore only have long range information.

Relationship between layer depth and context length We investigate how the performances of ELMo, BERT, and T-XL change at different layers as they are provided varying size of contexts. The results are shown in figure 6.5. We observe that in all networks, the middle layers perform the best for contexts longer than 15 words. In addition, the deepest layers across all networks show a sharp increase in performance at short-range context (fewer than 10 words), followed by

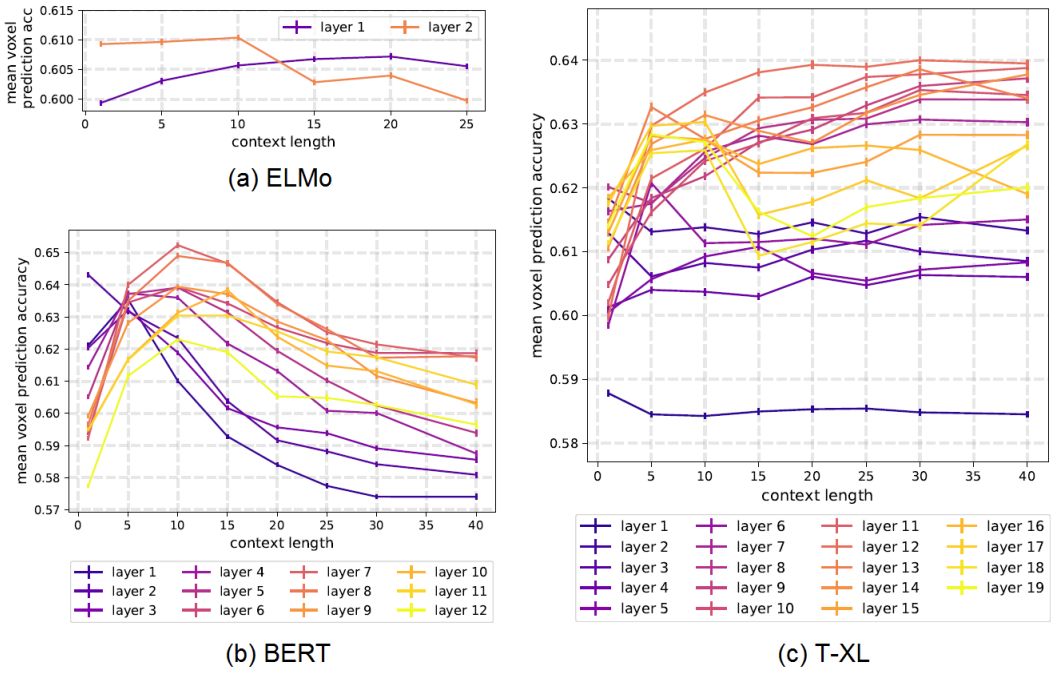


Figure 6.5: Performance of encoding models for all hidden layers in ELMo, BERT, and T-XL as the amount of context provided to the network is increased. Transformer-XL is the only model that continues to increase performance as the context length is increased. In all networks, the middle layers perform the best for contexts longer than 15 words. The deepest layers across all networks show a sharp increase in performance at short-range context (fewer than 10 words), followed by a decrease in performance.

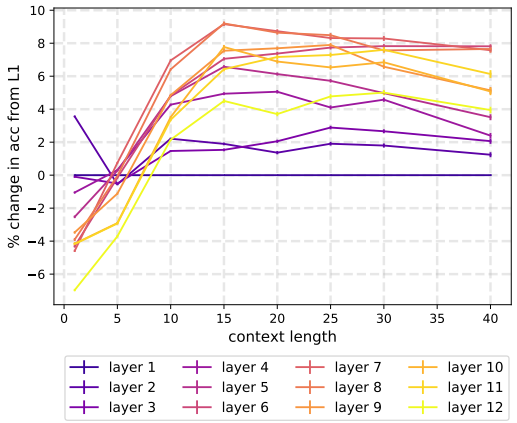


Figure 6.6: Change in encoding model performance of BERT layers from the performance of the first layer. When we adjust for the performance of the first layer, the performance of the remaining layers resemble that of T-XL more closely, as shown in Figure 6.5.

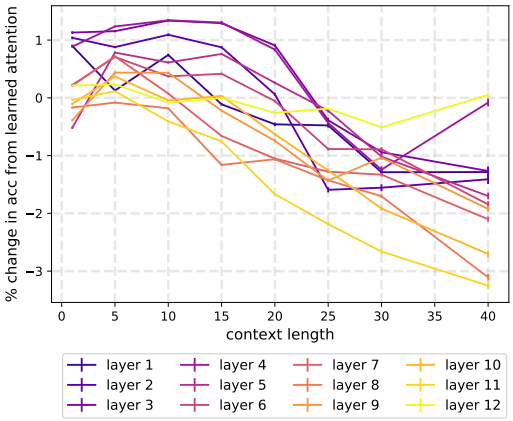


Figure 6.7: Change in encoding model performance of BERT layer l when the attention in layer l is made uniform. The performance of deep layers, other than the output layer, is harmed by the change in attention. Shallow layers benefit from the uniform attention for context lengths up to 25 words.

a decrease in performance. We further observe that T-XL is the only model that continues to increase performance as the context length is increased. T-XL was designed to represent long range information better than a usual transformer and our results suggest that it does. Finally, we observe that layer 1 in BERT behaves differently from the first layers in the other two networks. In figure 6.6, we show that when we instead examine the increase in performance of all subsequent layers from the performance of the first layer, the resulting context-layer relationships resemble the ones in T-XL. This suggests that BERT layer 1 combines the information from the token-level embeddings in a way that limits the retention of longer context information in the layer 1 representations.

Effect of attention on layer representation We further investigate the effect of attention across different layers by measuring the negative impact that removing its learned attention has on its brain prediction performance. Specifically we replaced the learned attention with uniform attention over the representations from the previous layer. More concretely, to alter the attention pattern at a single layer in BERT, for each attention head $h_i = Attn_i(QW_i^Q, KW_i^K, VW_i^V)$, we replace the pretrained parameter matrices W_i^Q, W_i^K, and W_i^V for this layer, such that the attention $Attn(Q, K, V)$, defined as $softmax(QK/\sqrt{d_k})^T V$ (Vaswani et al., 2017), yields equal probability over the values in value matrix V (here d_k denotes the dimensionality of the keys and queries). To this end, for a single layer, we replace W_i^Q and W_i^K with zero-filled matrices and W_i^V with the identity matrix. We only alter a single layer at a time, while keeping all other parameters of the pretrained BERT fixed. In figure 6.7, we present the change in performance of each layer with

uniform attention when compared to pretrained attention. The performance of deep layers, other than the output layer, is harmed by the change in attention. However, surprisingly and against our expectations, shallow layers benefit from the uniform attention for context lengths up to 25 words.

6.5 Applying insight from brain interpretations to NLP tasks

The observation that the layers in the first half of the base BERT model benefit from uniform attention for predicting brain activity was surprising because it suggests that applying uniform attention in the shallow layers of BERT allows BERT to encode brain-relevant language information better than the original BERT model, which was trained on millions of text documents. Because this observation was so surprising, we wanted to test the uniform-attention BERT models' understanding of language further using a linguistic task. We settled on a linguistic task that has previously been used to quantify BERT's syntactic understanding (Goldberg, 2019). This task evaluates whether an NLP model is able to predict the correct verb in a sentence that agrees with the number of the subject in the sentence (i.e. subject-verb agreement), across different types of sentences in a dataset introduced by Marvin et al., 2018. For example, in the sentence "the game that the guard hates [MASK] bad.", the task is to predict the verb that occurs in the masked position. The performance is evaluated based on whether the NLP model puts higher probability on the correct verb ("is" in the example) versus the incorrect verb ("are" in the example). This is the same task that BERT is originally trained to do–to predict a word in a specific context–but the data distribution which we use to evaluate the model is very different from the one BERT is trained on, so here we are specifically testing its ability to generalize to a new data distribution. Examples of all types of sentences in the dataset are shown in Table 6.2.

We adopt the evaluation protocol of Goldberg (2019), in which BERT is first fed a complete sentence where the single focus verb is masked (e.g. `[CLS] the game that the guard hates [MASK] bad .`), then the prediction for the masked position is obtained using the pretrained language-modeling head, and lastly the accuracy is obtained by comparing the scores for the original correct verb (e.g. `is`) to the score for the incorrect verb (i.e. the verb that is wrongly numbered) (e.g. `are`). We make the attention in layers 1 through 6 in base BERT uniform, a single layer at a time while keeping the remaining parameters fixed as described in Section 6.4, and evaluate on the 13 sentence types. We present the results of altering layers 1,2, and 6 in Table 6.3. We observe that the altered models significantly outperform the pretrained model ('base') in 8 of the 13 sentence types and achieve parity in 4 of the remaining 5 types (paired t-test, significance level 0.01, FDR controlled for multiple comparisons (Benjamini et al., 1995)). Performance of altering layers 3-5 is similar and is presented in Supplementary Table D.1. We contrast the performance of these layers with that of a model with uniform attention at layer 11, which is the model that suffers the most from this change for predicting the brain activity as shown in Figure 6.7. We observe that this model also performs poorly on the linguistic task as it performs on par or worse than the base model in 12 of the 13 types of sentences.

Subject-verb agreement type	Examples of (*in)correct sentences
Simple	The author **laughs/*laugh**.
In a sentential complement	The bankers knew the officer **smiles/*smile**.
Short VP coordination	The senator smiles and **laughs/*laugh**.
Long VP coordination	The manager writes every day and **likes/*like** to watch TV.
Across a prepositional phrase	The farmer near the parents **smiles/*smile**.
Across a subject relative clause	The officers that love the skater **smile/*smiles**.
Across an object relative clause	The farmer that the parents love **swims/*swim**.
Across an object relative (no that)	The farmer the parents love **swims/*swim**.
In an object relative clause	The farmer that the parents **love/*loves** swims.
In an object relative (no that)	The farmer the parents **love/*loves** swims.
Reflexive anaphora	
Simple	The senators embarrassed **themselves/*herself**.
In a sentential complement	The bankers thought the pilot embarrassed **himself/*themselves**.
Across a relative clause	The manager that the architects like doubted **himself/*themselves**.

Table 6.2: Types of sentences in the dataset introduced by Marvin et al., 2018, along with example sentences. The correct and incorrect word pairs in each example appear in bold, with the incorrect word signified by a preceding asterisk.

6.6 Discussion

We introduced an approach to use brain activity recordings of subjects reading naturalistic text to interpret different representations derived from neural networks. We used MEG to show that the (non-contextualized) word embedding of ELMo contains information about word length and part of speech as a proof of concept. We used fMRI to show that different network representation (for ELMo, USE, BERT, and T-XL) encode information relevant to language processing at different context lengths. USE long-range context representations perform differently from the other model and do not also include short-range information. The transformer models (BERT and T-XL) both capture the most brain-relevant context information in their middle layers. T-XL, by combining both recurrent properties and transformer properties, has representations that don't degrade in performance when very long context is used, unlike purely recurrent models (e.g. ELMo) or transformers (e.g. BERT).

We found that applying uniform attention in the shallow layer actually improved the brain prediction performance over using the original pre-trained attention. After this observation, we tested how the same alterations affect BERT's ability to generalize to a new data distribution from a linguistic syntactic task. We observed that the uniform-attention BERT significantly outperforms pretrained BERT across the majority of sentence types in the linguistic task. This result suggests that altering an NLP model to better align with brain recordings of people processing language may lead to better language understanding by the NLP model.

condition	uni L1	uni L2	uni L6	uni L11	base	count
simple	**1.00**	**1.00**	**1.00**	0.98	**1.00**	120
in a sentential complement	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**	1440
short VP coordination	0.88	0.90	**0.91**	0.88	0.89	720
long VP coordination	0.96	0.97	**1.00****	0.96	0.98	400
across a prepositional phrase	0.86	**0.93****	0.88	0.82	0.85	19440
across a subject relative clause	0.83	0.83	**0.85****	0.83	0.84	9600
across an object relative clause	0.87	0.91	**0.92****	0.86	0.89	19680
across an object relative clause (no that)	**0.87**	0.80	**0.87**	0.84	0.86	19680
in an object relative clause	**0.97****	0.95	0.91	0.93	0.95	15960
in an object relative clause (no that)	**0.83****	0.72	0.74	0.72	0.79	15960
reflexive anaphora: simple	0.91	0.94	**0.99****	0.95	0.94	280
reflexive anaphora: in a sent. complem.	0.88	0.85	0.86	0.85	**0.89**	3360
reflexive anaphora: across rel. clause	0.79	**0.84****	0.79	0.76	0.80	22400

Table 6.3: Performance of models with altered attention on subject-verb agreement across various sentence types (tasks by Marvin et al. (2018)). Best performance per task is made bold, and marked with ** when difference from 'base' performance is statistically significant. The altered models for the shallow layers significantly outperform the pretrained model ('base') in 8 of the 13 tasks and achieve parity in 4 of the remaining 5 tasks.

Future work We hope that as naturalistic brain experiments become more popular and data more widely shared, aligning brain activity with neural network will become a research area. Our next steps are to expand the analysis using MEG to uncover new aspects of word-embeddings and to derive more informative fMRI brain priors that contain specific conceptual information that is linked to brain areas, and use them to study the high level semantic information in network representations.

6.7 Takeaways

The contributions of this chapter can be summarized as follows:

1. We present a new method to interpret network representations and a proof of concept for it.
2. We use our method to analyze and provide hypotheses about 4 popular pretrained models–ELMo, BERT, USE and Transformer-XL.
3. We find the middle layers of transformers are better at predicting brain activity than other layers. We find that Transformer-XL's performance doesn't degrade as context is increased, unlike the other models'. We find that using uniform attention in early layers of BERT (removing the pretrained attention on the previous layer) leads to better prediction of brain activity.
4. We show that when BERT is altered to better align with brain recordings (by removing the pretrained attention in the shallow layers), it is also able to perform better at linguistic tasks

that probe its syntactic understanding (Marvin et al., 2018) on a new data distribution. This result presents the first evidence that brain recordings of people understanding language can be used to improve the generalization performance of a neural network NLP model.

Chapter 7

Conclusion

This dissertation presented a data-driven framework that establishes a direct connection between brain recordings of people comprehending language and natural language processing (NLP) systems. We presented evidence that this connection can be beneficial for both neurolinguistics and NLP. Specifically, we showed that this framework can utilize recent successes in neural networks for NLP to enable scientific discovery about context- and task-dependent meaning composition in the brain, and we presented the first evidence that brain activity measurements of people reading can be used to improve the generalization performance of a popular deep neural network language model. These investigations also contributed advances in cognitive modeling that may be useful beyond the study of language. In short, this dissertation involved multidisciplinary investigations and has made contributions to cognitive neuroscience, neurolinguistics, and natural language processing.

7.1 Summary of Contributions

7.1.1 Advances in Cognitive Modeling

Improved scientific inference for encoding models with rich stimulus representations. Cognitive neuroscience has taken a precipitous dive into relating stimuli representations derived from a neural network to brain recordings, largely utilizing encoding models that are trained to predict the brain recordings as a function of the neural network representations. In language alone, a quickly growing number of studies have used neural networks to study single word meaning, word sequence meaning, or different levels of language processing (Wehbe, Vaswani, et al., 2014; Jain et al., 2018; Toneva and Wehbe, 2019; Abnar et al., 2019; Beinborn et al., 2019; Hollenstein et al., 2019; Gauthier et al., 2019; Caucheteux and King, 2020). While we believe in the promise of this computational modeling approach, we do also recognize that it has limitations. As we move towards richer stimulus representations, it becomes important to reexamine what inferences we're able to make from our existing computational tools and adjust our toolbox accordingly. In Chapter 3, we first discuss limitations of encoding models when trained as a function of rich stimulus representations and propose two new tools that, when used together, can allow us to make stronger

scientific inferences about what stimulus-related information is processed by a specific brain region or timepoint. These new tools are not language-specific and can benefit varied computational investigations that utilize rich stimulus representations, such as ones obtained from neural networks.

Improved scientific inference in naturalistic experiments. Naturalistic stimuli have taken the cognitive neuroscience world by storm (Sonkusare et al., 2019; Nastase, Goldstein, et al., 2020; Hamilton et al., 2020) and also make up the majority of the datasets (i.e. three of four datasets) that we investigate in this dissertation. This naturalistic setting enables studying processing that is more easily generalizable to the everyday world. However, due to correlations that occur in naturalistic stimuli, these experiments have difficulty inferring what characteristics of the stimulus are truly related to the observed brain recordings. In contrast to the naturalistic setting, traditional neuroscience experiments carefully control characteristics of the experimental stimuli that are not of interest (e.g. word length or frequency) so that any observed variation in the resulting brain recordings would be due to the characteristic that the neuroscientist is actually interested in studying (e.g. word surprisal). In this dissertation, we proposed a solution to this limitation of naturalistic experiments–introducing post-hoc computational controls. In Chapter 4, we applied computational controls to a naturalistic reading paradigm to disentangle the supra-word meaning from the meaning of individual words. Computational controls are a general approach that is not limited to the study of language or to using rich stimuli representations, such as ones from neural networks, and we hope that they will enable scientific inferences in varied naturalistic settings.

Modeling task-dependent processing. As we move towards more naturalistic experimental paradigms, it is important to consider how the task a participants is asked to perform affects the corresponding brain recordings. In Chapter 5, we provide the first methodology that predicts brain recordings as a function of both the observed stimulus and a question task. All learning is performed in a zero-shot setting, in which neither the stimulus nor the question used to evaluate the learned models is seen during training (i.e. not just as the specific stimulus-question pair but also in combination with any other question/stimulus). This work is the first to successfully apply zero-shot learning to this question, which is important because it tests the generalization of the results beyond the experimental stimuli and tasks. We hope that this methodology will encourage neuroscientists to formulate mechanistic computational hypotheses about the effect of a question on the processing of a stimulus, as a first step towards understanding the effect of goal-directed tasks on brain activity.

7.1.2 NLP $\rightarrow$ Neurolinguistics

Processing of supra-word meaning. In Chapter 4, we investigate the neural basis of a facet of composed meaning in language, that we term supra-word meaning (i.e. the multi-word meaning that is beyond the meaning of individual words). Using fMRI recordings, we reveal that hubs thought to process lexical-level meaning also maintain supra-word meaning, suggesting a common substrate for lexical and combinatorial semantics. However, surprisingly, we find that supra-word meaning is difficult to detect in MEG. The difference between the fMRI and MEG results suggests

that the processing of supra-word meaning may be based on neural mechanisms that are not related to synchronized cell firing, as is the MEG signal. This investigation was enabled by using powerful NLP systems as model organisms for language comprehension, which allowed us to create numerical representations of supra-word meaning.

fMRI and MEG reveal different aspects of meaning composition. In Chapter 4, we identify potential limitations on the type of information that is detectable in MEG. While high temporal imaging resolution is key to reaching a mechanistic level of understanding of language processing, our findings suggest that a modality other than MEG may be necessary to detect long-range contextual information. Further, the fact that an aspect of meaning can be predictive in one imaging modality and invisible in the other calls for caution while interpreting findings about the brain from one modality alone, as some parts of the puzzle are systematically hidden. Our results also suggest that the imaging modality may impact the ability to decode the contextualized meaning of words, which is central to brain-computer interfaces (BCI) that aim to decode attempted speech.

Task-dependent processing in question answering. In Chapter 5, we investigate the effect of a question task on the processing of a concrete noun by predicting the millisecond-resolution MEG brain activity as a function of both the semantics of the noun and the task. We show that incorporating the task semantics (i.e., the specific question asked) significantly improved the prediction of MEG recordings, across participants. The improvement occurs $475 - 550$ms after the participants first see the word, which corresponds to what is considered to be the ending time of semantic processing for a word. These results suggest that only the end of semantic processing of a word is task-dependent. This finding may inspire new NLP training algorithms or architectures that keep some computation task-independent, in contrast to current transfer learning approaches for NLP that tune all parameters of a pretrained model when training to perform a specific task.

7.1.3 Neurolinguistics $\rightarrow$ NLP

Interpreting information encoded by NLP systems. In Chapter 6, we develop a method that uses prior neurolinguistic evidence to evaluate the presence of specific brain-relevant information in the representations of an NLP model. The method presents the same text word-by-word to a person in a neuroimaging device and an NLP model, and measures how well the network-derived representations align with the brain recordings in relevant brain regions. This work showed that we can use this method and a snapshot of brain activity, captured by functional magnetic resonance imaging, to reveal how much context is encoded in the representations derived from 4 popular pretrained NLP models. As the provided context length increased, the activity in context-processing brain regions best aligned with representations from a model that uses both recurrence and self-attention, suggesting that this model is best able to encode long-range context.

Improved generalization in NLP systems. In Chapter 6, we further showed that altering a state-of-the-art pretrained model to better predict fMRI recordings also significantly improved performance on syntactic NLP tasks (e.g. subject-verb alignment) across a variety of sentences. These

results are the first evidence that fMRI recordings of people reading can be used to improve the generalization of a neural network NLP model. We hope that this evidence can serve as a first step towards future brain-guided NLP systems which would benefit from incorporating brain recordings directly during training.

7.2 Future Research Directions

These contributions inspire a multitude of research directions that we hope will impact both neurolinguistics and natural language processing in the future. Below is a summary of the directions that are of particular interest to the author.

Improvements in brain-guided NLP systems beyond generalization. In Chapter 6, we showed that an intervention in an NLP system that improved alignment with fMRI recordings also resulted in an NLP system that better generalized to a new data distribution. A natural next step is to incorporate brain recordings directly during the training of the NLP system. We have taken some initial steps towards this direction (Schwartz et al., 2019). One area that is still underexplored with respect to this direction is how to quantify the changes in the NLP system when incorporating brain recordings during training. Thus far we have considered quantifying these changes by measuring the generalization to a new data distribution within the same training task (Toneva and Wehbe, 2019) and the performance of transferring the trained NLP system to natural language understanding tasks (Schwartz et al., 2019) (i.e. the GLUE benchmark tasks (Wang, Singh, et al., 2018)). There may be other ways in which incorporating brain recordings during training may benefit the NLP system.

For instance, while humans are able to learn new linguistic concepts with very few examples, current NLP systems still require large amounts of experience and have trouble generalizing new concepts (Lake et al., 2019). There is some recent evidence that incorporating fMRI recordings into a prediction framework that aims to predict cognitive-relevant labels from complex naturalistic stimuli (i.e.. short movies) results in an increased ability of the model to transfer to new datasets (Nishida et al., 2019). One exciting direction is to investigate whether a brain-guided NLP system is more easily adaptable to new tasks.

Though Schwartz et al., 2019 showed that aligning BERT to the brain through fine-tuning on fMRI recordings did not result in significant gains on downstream natural language understanding tasks, it would be informative to observe how the accuracy on each task of both the vanilla and brain-aligned systems changes with the number of samples the systems see during training (i.e. quantify the sample efficiency). One prediction is that, if the brain-alignment has made the model representations more adaptive, the brain-aligned model will require fewer examples to reach the same level of accuracy as the vanilla model. Another interesting question is whether the brain-aligned system is better able to retain previously learned task-related information when trained on a different task. Much like other deep neural network-based systems, BERT has been shown to suffer from catastrophic forgetting of previously learned task-related information when trained on a different natural language processing task, without retraining on previous tasks (Yogatama et al., 2019). We can evaluate the brain-alignments' ability to combat catastrophic forgetting by training

on a sequence of tasks, and observing how the system performs on the previously trained-on tasks. We can then compare these performances to those of the original system, when trained in the same continual learning regime.

New training paradigms and architectures. The introduction of brain recordings during training of NLP systems also opens new questions about how best to incorporate the bias from brain recordings into the NLP system. The current best practices approach in NLP of fine-tuning all parameters of the network may not be the optimal approach when it comes to incorporating the signal from brain recordings, which is noisier and sparser than text data. For instance, focusing the effect of the brain recordings on certain parts of the neural network, rather than distributing it across all parameters, may be more beneficial.

Understanding language comprehension as a task. Three of the four brain imaging datasets in this dissertation were recorded while participants were asked to merely comprehend the stimulus (e.g. a chapter of a book or a movie) without performing any explicit task. An interesting future direction is to observe how the alignment of an NLP system with the brain recordings changes as participants are asked to perform specific linguistic tasks during the brain recording. Additionally, interventions in an NLP system that ablate or introduce new information pathways may reveal the sufficient or necessary computations that underlie various linguistic tasks, and one day perhaps even language comprehension.

Role of memory in language comprehension. The findings of Chapter 5 about the ability of the brain to keep some processing invariant from the task may be important for reusing previously learned concepts to perform new tasks, and may be one reason why the brain can learn so quickly, flexibly, and retain previously important information. Moving forward, we believe that understanding how to form and use task-invariant memories will lead to progress on several core challenges in bridging human and machine intelligence. Neurolinguistics would also benefit from an investigation into the role of different types of memory (e.g. semantic, episodic, and working memory) in language comprehension, as memory is essential to understanding the meaning of sequences of words. To make sense of the observed words, our brains must rely on memory to retrieve the previously learned meaning of each word and to maintain important context- and task-dependent information as new words are perceived.

Beyond language: integrating information from multiple modalities. The brain effortlessly integrates information from multiple sensory modalities with its internal representations (i.e. memory) to guide behavior. In contrast, machine learning models have trouble learning representations that can generalize to different tasks within the same modality, let alone across modalities. One long-term direction of the deeper investigation of memory in the brain is to understand the computational mechanism behind encoding generalizeable memory representations and the mechanisms that retrieve such memory representations according to the current modality context. An interesting short-term research direction is to take full advantage of multi-modal brain recording experiments, such as the Courtois NeuroMod dataset of people watching movies that we use in this dissertation,

and model all modalities involved (e.g. language, audition of non-language stimuli, and vision) instead of focusing on a single modality at a time.

Appendix A

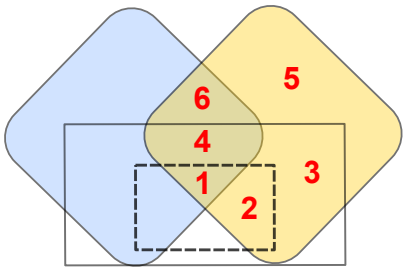
Supplementary Results for Chapter 3

Relationship of Special Cases A-C in Figure 3.2 to Most General Case

In Figure A.1, we present a Venn diagram that captures all possible relationships among two brain measurements, the presented stimulus, and the stimulus representation. All possible cases can be obtained by varying the amount of information that makes up each of the regions annotated with a red number, as well as the analogous regions in the blue brain source 1. In the main paper, we focus on three specific cases, presented in Figure 3.2 and replicated at the bottom of Figure A.1 for ease of visualization. These cases were selected because they are all possible candidates for the underlying relationships that lead to the real-world example of the inference problem described in Chapter 3.4, because an encoding model would perform equally well at predicting both sources in each of these 3 cases.

Metric Normalizations

The main metrics of interest defined in Chapter 3.4.2 are encoding model performance, source generalization, and source residuals. In the simple setting where all annotated regions in Figure A.1 are independent of each other, the encoding model performance is proportional to annotated regions **1 + 2**, source generalization to **1**, and source residuals to **2 + 3** (and to the analogous **2 + 3** regions on the blue brain source side). For some scientific questions, it may be more informative to normalize these metrics in different ways. For example, one may normalize the source generalization by the encoding model performance to compute the proportion of $\frac{\mathbf{1}}{\mathbf{1+2}}$ (i.e. the proportion of information shared between a brain source and the stimulus representation that is also shared by a second brain source). This metric is identical to the one proposed by Toneva, Mitchell, et al. (2020). Another type of normalization that we find informative in the current work is the intersubject correlation (ISC), which is proportional to **1 + 2 + 3 + 4** (i.e. the information shared between a brain source and the stimulus). This metric can be thought of as an estimate of the maximum possible performance (i.e. the noise ceiling). A similar metric was used as an estimate of the

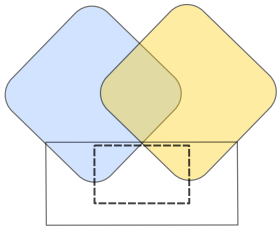


brain source 1
brain source 2
stimulus
stimulus representation

$$Y_{i,P} = \alpha \times \underbrace{g_{12,P}(X)}_{\textbf{1}\ \text{joint signal}} + (1-\alpha) \times \underbrace{g_{i,P}(X)}_{\textbf{2}\ \text{unique signal}} + \alpha \times \underbrace{N_{i,P}}_{\text{unique noise}} + (1-\alpha) \times \underbrace{N_{12,P}}_{\text{joint noise}} \quad (7)$$

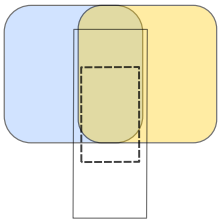
$$\text{where } N_{i,P} = \delta \times \underbrace{h_{i,P}(Z)}_{\textbf{3}} + (1-\delta) \times \underbrace{\epsilon_{i,P}}_{\textbf{5}} \text{ and } N_{12,P} = \delta \times \underbrace{h_{12,P}(Z)}_{\textbf{4}} + (1-\delta) \times \underbrace{\epsilon_{12,P}}_{\textbf{6}}.$$

Case A: unique information w.r.t. stimulus



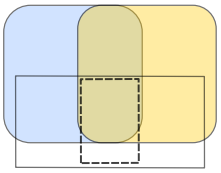
1 and **4** small

Case B: shared information w.r.t. stimulus



2 and **3** small

Case C: limited stimulus representation



2 and **4** small

Figure A.1: (Top) Most general Venn diagram that captures all possible underlying relationships between two brain measurements, the presented stimulus, and the stimulus representation. (Middle) Annotated data generation model in Eq. 3.8. (Bottom) Special cases considered in the main paper, that we argue cannot be disambiguated solely through encoding model performance.

noise ceiling by Wehbe, Blank, et al. (2020), though the authors did not make the connection to ISC explicitly. Note that the ISC across a dataset of more than two subjects is computed as the average of the pairwise ISC (i.e. the ISC for 1 of 6 subjects is the average across the ISC computed between that subject and the remaining 5 subjects). Following previous work (Hsu et al., 2004; Lescroart et al., 2019), we normalize all of our metrics by the square-root of the noise-ceiling, yielding normalized correlation values.

We hope that the conceptual breakdown of the different possible relations that we present in Figure 3.2 and Figure A.1 will help other researchers choose the most relevant normalization for their questions of interest.

Results for Human Connectome Project

HCP: short movies. We use publicly available data from the Human Connectome Project (HCP) 7T dataset, with healthy participants between 22-36 years old (Van Essen et al., 2013). HCP fMRI data comes minimally pre-processed as FIX-Denoised data (Glasser et al., 2013; Griffanti et al., 2014; Salimi-Khorshidi et al., 2014). We focus our analysis on only 90 participants for now, and are not using the remaining due to another project. Each participant watched naturalistic audio-visual video clips in English during 4 scans. Each scan was just over 15 minutes long, 60 minutes and 55 seconds of data were recorded. The fMRI sampling rate (TR) was 1 second.

In Figures A.2-A.7, we present the averaged and individual-level results for two participants in the HCP dataset, and show them against the averaged results for the Courtois NeuroMod dataset. We observe striking similarities across the two datasets, especially for the source generalization and source residuals.

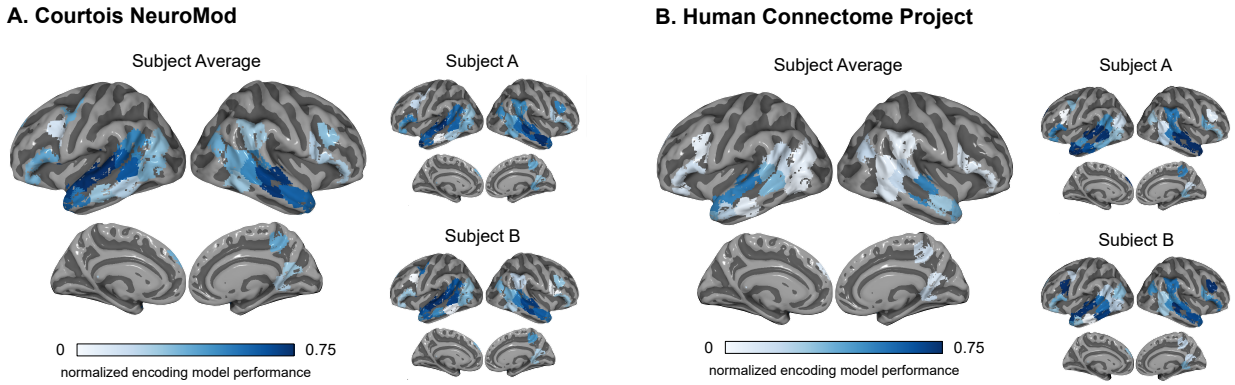


Figure A.2: Encoding performance at 33 significantly predicted ROIs (corrected at level 0.05).

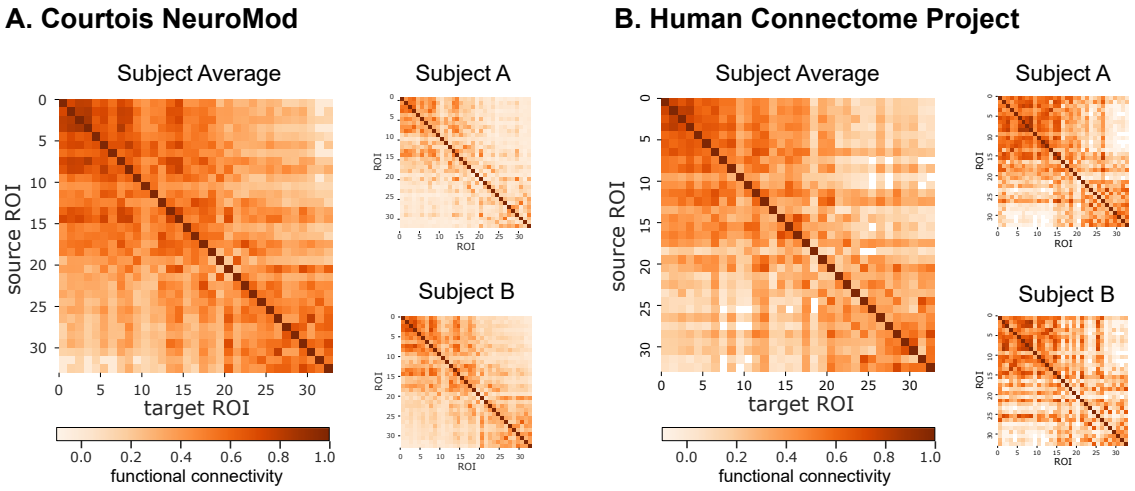


Figure A.3: Function connectivity of the 33 language ROIs with significant encoding model performance. Non-significant values are presented as 0's (t-test, corrected for multiple comparisons across regions at alpha level 0.05)

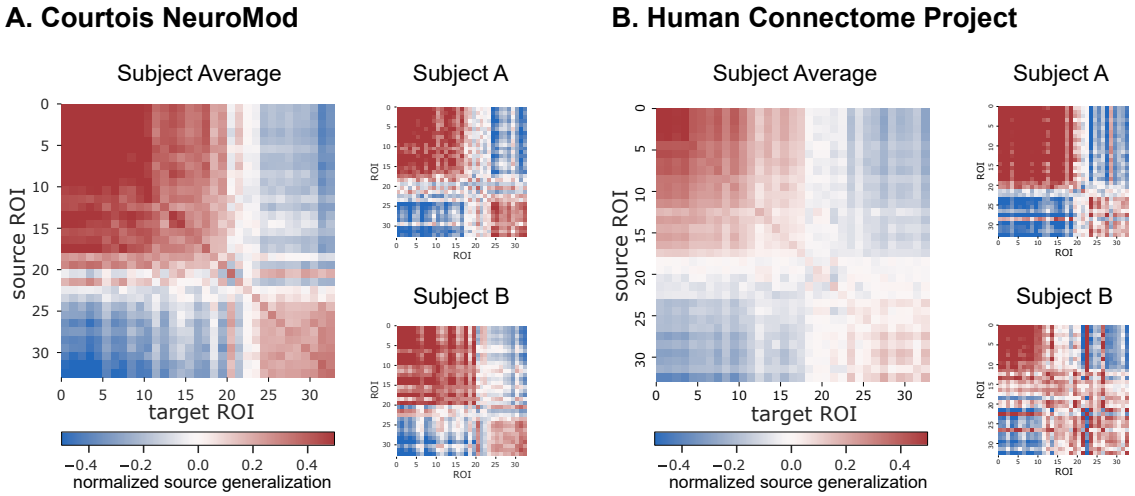


Figure A.4: Source Generalization. ROI pairs with high norm. source generalization (red) process information captured by the stimulus representations in a similar way. Pairs with high norm. source generalization are consistent at the group and individual level in both datasets.

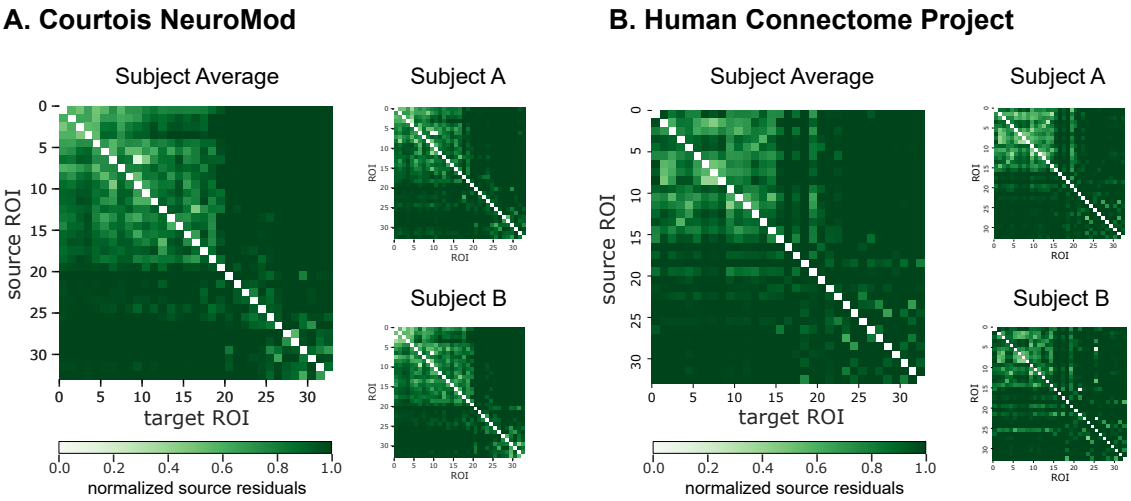


Figure A.5: Source Residuals. ROI pairs with high norm. source residuals (dark green) are processing unique information related to the stimulus representations. These ROI pairs with high norm. source residuals are consistent at the group and individual level in both datasets.

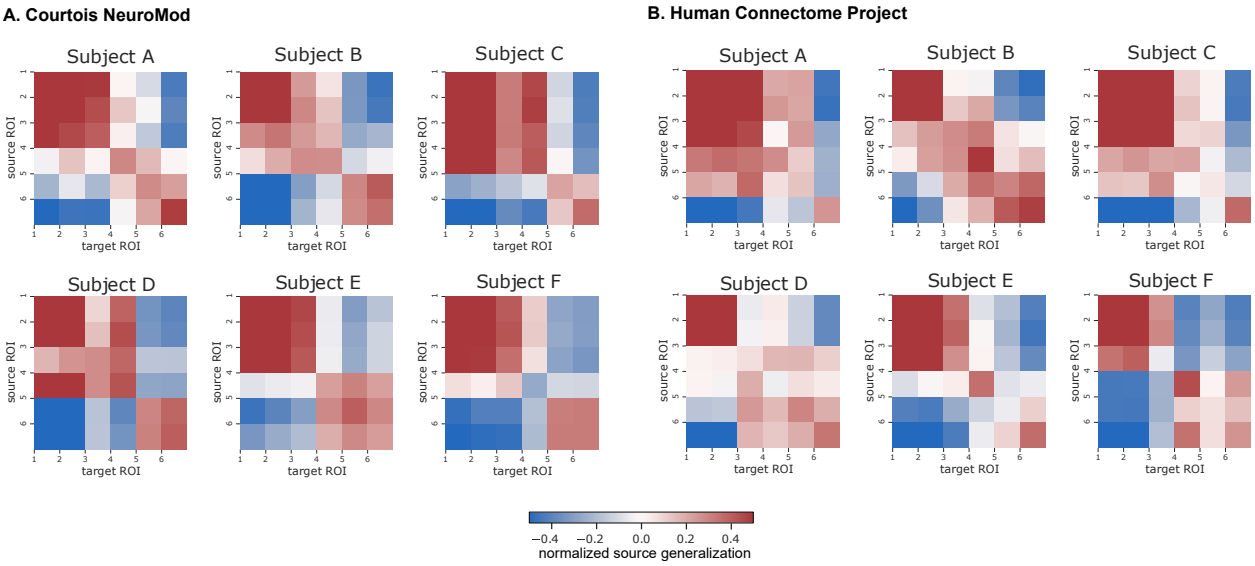


Figure A.6: Proposed Framework Example Individual Level Source Generalization. This figure shows the normalized source generalization for the six ROIs in the example using the proposed framework for participants A-F in both datasets. The ROI pairs with high normalized source generalization (red) are consistent across participants A-F in both datasets. They are also consistent with the group level presented in the main text.

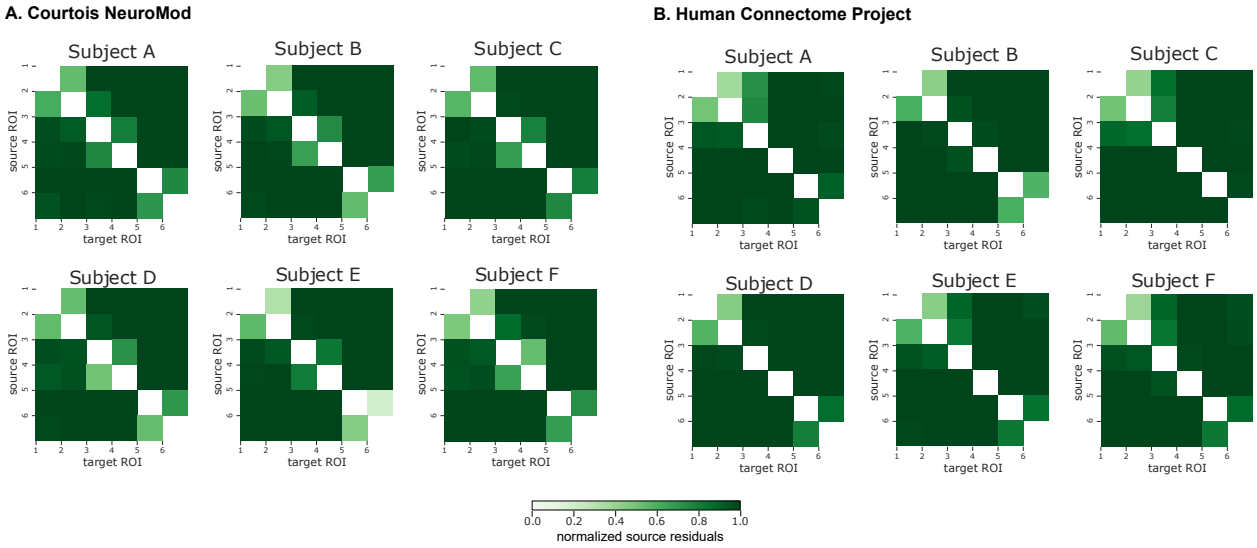


Figure A.7: Proposed Framework Example Individual Level Source Residuals. This figure shows the normalized source residuals for the six ROIs in the example using the proposed framework for participants A-F in both datasets. The ROI pairs with high normalized source residuals (green) are consistent across participants A-F in both datasets. They are also consistent with the group level presented in the main text.

Appendix B

Supplementary Results for Chapter 4

Experiments revealing shared information among NLP embeddings

For each of the 3 NLP embedding types (i.e. context(t-1) embedding, word(t-1) embedding, word(t) embedding), we train an encoding model taking as input each NLP embedding and predicting as output the word embedding for word(i), where $i \in [t-6, t+2]$. We evaluate the predictions of the encoding models using Pearson correlation, and obtain an average correlation over the four cross-validation folds.

Additional supra-word meaning results

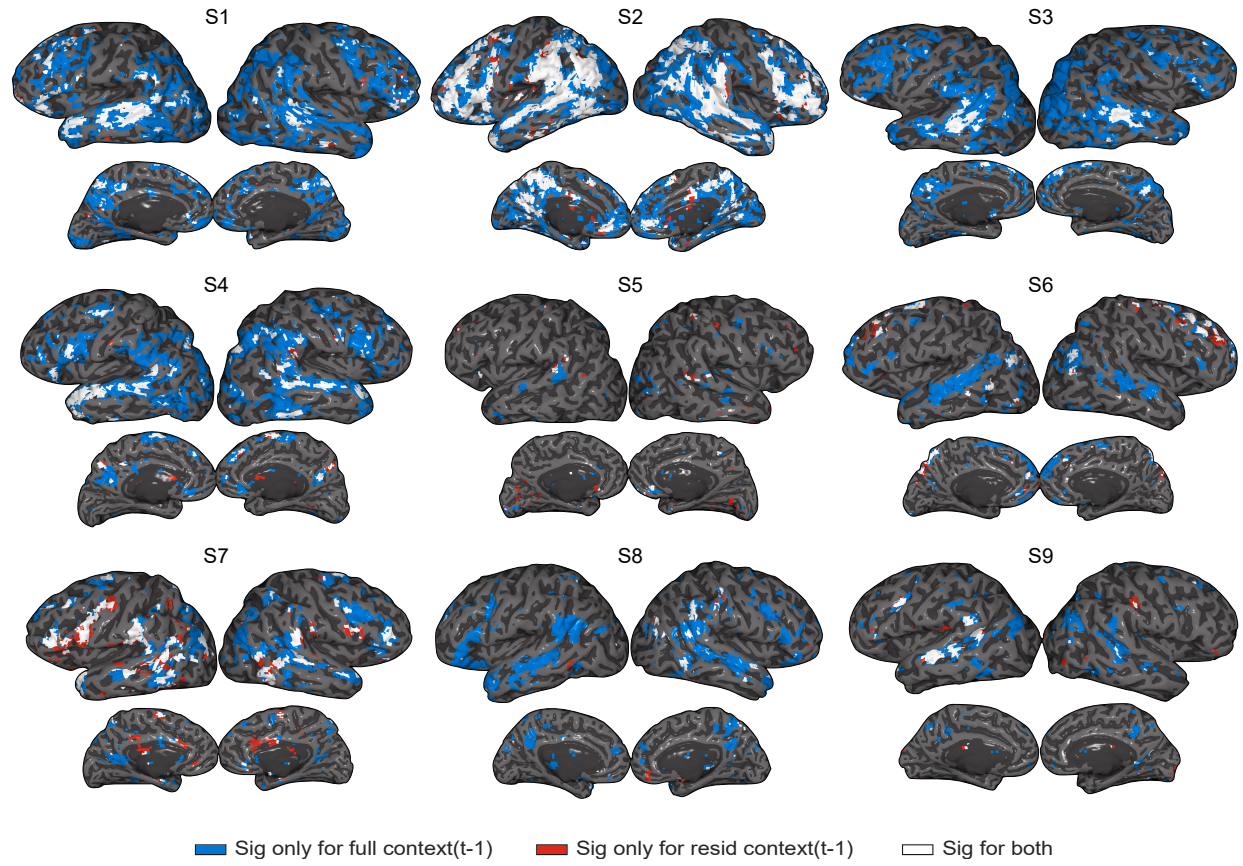


Figure B.1: Qualitative visualization of the voxels that are significantly predicted by the full contextualized representation (in blue), the residual contextualized representation (in red), or both (in white). A voxel is determined to be significantly predicted through a permutation test and FDR correction for multiple comparisons at the 0.01 level. Large parts of the language system, spanning the temporal cortex and the inferior frontal cortex, are significantly predicted by the full context embeddings. The voxels significantly predicted by the context residual are largely a subset of those predicted by the full context embeddings.

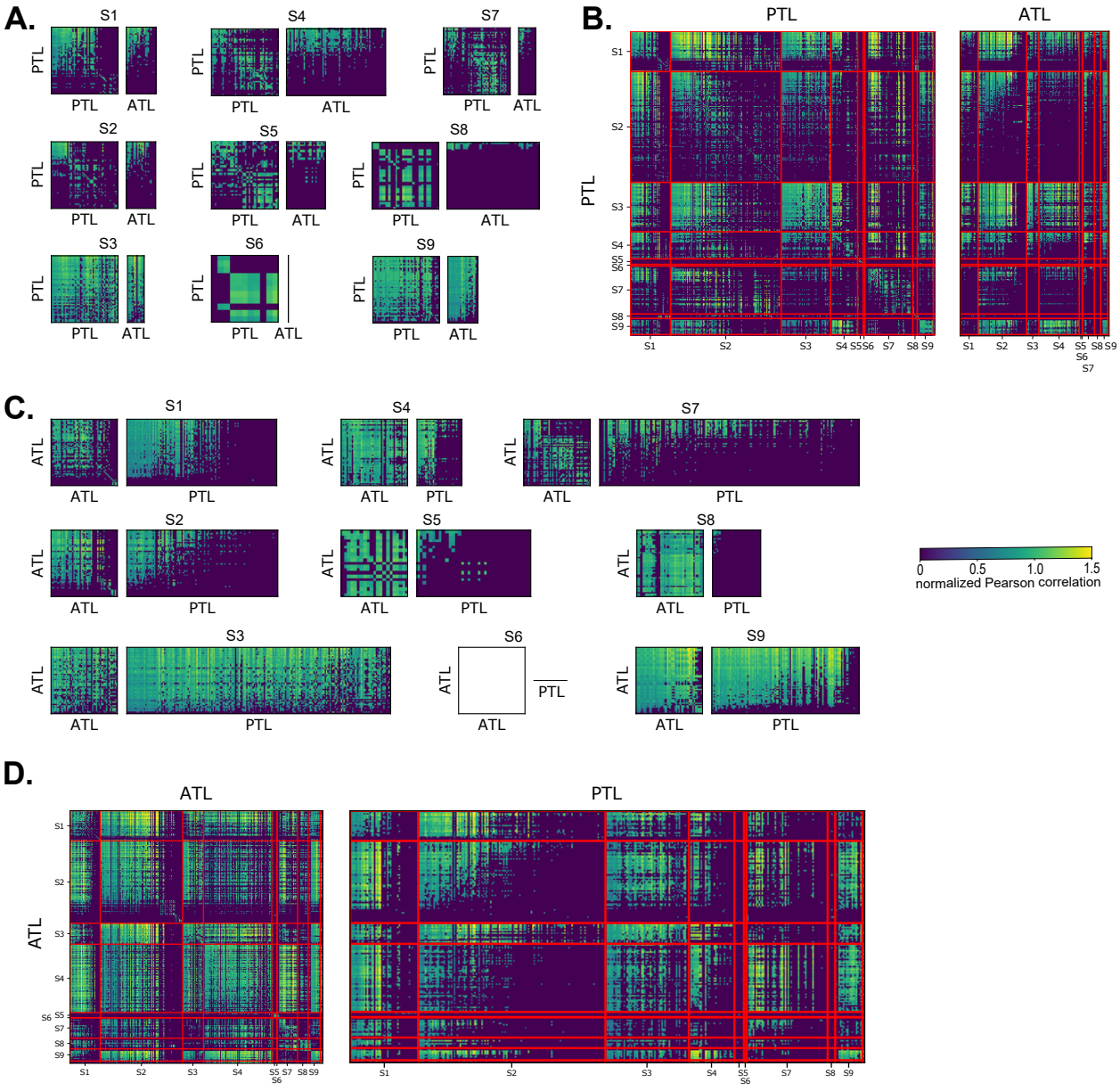

A.
S1
S4
S7
S2
S5
S8
S3
S6
S9
PTL
ATL
B.
PTL
ATL
S1
S2
S3
S4
S5
S6
S7
S8
S9
C.
S1
S4
S7
S2
S5
S8
S3
S6
S9
ATL
PTL
0
0.5
1.0
1.5
normalized Pearson correlation
D.
ATL
PTL

Figure B.2: Source Generalization Matrices for all 9 participants. Models trained to predict PTL voxels are used to predict PTL and ATL voxels (within-participant (**A**), and across-participants (**B**)). Models trained to predict ATL voxels are used to predict ATL and PTL voxels (within-participant (**C**), and across-participants (**D**)). Note that the block diagonal matrices of the across-participants correlations (in **B/D**) are equivalent to the plots in **A/C**. Only voxels that are significantly predicted by the context residual are included in this analysis. Note that the participant S6 does not have any significantly predicted voxels in the ATL. Correlations are normalized by dividing the performance of a model trained on voxel i at predicting the target voxel j by the performance of a model trained on the target voxel j. PTL cross-voxel correlations form two clusters: voxels in a cluster can predict each other but not the other cluster's voxels. Across participants, only one of these clusters has voxels that predict ATL voxels.

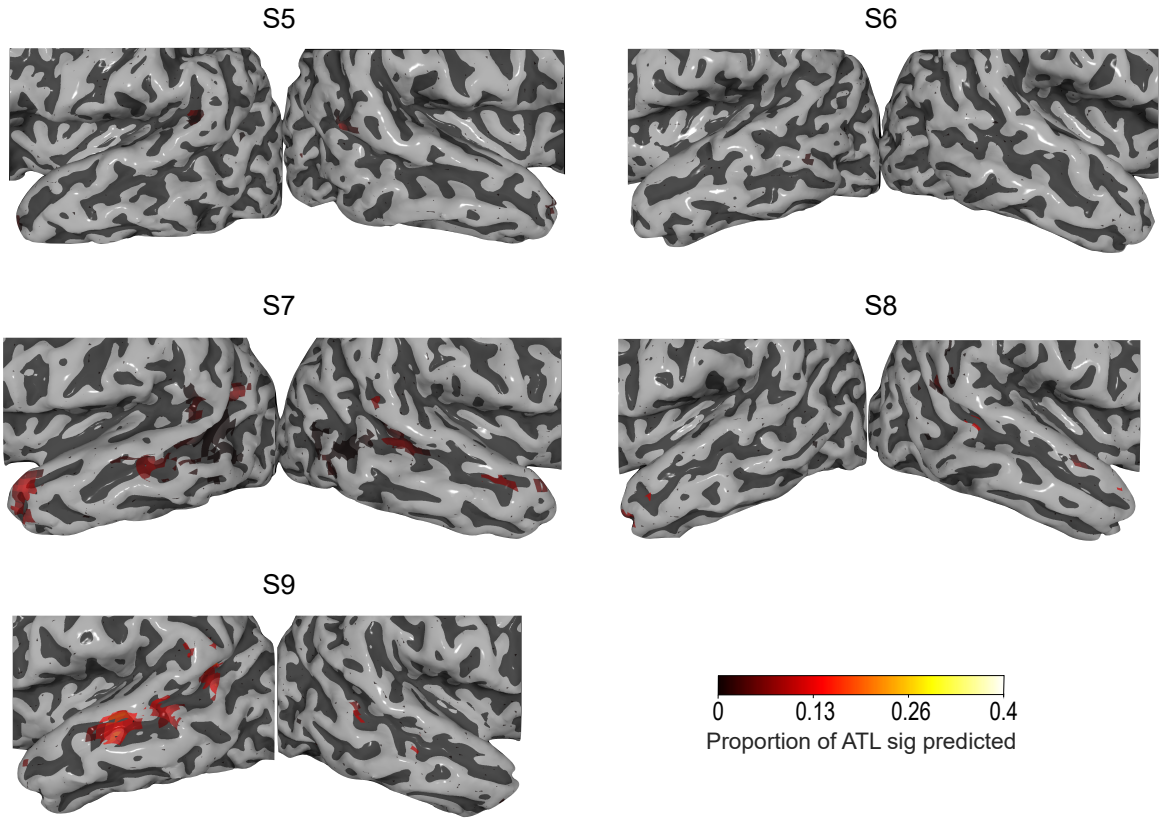


Figure B.3: Performance of encoding models trained on ATL and PTL voxels at predicting other participants' ATL for the remaining 5 participants. All participants who have more than a few significantly predicted voxels (6 out of 9 participants) show a cluster of predictive voxels in the pSTS.

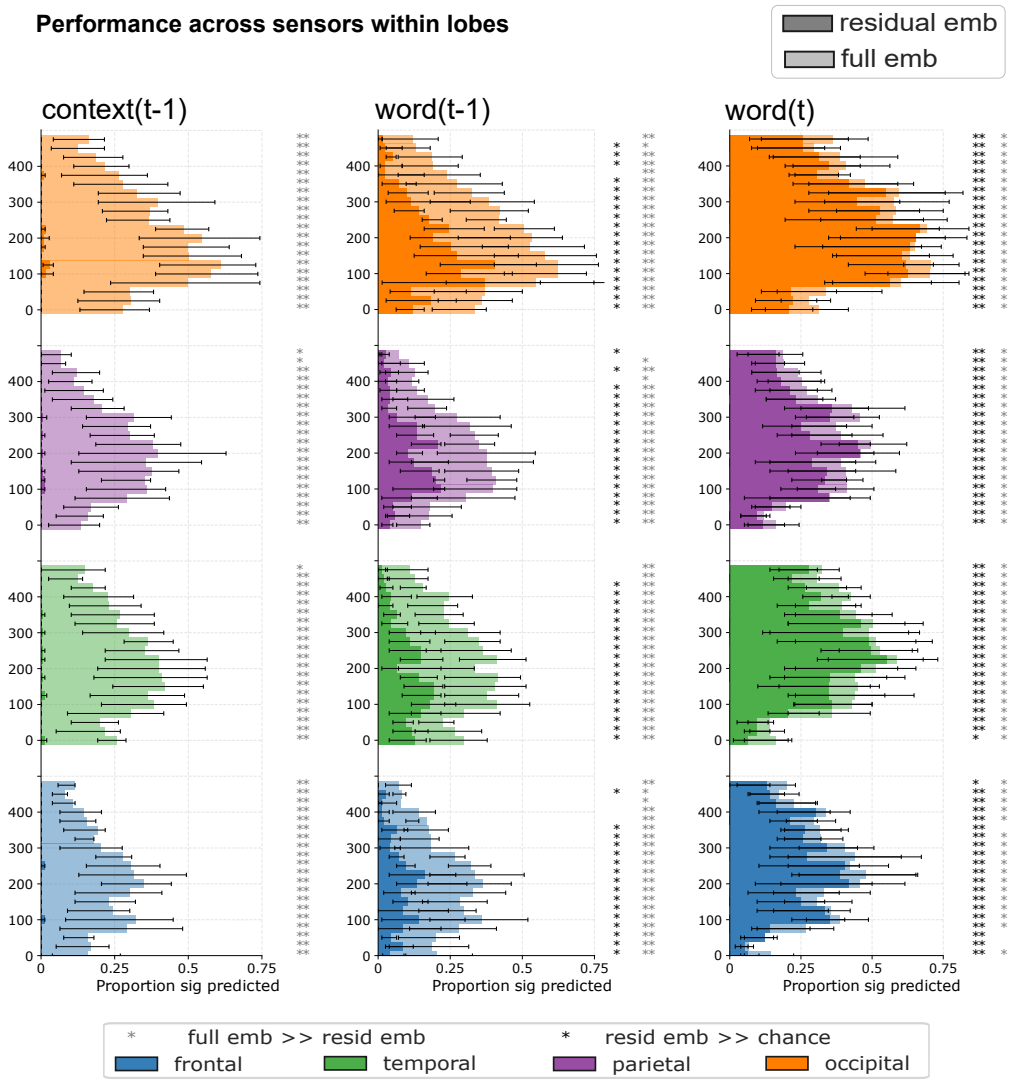


Figure B.4: Proportions of significantly predicted MEG sensors for each timepoint, divided by lobe. All subplots present the median across participants and errorbars signify the medians' 95% confidence intervals. Residual embeddings performance is compared with that of full embeddings (darker and lighter colors respectively, FDR corrected, $p < 0.05$). Removing the shared information among the full current word, the previous word and the context embeddings results in a significant decrease in performance for all embeddings and lobes. The decrease in performance for the context embedding (left column) is the most drastic, with no timewindows being significant for the residual context embedding across lobes.

Appendix C

Supplementary Results for Chapter 5

Relevant word features determined by the precomputed attention.

The precomputed attention is described in Section 5.3.4, and assigns a different relevance score (between 0 to 1) to each word feature for each question. Here we list the top 5 most relevant word features (i.e. with highest attention scores) for each of the 20 experimental question, as determined by the precomputed attention. The features are listed in decreasing order of importance (the first is the most important).

1. 'Can you hold it?':
    - CAN IT BE EASILY MOVED?
    - IS IT LIGHTWEIGHT?
    - WOULD YOU FIND IT IN A HOUSE?
    - CAN YOU TOUCH IT?
    - CAN YOU BUY IT?
2. 'Can you hold it in one hand?':
    - CAN IT BE EASILY MOVED?
    - IS IT LIGHTWEIGHT?
    - WOULD YOU FIND IT IN A HOUSE?
    - DO YOU HOLD IT TO USE IT?
    - CAN YOU BUY IT?
3. 'Can you pick it up?':
    - CAN IT BE EASILY MOVED?
    - IS IT LIGHTWEIGHT?
    - CAN YOU BUY IT?

- WOULD YOU FIND IT IN A HOUSE?
- DO YOU HOLD IT TO USE IT?

4. 'Is it bigger than a loaf of bread?':
    - IS IT HEAVY?
    - IS IT TALLER THAN A PERSON?
    - IS IT LONG?
    - DOES IT COME IN DIFFERENT SIZES?
    - IS IT USUALLY OUTSIDE?
5. 'Is it bigger than a microwave oven?':
    - IS IT TALLER THAN A PERSON?
    - IS IT HEAVY?
    - IS IT BIGGER THAN A BED?
    - IS IT LONG?
    - IS IT USUALLY OUTSIDE?
6. 'Is it bigger than a car?':
    - IS IT BIGGER THAN A BED?
    - IS IT BIGGER THAN A HOUSE?
    - IS IT TALLER THAN A PERSON?
    - IS IT HEAVY?
    - IS IT LONG?
7. 'Can it keep you dry?':
    - DOES IT PROVIDE SHADE?
    - IS IT A BUILDING?
    - DOES IT PROVIDE PROTECTION?
    - CAN YOU TOUCH IT?
    - ARE THERE MANY VARIETIES OF IT?
8. 'Could you fit inside it?':
    - IS IT BIGGER THAN A BED?
    - IS IT TALLER THAN A PERSON?
    - DOES IT PROVIDE SHADE?
    - IS IT A BUILDING?
    - IS IT BIGGER THAN A HOUSE?
9. 'Does it have at least one hole?':

- DOES IT HAVE A FRONT AND A BACK?
- IS IT SYMMETRICAL?
- DOES IT HAVE PARTS?
- DOES IT COME IN DIFFERENT SIZES?
- DOES IT HAVE INTERNAL STRUCTURE?

10. 'Is it hollow?':
    - IS IT A BUILDING?
    - DOES IT HAVE FLAT / STRAIGHT SIDES?
    - DOES IT OPEN?
    - DOES IT COME IN DIFFERENT SIZES?
    - CAN YOU TOUCH IT?
11. 'Is part of it made of glass?':
    - DOES IT HAVE WIRES OR A CORD?
    - DOES IT USE ELECTRICITY?
    - IS IT A BUILDING?
    - DOES IT HAVE FLAT / STRAIGHT SIDES?
    - DOES IT HAVE WRITING ON IT?
12. 'Is it made of metal?':
    - IS IT SILVER?
    - IS IT MECHANICAL?
    - WAS IT INVENTED?
    - IS IT SHINY?
    - DOES IT HAVE A HARD OUTER SHELL?
13. 'Is it manufactured?':
    - WAS IT INVENTED?
    - DOES IT HAVE WRITING ON IT?
    - DOES IT HAVE FLAT / STRAIGHT SIDES?
    - CAN YOU USE IT?
    - CAN YOU BUY IT?
14. 'Is it manmade?':
    - WAS IT INVENTED?
    - DOES IT HAVE FLAT / STRAIGHT SIDES?
    - DOES IT HAVE WRITING ON IT?

- CAN YOU USE IT?
- ARE THERE MANY VARIETIES OF IT?

15. 'Is it alive?':
    - IS IT CONSCIOUS?
    - IS IT AN ANIMAL?
    - IS IT WARM BLOODED?
    - DOES IT HAVE EARS?
    - IS IT WILD?
16. 'Was it ever alive?':
    - IS IT AN ANIMAL?
    - IS IT WILD?
    - CAN IT BITE OR STING?
    - IS IT CURVED?
    - IS IT CONSCIOUS?
17. 'Does it grow?':
    - IS IT WILD?
    - IS IT CONSCIOUS?
    - IS IT AN ANIMAL?
    - CAN IT BITE OR STING?
    - IS IT WARM BLOODED?
18. 'Does it have feelings?':
    - IS IT CONSCIOUS?
    - DOES IT HAVE EARS?
    - DOES IT HAVE A BACKBONE?
    - IS IT WARM BLOODED?
    - DOES IT HAVE A FACE?
19. 'Does it live in groups?':
    - IS IT AN ANIMAL?
    - CAN IT JUMP?
    - IS IT A HERBIVORE?
    - IS IT WILD?
    - IS IT CONSCIOUS?
20. 'Is it hard to catch?:

- IS IT FAST?
- IS IT A PREDATOR?
- IS IT AN ANIMAL?
- CAN IT JUMP?
- IS IT USUALLY OUTSIDE?

Sensor-timepoint results for H4.1 for 25ms windows.

We present the sensor-timepoint results for H3 for 25ms time-windows in Figure C.1. They follow the general trend of the results from 50ms time-windows presented in Figure 5.7.

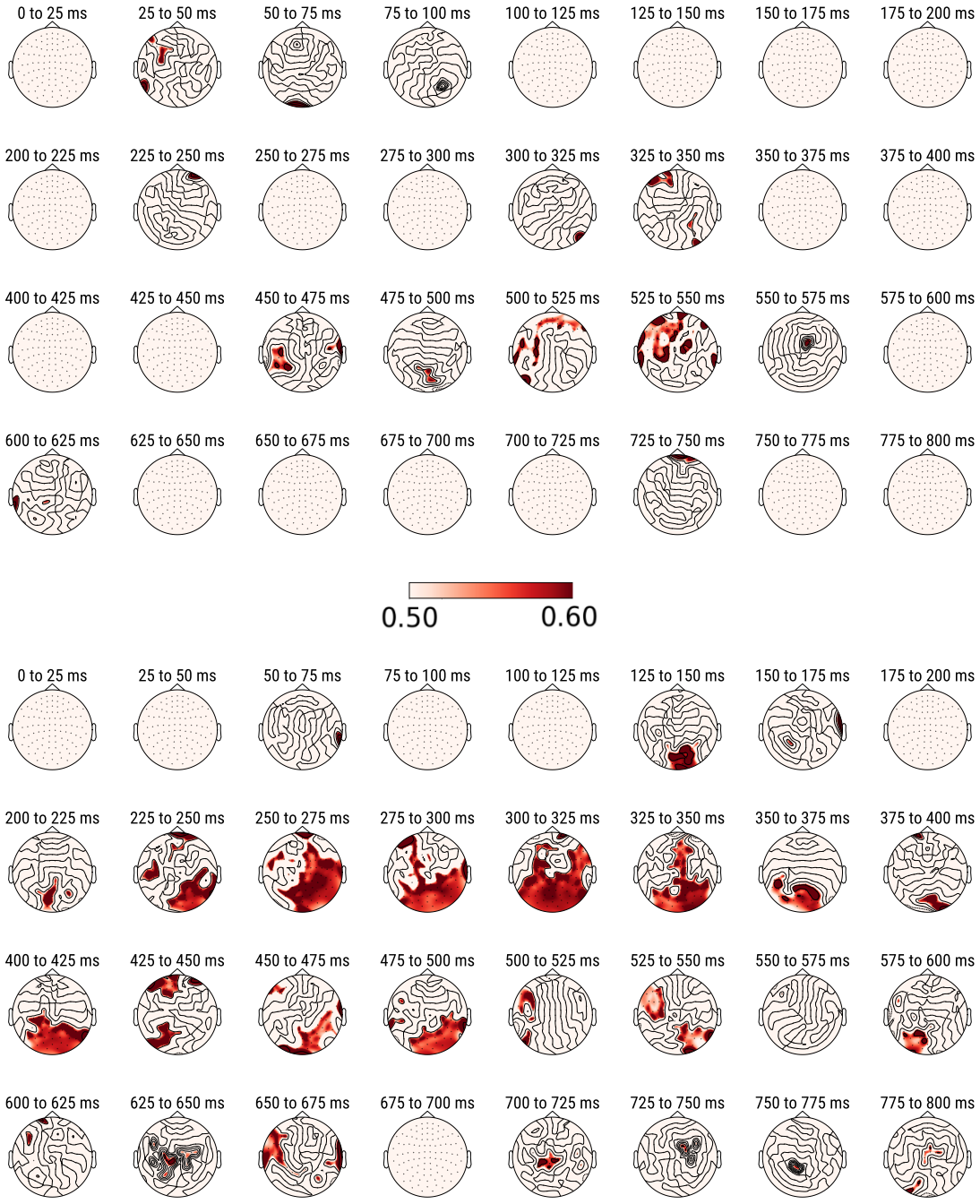


Figure C.1: Mean 2v2 accuracy across subjects of predicting sensor-timepoints in 25ms time-windows using H4.1, when predicting the brain recordings for two stimulus-question pairs that share the same word (Top) and the same question (Bottom). The displayed accuracies are significantly greater than chance.

Stimuli and questions for 20question dataset

Questions in experiment:

1. Can you hold it?
2. Can you hold it in one hand?
3. Can you pick it up?
4. Is it bigger than a loaf of bread?
5. Is it bigger than a microwave oven?
6. Is it bigger than a car?
7. Can it keep you dry?
8. Could you fit inside it?
9. Does it have at least one hole?
10. Is it hollow?
11. Is part of it made of glass?
12. Is it made of metal?
13. Is it manufactured?
14. Is it manmade?
15. Is it alive?
16. Was it ever alive?
17. Does it grow?
18. Does it have feelings?
19. Does it live in groups?
20. Is it hard to catch?

Stimuli in experiment: dog, horse, arm, eye, foot, hand, leg, apartment, barn, church, house, igloo, arch, chimney, closet, door, window, coat, dress, pants, shirt, skirt, bed, chair, desk, dresser, table, ant, bee, beetle, butterfly, fly, bottle, cup, glass, knife, spoon, bell, key, refrigerator, telephone, watch, chisel, hammer, pliers, saw, screwdriver, carrot, celery, corn, lettuce, tomato, airplane, bicycle, car, train, truck

All MTurk questions:

- Is it an animal?
- Is it a body part?
- Is it a building?
- Is it a building part?
- Is it clothing?
- Is it furniture?
- Is it an insect?
- Is it a kitchen item?
- Is it manmade?
- Is it a tool?
- Can you eat it?
- Is it a vehicle?
- Is it a person?
- Is it a vegetable / plant?
- Is it a fruit?
- Is it made of metal?
- Is it made of plastic?
- Is part of it made of glass?
- Is it made of wood?
- Is it shiny?
- Can you see through it?
- Is it colorful?
- Does it change color?
- Is one more than one colored?
- Is it always the same color(s)?
- Is it white?
- Is it red?
- Is it orange?
- Is it flesh-colored?
- Is it yellow?
- Is it green?
- Is it blue?
- Is it silver?
- Is it brown?
- Is it black?
- Is it curved?
- Is it straight?
- Is it flat?
- Does it have a front and a back?
- Does it have a flat / straight top?
- Does it have flat / straight sides?
- Is taller than it is wide/long?
- Is it long?
- Is it pointed / sharp?
- Is it tapered?
- Is it round?
- Does it have corners?
- Is it symmetrical?

- Is it hairy?
- Is it fuzzy?
- Is it clear?
- Is it smooth?
- Is it soft?
- Is it heavy?
- Is it lightweight?
- Is it dense?
- Is it slippery?
- Can it change shape?
- Can it bend?
- Can it stretch?
- Can it break?
- Is it fragile?
- Does it have parts?
- Does it have moving parts?
- Does it come in pairs?
- Does it come in a bunch/pack?
- Does it live in groups?
- Is it part of something larger?
- Does it contain something else?
- Does it have internal structure?
- Does it open?
- Is it hollow?
- Does it have a hard inside?
- Does it have a hard outer shell?
- Does it have at least one hole?
- Is it alive?
- Was it ever alive?
- Is it a specific gender?
- Is it manufactured?
- Was it invented?
- Was it around 100 years ago?
- Are there many varieties of it?
- Does it come in different sizes?
- Does it grow?
- Is it smaller than a golfball?
- Is it bigger than a loaf of bread?
- Is it bigger than a microwave oven?
- Is it bigger than a bed?
- Is it bigger than a car?
- Is it bigger than a house?
- Is it taller than a person?
- Does it have a tail?
- Does it have legs?
- Does it have four legs?
- Does it have feet?
- Does it have paws?
- Does it have claws?
- Does it have horns / thorns / spikes?
- Does it have hooves?
- Does it have a face?
- Does it have a backbone?
- Does it have wings?
- Does it have ears?
- Does it have roots?
- Does it have seeds?
- Does it have leaves?
- Does it come from a plant?
- Does it have feathers?
- Does it have some sort of nose?
- Does it have a hard nose/beak?
- Does it contain liquid?
- Does it have wires or a cord?
- Does it have writing on it?
- Does it have wheels?
- Does it make a sound?
- Does it make a nice sound?
- Does it make sound continuously when active?
- Is its job to make sounds?
- Does it roll?
- Can it run?
- Is it fast?
- Can it fly?
- Can it jump?
- Can it float?
- Can it swim?
- Can it dig?
- Can it climb trees?
- Can it cause you pain?
- Can it bite or sting?
- Does it stand on two legs?
- Is it wild?
- Is it a herbivore?
- Is it a predator?
- Is it warm blooded?
- Is it a mammal?
- Is it nocturnal?
- Does it lay eggs?
- Is it conscious?
- Does it have feelings?
- Is it smart?
- Is it mechanical?
- Is it electronic?
- Does it use electricity?
- Can it keep you dry?
- Does it provide protection?
- Does it provide shade?

- Does it cast a shadow?
- Do you see it daily?
- Is it helpful?
- Do you interact with it?
- Can you touch it?
- Would you avoid touching it?
- Can you hold it?
- Can you hold it in one hand?
- Do you hold it to use it?
- Can you play it?
- Can you play with it?
- Can you pet it?
- Can you use it?
- Do you use it daily?
- Can you use it up?
- Do you use it when cooking?
- Is it used to carry things?
- Can you pick it up?
- Can you control it?
- Can you sit on it?
- Can you ride on/in it?
- Is it used for transportation?
- Could you fit inside it?
- Is it used in sports?
- Do you wear it?
- Can it be washed?
- Is it cold?
- Is it cool?
- Is it warm?
- Is it hot?
- Is it unhealthy?
- Is it hard to catch?
- Can you peel it?
- Can you walk on it?
- Can you switch it on and off?
- Can it be easily moved?
- Do you drink from it?
- Does it go in your mouth?
- Is it tasty?
- Is it used during meals?
- Does it have a strong smell?
- Does it smell good?
- Does it smell bad?
- Is it usually inside?
- Is it usually outside?
- Would you find it on a farm?
- Would you find it in a school?
- Would you find it in a zoo?
- Would you find it in an office?
- Would you find it in a restaurant?
- Would you find in the bathroom?
- Would you find it in a house?
- Would you find it near a road?
- Would you find it in a dump/landfill?
- Would you find it in the forest?
- Would you find it in a garden?
- Would you find it in the sky?
- Do you find it in space?
- Does it live above ground?
- Does it get wet?
- Does it live in water?
- Can it live out of water?
- Do you take care of it?
- Does it make you happy?
- Do you love it?
- Would you miss it if it were gone?
- Is it scary?
- Is it dangerous?
- Is it friendly?
- Is it rare?
- Can you buy it?
- Is it valuable?

Appendix D

Supplementary Results for Chapter 6

Complete attention results for uniform BERT

condition	uni L3	uni L4	uni L5	base	count
simple	0.96	1.00	0.99	1.00	120
in a sentential complement	0.83	0.83	0.84	0.83	1440
short VP coordination	0.91	0.88	0.88	0.89	720
long VP coordination	0.95	0.95	0.96	0.98	400
across a prepositional phrase	0.88	0.86	0.80	0.85	19440
across a subject relative clause	0.84	0.84	0.83	0.84	9600
across an object relative clause	0.90	0.86	0.83	0.89	19680
across an object relative clause (no that)	0.75	0.72	0.75	0.86	19680
in an object relative clause	0.96	0.92	0.91	0.95	15960
in an object relative clause (no that)	0.70	0.69	0.74	0.79	15960
reflexive anaphora: simple	0.99	0.98	1.00	0.94	280
reflexive anaphora: in a sent. complem.	0.88	0.87	0.86	0.89	3360
reflexive anaphora: across a rel. clause	0.82	0.68	0.66	0.80	22400

Table D.1: Performance of models with uniformly-altered attention in layers 3-5 in BERT on a range of syntactic tasks by Marvin et al. (2018). 'Base' refers to pretrained BERT.

Bibliography

[1] Mostafa Abdou, Ana Valeria Gonzalez, Mariya Toneva, Daniel Hershcovich, and Anders Søgaard. "Does injecting linguistic structure into language models lead to better alignment with brain recordings?" In: *arXiv preprint arXiv:2101.12608* (2021).

[2] Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. "Blackbox meets blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains". In: *arXiv preprint arXiv:1906.01539* (2019).

[3] Mark Allen, William Badecker, and Lee Osterhout. "Morphological analysis in sentence processing: An ERP study". In: *Language and Cognitive Processes* 18.4 (2003), pp. 405–430.

[4] Elissa M Aminoff, Mariya Toneva, Abhinav Shrivastava, Xinlei Chen, Ishan Misra, Abhinav Gupta, and Michael J Tarr. "Applying artificial vision models to human scene understanding". In: *Frontiers in computational neuroscience* 9 (2015), p. 8.

[5] Ery Arias-Castro, Shiyun Chen, et al. "Distribution-free multiple testing". In: *Electronic Journal of Statistics* 11.1 (2017), pp. 1983–2001.

[6] Rina Foygel Barber, Emmanuel J Candès, et al. "Controlling the false discovery rate via knockoffs". In: *The Annals of Statistics* 43.5 (2015), pp. 2055–2085.

[7] Marco Baroni. "Linguistic generalization and compositionality in modern artificial neural networks". In: *Philosophical Transactions of the Royal Society B* 375.1791 (2020), p. 20190307.

[8] Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. "Entailment above the word level in distributional semantics". In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012, pp. 23–32.

[9] Marco Baroni, Raffaella Bernardi, Roberto Zamparelli, et al. "Frege in space: A program for compositional distributional semantics". In: *Linguistic Issues in language technology* 9.6 (2014), pp. 5–110.

[10] Marco Baroni and Roberto Zamparelli. "Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space". In: *Proceedings of the 2010 conference on empirical methods in natural language processing*. 2010, pp. 1183–1193.

[11] Lawrence W Barsalou. "Grounded cognition". In: *Annu. Rev. Psychol.* 59 (2008), pp. 617–645.

[12] LW Barsalou. “Perceptual symbol systems. Behavioral and Brain Sciences22: 577609.[BRC](2003) Situated simulation in the human conceptual system”. In: *Language and* (1999).

[13] Lisa Beinborn, Samira Abnar, and Rochelle Choenni. “Robust Evaluation of Language-Brain Encoding Experiments”. In: *arXiv preprint arXiv:1904.02547* (2019).

[14] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[15] Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. “Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies”. In: *Cerebral cortex* 19.12 (2009), pp. 2767–2796.

[16] Esti Blanco-Elorrieta and Liina Pylkkänen. “Bilingual language switching in the laboratory versus in the wild: The spatiotemporal dynamics of adaptive language control”. In: *Journal of Neuroscience* 37.37 (2017), pp. 9022–9036.

[17] Idan Blank and Evelina Fedorenko. “Domain-general brain regions do not track linguistic input as closely as language-selective regions”. In: *Journal of Neuroscience* (2017), pp. 3642–16.

[18] J. Brennan, Y. Nir, U. Hasson, R. Malach, D.J. Heeger, and L. Pylkkänen. “Syntactic structure building in the anterior temporal lobe during natural story listening”. In: *Brain and language* (2010).

[19] Peter Bright, HE Moss, Emmanuel A Stamatakis, and Lorraine K Tyler. “Longitudinal studies of semantic dementia: the relationship between structural and functional changes over time”. In: *Neuropsychologia* 46.8 (2008), pp. 2177–2188.

[20] Gijs Joost Brouwer and David J Heeger. “Categorical clustering of the neural representation of color”. In: *Journal of Neuroscience* 33.39 (2013), pp. 15454–15465.

[21] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. “Decomposing lexical and compositional syntax and semantics with deep language models”. In: *arXiv preprint arXiv:2103.01620* (2021).

[22] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. “GPT-2’s activations predict the degree of semantic comprehension in the human brain”. In: *bioRxiv* (2021).

[23] Charlotte Caucheteux and Jean-Rémi King. “Language processing in brains and deep neural networks: computational convergence and its limits”. In: *BioRxiv* (2020).

[24] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. “Universal sentence encoder”. In: *arXiv preprint arXiv:1803.11175* (2018).

[25] Linda L Chao, James V Haxby, and Alex Martin. “Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects”. In: *Nature neuroscience* 2.10 (1999), pp. 913–919.

[26] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. “One billion word benchmark for measuring progress in statistical language modeling”. In: *arXiv preprint arXiv:1312.3005* (2013).

[27] Yining Chen, Sorcha Gilroy, Kevin Knight, and Jonathan May. "Recurrent neural networks as weighted language recognizers". In: *arXiv preprint arXiv:1711.05408* (2017).

[28] Noam Chomsky. *Syntactic structures*. Walter de Gruyter, 2002.

[29] Stephen Clark. *Type-Driven Syntax and Semantics for Composing Meaning Vectors*. 2013.

[30] Jonathan D Cohen, Kevin Dunbar, and James L McClelland. "On the control of automatic processes: a parallel distributed processing account of the Stroop effect." In: *Psychological review* 97.3 (1990), p. 332.

[31] Alexis Conneau, German Kruszewski, Guillaume Lample, Loic Barrault, and Marco Baroni. "What you can cram into a single vector: Probing sentence embeddings for linguistic properties". In: *arXiv preprint arXiv:1805.01070* (2018).

[32] Seana Coulson, Jonathan W King, and Marta Kutas. "Expect the unexpected: Event-related brain response to morphosyntactic violations". In: *Language and cognitive processes* 13.1 (1998), pp. 21–58.

[33] Logan Cross, Jeff Cockburn, Yisong Yue, and John P O'Doherty. "Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments". In: *Neuron* 109.4 (2021), pp. 724–738.

[34] Tolga Cukur, Shinji Nishimoto, Alexander G Huth, and Jack L Gallant. "Attention during natural vision warps semantic representation across the human brain". In: *Nature Neuroscience* 16.6 (2013), p. 763.

[35] Jeff Da and Jungo Kasai. "Understanding Commonsense Inference Aptitude of Deep Contextual Representations". In: *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*. 2019, pp. 1–12.

[36] Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context". In: *arXiv preprint arXiv:1901.02860* (2019).

[37] Hanna Damasio, Thomas J Grabowski, Daniel Tranel, Richard D Hichwa, and Antonio R Damasio. "A neural basis for lexical retrieval". In: *Nature* 380.6574 (1996), pp. 499–505.

[38] Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. "The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality". In: *Journal of Neuroscience* 39.39 (2019), pp. 7722–7736.

[39] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[40] Manoj K. Doss, Darrick G. May, Matthew W. Johnson, John M. Clifton, Sidnee L. Hedrick, Thomas E. Prisinzano, Roland R. Griffiths, and Frederick S. Barrett. "The Acute Effects of the Atypical Dissociative Hallucinogen Salvinorin A on Functional Connectivity in the Human Brain". In: *Scientific Reports* 10.1 (Dec. 2020), p. 16392. ISSN: 20452322. DOI: 10.1038/s41598-020-73216-8. URL: https://doi.org/10.1038/s41598-020-73216-8.

[41] John Duncan. "Attention, intelligence, and the frontal lobes." In: (1995).

[42] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[43] Oscar Esteban, Ross Blair, Christopher J. Markiewicz, Shoshana L. Berleant, Craig Moodie, Feilong Ma, Ayse Ilkay Isik, Asier Erramuzpe, Mathias Kent James D. andGoncalves, Elizabeth DuPre, Kevin R. Sitek, Daniel E. P. Gomez, Daniel J. Lurie, Zhifang Ye, Russell A. Poldrack, and Krzysztof J. Gorgolewski. "fMRIPrep". In: *Software* (2018). DOI: 10.5281/zenodo.852659.

[44] Oscar Esteban, Christopher Markiewicz, Ross W Blair, Craig Moodie, Ayse Ilkay Isik, Asier Erramuzpe Aliaga, James Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit Ghosh, Jessey Wright, Joke Durnez, Russell Poldrack, and Krzysztof Jacek Gorgolewski. "fMRIPrep: a robust preprocessing pipeline for functional MRI". In: *Nature Methods* (2018). DOI: 10.1038/s41592-018-0235-4.

[45] E. Fedorenko, P.-J. Hsieh, A. Nieto-Castanon, S. Whitfield-Gabrieli, and N. Kanwisher. "New method for fMRI investigations of language: Defining ROIs functionally in individual subjects". In: *Journal of Neurophysiology* 104.2 (2010), pp. 1177–1194.

[46] Evelina Fedorenko, Terri L Scott, Peter Brunner, William G Coon, Brianna Pritchett, Gerwin Schalk, and Nancy Kanwisher. "Neural correlate of the construction of sentence meaning". In: *Proceedings of the National Academy of Sciences* 113.41 (2016), E6256–E6262.

[47] Evelina Fedorenko and Sharon L Thompson-Schill. "Reworking the language network". In: *Trends in cognitive sciences* 18.3 (2014), pp. 120–126.

[48] Emily S. Finn, Xilin Shen, Dustin Scheinost, Monica D. Rosenberg, Jessica Huang, Marvin M. Chun, Xenophon Papademetris, and R. Todd Constable. "Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity". In: *Nature Neuroscience* 18.11 (Nov. 2015), pp. 1664–1671. ISSN: 15461726. DOI: 10.1038/nn.4135. URL: https://www.nature.com/articles/nn.4135.

[49] Bruce Fischl. "FreeSurfer". In: *Neuroimage* 62.2 (2012), pp. 774–781.

[50] Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. "The ERP response to the amount of information conveyed by words in sentences". In: *Brain and language* 140 (2015), pp. 1–11.

[51] Steven M Frankland and Joshua D Greene. "An architecture for encoding sentence meaning in left mid-superior temporal cortex". In: *Proceedings of the National Academy of Sciences* 112.37 (2015), pp. 11732–11737.

[52] Angela D Friederici. "The brain basis of language processing: from structure to function". In: *Physiological reviews* 91.4 (2011), pp. 1357–1392.

[53] Kunihiko Fukushima and Sei Miyake. "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition". In: *Competition and cooperation in neural nets*. Springer, 1982, pp. 267–285.

[54] Alona Fyshe. "Studying language in context using the temporal generalization method". In: *Philosophical Transactions of the Royal Society B* 375.1791 (2020), p. 20180531.

[55] Alona Fyshe, Brian Murphy, Partha Talukdar, and Tom Mitchell. "Documents and dependencies: an exploration of vector space models for semantic composition". In: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. 2013, pp. 84–93.

[56] Alona Fyshe, Gustavo Sudre, Leila Wehbe, Nicole Rafidi, and Tom M Mitchell. "The lexical semantics of adjective–noun phrases in the human brain". In: *Human brain mapping* 40.15 (2019), pp. 4457–4469.

[57] Alona Fyshe, Partha P Talukdar, Brian Murphy, and Tom M Mitchell. "Interpretable semantic vectors from a joint model of brain-and text-based meaning". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Vol. 1. 2014, pp. 489–499.

[58] Gabriela Gan, Christian Büchel, and Frédéric Isel. "Effect of language task demands on the neural response during lexical access: a functional magnetic resonance imaging study". In: *Brain and behavior* 3.4 (2013), pp. 402–416.

[59] James S Gao, Alexander G Huth, Mark D Lescroart, and Jack L Gallant. "Pycortex: an interactive surface visualizer for fMRI". In: *Frontiers in neuroinformatics* 9 (2015), p. 23.

[60] Siyuan Gao, Abigail S. Greene, R. Todd Constable, and Dustin Scheinost. "Combining multiple connectomes improves predictive modeling of phenotypic measures". In: *NeuroImage* 201 (Nov. 2019), p. 116038. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2019.116038.

[61] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. "AllenNLP: A Deep Semantic Natural Language Processing Platform". In: *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. 2018, pp. 1–6.

[62] Jon Gauthier and Roger Levy. "Linking artificial and human neural representations of language". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 529–539.

[63] Matthew F. Glasser, Stamatios N. Sotiropoulos, J. Anthony Wilson, Timothy S. Coalson, Bruce Fischl, Jesper L. Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R. Polimeni, David C. Van Essen, and Mark Jenkinson. "The minimal preprocessing pipelines for the Human Connectome Project". In: *NeuroImage* 80 (Oct. 2013), pp. 105–124. ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.2013.04.127. URL: https://www.sciencedirect.com/science/article/pii/S1053811913005053?via%3Dihub.

[64] RF Goldberg, CA Perfetti, and W Schneider. "Distinct and common cortical activations for multimodal semantic categories". In: *Cognitive, Affective, & Behavioral Neuroscience* 6.3 (2006), pp. 214–222.

[65] Robert F Goldberg, Charles A Perfetti, and Walter Schneider. "Perceptual knowledge retrieval activates sensory brain regions". In: *The Journal of Neuroscience* 26.18 (2006), pp. 4917–4921.

[66] Yoav Goldberg. "Assessing BERT's Syntactic Abilities". In: *arXiv preprint arXiv:1901.05287* (2019).

[67] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. "Thinking ahead:

prediction in context as a keystone of language in humans and machines". In: *bioRxiv* (2021), pp. 2020–12.

[68] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. Vol. 1. 2. MIT press Cambridge, 2016.

[69] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. "MEG and EEG data analysis with MNE-Python". In: *Frontiers in neuroscience* 7 (2013), p. 267.

[70] Alex Graves. "Supervised sequence labelling". In: *Supervised sequence labelling with recurrent neural networks*. Springer, 2012, pp. 5–13.

[71] Abigail S. Greene, Siyuan Gao, Dustin Scheinost, and R. Todd Constable. "Task-induced brain state manipulation improves prediction of individual traits". In: *Nature Communications* 9.1 (Dec. 2018), pp. 1–13. ISSN: 20411723. DOI: 10.1038/s41467-018-04920-3. URL: www.nature.com/naturecommunications.

[72] Edward Grefenstette and Mehrnoosh Sadrzadeh. "Experimental support for a categorical compositional distributional model of meaning". In: *arXiv preprint arXiv:1106.4058* (2011).

[73] Ludovica Griffanti, Gholamreza Salimi-Khorshidi, Christian F. Beckmann, Edward J. Auerbach, Gwenaëlle Douaud, Claire E. Sexton, Eniko Zsoldos, Klaus P. Ebmeier, Nicola Filippini, Clare E. Mackay, Steen Moeller, Junqian Xu, Essa Yacoub, Giuseppe Baselli, Kamil Ugurbil, Karla L. Miller, and Stephen M. Smith. "ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging". In: *NeuroImage* 95 (July 2014), pp. 232–247. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2014.03.034. URL: /pmc/articles/PMC4154346/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4154346/.

[74] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. "Colorless green recurrent networks dream hierarchically". In: *arXiv preprint arXiv:1803.11138* (2018).

[75] Peter Hagoort. "How the brain solves the binding problem for language: a neurocomputational model of syntactic processing". In: *Neuroimage* 20 (2003), S18–S29.

[76] Peter Hagoort. "MUC (memory, unification, control) and beyond". In: *Frontiers in psychology* 4 (2013).

[77] Peter Hagoort. "The meaning-making mechanism (s) behind the eyes and between the ears". In: *Philosophical Transactions of the Royal Society B* 375.1791 (2020), p. 20190301.

[78] John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan R Brennan. "Finding Syntax in Human Encephalography with Beam Search". In: *arXiv preprint arXiv:1806.04127* (2018).

[79] Eric Halgren, Rupali P Dhond, Natalie Christensen, Cyma Van Petten, Ksenija Marinkovic, Jeffrey D Lewine, and Anders M Dale. "N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences". In: *Neuroimage* 17.3 (2002), pp. 1101–1116.

[80] Emma L Hall, Siân E Robson, Peter G Morris, and Matthew J Brookes. "The relationship between MEG and fMRI". In: *Neuroimage* 102 (2014), pp. 80–91.

[81] Liberty S Hamilton and Alexander G Huth. "The revolution will not be controlled: natural stimuli in speech neuroscience". In: *Language, Cognition and Neuroscience* 35.5 (2020), pp. 573–582.

[82] James A Hampton. "Conceptual combination". In: *Knowledge, concepts, and categories* (1997), pp. 133–159.

[83] Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. "Intersubject Synchronization of Cortical Activity during Natural Vision". In: *Science* 303.5664 (Mar. 2004), pp. 1634–1640. ISSN: 00368075. DOI: 10.1126/science.1089506. URL: http://science.sciencemag.org/.

[84] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. "On the interpretation of weight vectors of linear models in multivariate neuroimaging". In: *Neuroimage* 87 (2014), pp. 96–110.

[85] Martin N Hebart, Brett B Bankson, Assaf Harel, Chris I Baker, and Radoslaw M Cichy. "The representational dynamics of task and object processing in humans". In: *Elife* 7 (2018), e32816.

[86] Wendy A de Heer, Alexander G Huth, Thomas L Griffiths, Jack L Gallant, and Frédéric E Theunissen. "The hierarchical cortical organization of human speech processing". In: *Journal of Neuroscience* 37.27 (2017), pp. 6539–6557.

[87] Päivi Helenius, Riitta Salmelin, Elisabet Service, and John F Connolly. "Distinct time courses of word and context comprehension in the left temporal cortex." In: *Brain: a journal of neurology* 121.6 (1998), pp. 1133–1142.

[88] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. "Teaching machines to read and comprehend". In: *Advances in Neural Information Processing Systems*. 2015, pp. 1693–1701.

[89] Gregory Hickok and David Poeppel. "Neural basis of speech perception". In: *Neurobiology of Language*. Elsevier, 2016, pp. 299–310.

[90] Gregory Hickok and David Poeppel. "The cortical organization of speech processing". In: *Nature Reviews Neuroscience* 8.5 (2007), pp. 393–402.

[91] Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. "CogniVal: A Framework for Cognitive Word Embedding Evaluation". In: *arXiv preprint arXiv:1909.09001* (2019).

[92] Sture Holm. "A simple sequentially rejective multiple test procedure". In: *Scandinavian journal of statistics* (1979), pp. 65–70.

[93] Jeremy Howard and Sebastian Ruder. "Universal language model fine-tuning for text classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2018, pp. 328–339.

[94] Anne Hsu, Alexander Borst, and Frédéric E Theunissen. "Quantifying variability in neural responses and its application for the validation of model predictions". In: *Network: Computation in Neural Systems* 15.2 (2004), pp. 91–109.

[95] David H Hubel and Torsten N Wiesel. "Receptive fields and functional architecture of monkey striate cortex". In: *The Journal of physiology* 195.1 (1968), pp. 215–243.

[96] Annika Hultén, Jan-Mathijs Schoffelen, Julia Uddén, Nietzsche HL Lam, and Peter Hagoort. "How the brain makes sense beyond the processing of single words–An MEG study". In: *Neuroimage* 186 (2019), pp. 586–594.

[97] Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, Jack L Gallant, Wendy a De Heer, Thomas L Griffiths, and Jack L Gallant. "Natural speech reveals the semantic maps that tile human cerebral cortex". In: *Nature* 532.7600 (2016), pp. 453–458. DOI: 10.1038/nature17637.Natural.

[98] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. "A continuous semantic space describes the representation of thousands of object and action categories across the human brain". In: (2012). DOI: 10.1016/j.neuron.2012.10.014. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3556488/pdf/nihms418681.pdf.

[99] Shailee Jain and Alexander Huth. "Incorporating context into language encoding models for fMRI". In: *Advances in Neural Information Processing Systems*. 2018, pp. 6628–6637.

[100] Nancy Kanwisher and Ewa Wojciulik. "Visual attention: insights from brain imaging". In: *Nature Reviews Neuroscience* 1.2 (2000), pp. 91–100.

[101] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. "Identifying natural images from human brain activity". In: *Nature* 452.7185 (2008), p. 352.

[102] David Kemmerer. *Cognitive neuroscience of language*. Psychology Press, 2014.

[103] David Kemmerer. "Word classes in the brain: Implications of linguistic typology for cognitive neuroscience". In: *Cortex* 58 (2014), pp. 27–51.

[104] Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. "Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 284–294.

[105] Markus Kiefer. "Perceptual and semantic sources of category-specific effects: Event-related potentials during picture and word categorization". In: *Memory & Cognition* 29.1 (2001), pp. 100–116.

[106] Markus Kiefer, Eun-Jin Sim, Bärbel Herrnberger, Jo Grothe, and Klaus Hoenig. "The sound of concepts: four markers for a link between auditory and conceptual brain systems". In: *Journal of Neuroscience* 28.47 (2008), pp. 12224–12230.

[107] Jean-Rémi King and Stanislas Dehaene. "Characterizing the dynamics of mental representations: the temporal generalization method". In: *Trends in cognitive sciences* 18.4 (2014), pp. 203–210.

[108] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *Proceedings of the 3rd International Conference for Learning Representations*. ACM. 2015.

[109] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. "Skip-thought vectors". In: *Advances in neural information processing systems*. 2015, pp. 3294–3302.

[110] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. "Representational similarity analysis-connecting the branches of systems neuroscience". In: *Frontiers in Systems Neuroscience* 2 (2008), p. 4.

[111] Gina R Kuperberg. "Neural mechanisms of language comprehension: Challenges to syntax". In: *Brain research* 1146 (2007), pp. 23–49.

[112] Gina R Kuperberg, Phillip J Holcomb, Tatiana Sitnikova, Douglas Greve, Anders M Dale, and David Caplan. "Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies". In: *Journal of Cognitive Neuroscience* 15.2 (2003), pp. 272–293.

[113] Marta Kutas and Kara D Federmeier. "Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP)". In: *Annual review of psychology* 62 (2011), pp. 621–647.

[114] Marta Kutas, Cyma K Van Petten, and Robert Kluender. "Psycholinguistics electrified II (1994–2005)". In: *Handbook of psycholinguistics*. Elsevier, 2006, pp. 659–724.

[115] Brenden M Lake, Tal Linzen, and Marco Baroni. "Human few-shot learning of compositional instructions". In: *arXiv preprint arXiv:1901.04587* (2019).

[116] Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. "The emergence of number and syntax units in LSTM language models". In: *arXiv preprint arXiv:1903.07435* (2019).

[117] Matthew A Lambon Ralph, Gorana Pobric, and Elizabeth Jefferies. "Conceptual knowledge is underpinned by the temporal pole bilaterally: convergent evidence from rTMS". In: *Cerebral cortex* 19.4 (2009), pp. 832–838.

[118] Alessandro Lenci. "Distributional models of word meaning". In: *Annual review of Linguistics* 4 (2018), pp. 151–171.

[119] Yulia Lerner, Christopher J Honey, Lauren J Silbert, and Uri Hasson. "Topographic mapping of a hierarchy of temporal receptive windows using a narrated story". In: *The Journal of Neuroscience* 31.8 (2011), pp. 2906–2915.

[120] Mark D Lescroart and Jack L Gallant. "Human scene-selective areas represent 3D configurations of surfaces". In: *Neuron* 101.1 (2019), pp. 178–192.

[121] Hector Levesque, Ernest Davis, and Leora Morgenstern. "The winograd schema challenge". In: *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer. 2012.

[122] Long-Ji Lin. "Self-improving reactive agents based on reinforcement learning, planning and teaching". In: *Machine learning* 8.3-4 (1992), pp. 293–321.

[123] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. "Assessing the ability of LSTMs to learn syntax-sensitive dependencies". In: *arXiv preprint arXiv:1611.01368* (2016).

[124] Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. "Linguistic knowledge and transferability of contextual representations". In: *arXiv preprint arXiv:1903.08855* (2019).

[125] Bingjiang Lyu, Hun S Choi, William D Marslen-Wilson, Alex Clarke, Billi Randall, and Lorraine K Tyler. "Neural dynamics of semantic composition". In: *Proceedings of the National Academy of Sciences* 116.42 (2019), pp. 21318–21327.

[126] Joseph G Makin, David A Moses, and Edward F Chang. *Machine translation of cortical activity to text with an encoder–decoder framework*. Tech. rep. Nature Publishing Group, 2020.

[127] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[128] Rebecca Marvin and Tal Linzen. "Targeted syntactic evaluation of language models". In: *arXiv preprint arXiv:1808.09031* (2018).

[129] JL McClelland, BL McNaughton, R O'Reilly, and L Nadel. "Complementary roles of hippocampus and neocortex in learning and memory". In: *Society for Neuroscience Abstracts*. Vol. 18. 508.7. 1992, p. 1216.

[130] Michael McCloskey. "Networks and theories: The place of connectionism in cognitive science". In: *Psychological science* 2.6 (1991), pp. 387–395.

[131] BL McNaughton. "Comments in hippocampus symposium panel discussion". In: *Neurobiology of the hippocampus* (1983), pp. 609–610.

[132] Tomáš Mikolov et al. "Statistical language models based on neural networks". In: *Presentation at Google, Mountain View, 2nd April* 80 (2012), p. 26.

[133] Jeff Miller. "A warning about median reaction time." In: *Journal of Experimental Psychology: Human Perception and Performance* 14.3 (1988), p. 539.

[134] Jeff Mitchell and Mirella Lapata. "Composition in distributional models of semantics". In: *Cognitive science* 34.8 (2010), pp. 1388–1429.

[135] Tom Mitchell, Svetlana Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente Malave, Robert Mason, and Marcel Adam Just. "Predicting human brain activity associated with the meanings of nouns". In: *science* 320.5880 (2008), pp. 1191–1195.

[136] Brian Murphy, Leila Wehbe, and Alona Fyshe. "Decoding Language from the Brain". In: *Language, Cognition, and Computational Models* (2018), p. 53.

[137] Gregory L Murphy. "Noun phrase interpretation and conceptual combination". In: *Journal of memory and language* 29.3 (1990), pp. 259–288.

[138] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. *Encoding and decoding in fMRI*. May 2011. DOI: 10.1016/j.neuroimage.2010.07.073.

[139] Samuel A Nastase, Andrew C Connolly, Nikolaas N Oosterhof, Yaroslav O Halchenko, J Swaroop Guntupalli, Matteo Visconti di Oleggio Castello, Jason Gors, M Ida Gobbini, and James V Haxby. "Attention selectively reshapes the geometry of distributed semantic representation". In: *Cerebral Cortex* 27.8 (2017), pp. 4277–4291.

[140] Samuel A Nastase, Ariel Goldstein, and Uri Hasson. "Keep it real: rethinking the primacy of experimental control in cognitive neuroscience". In: *NeuroImage* 222 (2020), p. 117254.

[141] Lance Edward Nathan. "On the interpretation of concealed questions". PhD thesis. Massachusetts Institute of Technology, 2006.

[142] Satoshi Nishida, Yusuke Nakano, Antoine Blanc, and Shinji Nishimoto. "Brain-mediated Transfer Learning of Convolutional Neural Networks". In: *arXiv preprint arXiv:1905.10037* (2019).

[143] S. Nishimoto, A.T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J.L. Gallant. "Reconstructing visual experiences from brain activity evoked by natural movies". In: *Current Biology* (2011).

[144] S. Nishimoto, A.T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J.L. Gallant. "Reconstructing visual experiences from brain activity evoked by natural movies". In: *Current Biology* (2011).

[145] Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. "Voxelwise encoding models with non-spherical multivariate normal priors". In: *Neuroimage* 197 (2019), pp. 482–492.

[146] C. Pallier, A.D. Devauchelle, and S. Dehaene. "Cortical representation of the constituent structure of sentences". In: *Proceedings of the National Academy of Sciences* 108.6 (2011), pp. 2522–2527.

[147] Barbara Partee. "Noun phrase interpretation and type-shifting principles". In: *Formal semantics: The essential readings* (2002), pp. 357–381.

[148] Diane Pecher and Rolf A Zwaan. *Grounding cognition: The role of perception and action in memory, language, and thinking*. Cambridge University Press, 2005.

[149] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[150] Hao Peng, Roy Schwartz, Sam Thomson, and Noah A Smith. "Rational recurrences". In: *arXiv preprint arXiv:1808.09357* (2018).

[151] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep contextualized word representations". In: *Proceedings of NAACL-HLT*. 2018, pp. 2227–2237.

[152] Gorana Pobric, Stefan R Schweinberger, and Michal Lavidor. "Magnetic stimulation of the right visual cortex impairs form-specific priming". In: *Journal of cognitive neuroscience* 19.6 (2007), pp. 1013–1020.

[153] Russell A Poldrack. "Can cognitive processes be inferred from neuroimaging data?" In: *Trends in cognitive sciences* 10.2 (2006), pp. 59–63.

[154] Liina Pylkkänen. “Neural basis of basic composition: what we have learned from the red–boat studies and their extensions”. In: *Philosophical Transactions of the Royal Society B* 375.1791 (2020), p. 20190299.

[155] Maxim Rabinovich, Aaditya Ramdas, Michael I Jordan, and Martin J Wainwright. “Optimal rates and tradeoffs in multiple testing”. In: *arXiv preprint arXiv:1705.05391* (2017).

[156] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.

[157] Aniketh Janardhan Reddy and Leila Wehbe. “Syntactic representations in the human brain: beyond effort-based metrics”. In: *bioRxiv* (2020).

[158] Monica D. Rosenberg, Emily S. Finn, Dustin Scheinost, Xenophon Papademetris, Xilin Shen, R. Todd Constable, and Marvin M. Chun. “A neuromarker of sustained attention from whole-brain functional connectivity”. In: *Nature Neuroscience* 19.1 (Dec. 2015), pp. 165–171. ISSN: 15461726. DOI: 10.1038/nn.4179. URL: https://www.nature.com/articles/nn.4179.

[159] J.K. Rowling. *Harry Potter and the Sorcerer’s Stone*. Harry Potter US. Pottermore Limited, 2012. ISBN: 9781781100271. URL: http://books.google.com/books?id=wrOQLV6xB-wC.

[160] Gholamreza Salimi-Khorshidi, Gwenaëlle Douaud, Christian F. Beckmann, Matthew F. Glasser, Ludovica Griffanti, and Stephen M. Smith. “Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers”. In: *NeuroImage* 90 (Apr. 2014), pp. 449–468. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2013.11.046. URL: /pmc/articles/PMC4019210/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4019210/.

[161] Riitta Salmelin. “Clinical neurophysiology of language: the MEG approach”. In: *Clinical Neurophysiology* 118.2 (2007), pp. 237–254.

[162] Martin Schrimpf, Idan A Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy G Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. “Artificial Neural Networks Accurately Predict Language Processing in the Brain”. In: *BioRxiv* (2020).

[163] Dan Schwartz, Mariya Toneva, and Leila Wehbe. “Inducing brain-relevant bias in natural language processing models”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 14123–14133.

[164] Elisabet Service, Päivi Helenius, Sini Maury, and Riitta Salmelin. “Localization of syntactic and semantic brain responses using magnetoencephalography”. In: *Journal of Cognitive Neuroscience* 19.7 (2007), pp. 1193–1205.

[165] X. Shen, F. Tokoglu, X. Papademetris, and R. T. Constable. “Groupwise whole-brain parcellation from resting-state fMRI data for network node identification”. In: *NeuroImage* 82 (Nov. 2013), pp. 403–415. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2013.05.081. URL: /pmc/articles/PMC3759540/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3759540/.

[166] Kevin Sheppard, Stanislav Khrapov, Gábor Lipták, mikedeltalima, Rob Capellini, Hugle, esvhd, Alex Fortin, JPN, Austin Adams, jbrockmendel, M. Rabba, Michael E. Rose, Tom Rochette, Xavier RENE-CORAIL, and syncoding. *bashtage/arch: Release 4.15*. Version 4.15.

June 2020. DOI: 10.5281/zenodo.3906869. URL: https://doi.org/10.5281/zenodo.3906869.

[167] Peng Shi and Jimmy Lin. "Simple bert models for relation extraction and semantic role labeling". In: *arXiv preprint arXiv:1904.05255* (2019).

[168] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587 (2016), pp. 484–489.

[169] W Kyle Simmons, Vimal Ramjee, Michael S Beauchamp, Ken McRae, Alex Martin, and Lawrence W Barsalou. "A common neural substrate for perceiving and knowing about color". In: *Neuropsychologia* 45.12 (2007), pp. 2802–2810.

[170] Erez Simony, Christopher J. Honey, Janice Chen, Olga Lositsky, Yaara Yeshurun, Ami Wiesel, and Uri Hasson. "Dynamic reconfiguration of the default mode network during narrative comprehension". In: *Nature Communications* 7.1 (July 2016), pp. 1–13. ISSN: 20411723. DOI: 10.1038/ncomms12141. URL: https://www.nature.com/articles/ncomms12141.

[171] Michael A Skeide and Angela D Friederici. "The ontogeny of the cortical language network". In: *Nature Reviews Neuroscience* 17.5 (2016), p. 323.

[172] Edward E Smith and Daniel N Osherson. "Conceptual combination with prototype concepts". In: *Cognitive science* 8.4 (1984), pp. 337–361.

[173] Edward E Smith, Edward J Shoben, and Lance J Rips. "Structure and process in semantic memory: A featural model for semantic decisions." In: *Psychological review* 81.3 (1974), p. 214.

[174] Anders Søgaard. "Evaluating word embeddings with fMRI and eye-tracking". In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. 2016, pp. 116–121.

[175] Saurabh Sonkusare, Michael Breakspear, and Christine Guo. "Naturalistic stimuli in neuroscience: critically acclaimed". In: *Trends in cognitive sciences* 23.8 (2019), pp. 699–714.

[176] N.K. Speer, J.R. Reynolds, K.M. Swallow, and J.M. Zacks. "Reading stories activates neural representations of visual and motor experiences". In: *Psychological Science* 20.8 (2009), pp. 989–999.

[177] Donald T Stuss. "Frontal lobes and attention: processes and networks, fractionation and integration". In: *Journal of the International Neuropsychological Society* 12.2 (2006), pp. 261–271.

[178] 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". In: *NeuroImage* 62 (2012), pp. 451–463.

[179] 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". In: *NeuroImage* 62.1 (2012), pp. 451–463.

[180] Samu Taulu, Matti Kajola, and Juha Simola. "Suppression of interference and artifacts by the signal space separation method". In: *Brain topography* 16.4 (2004), pp. 269–275.

[181] Samu Taulu and Juha Simola. "Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements". In: *Physics in Medicine & Biology* 51.7 (2006), p. 1759.

[182] Ian Tenney, Dipanjan Das, and Ellie Pavlick. "BERT rediscovers the classical NLP pipeline". In: *arXiv preprint arXiv:1905.05950* (2019).

[183] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel Bowman, Dipanjan Das, et al. "What do you learn from context? Probing for sentence structure in contextualized word representations". In: *7th International Conference on Learning Representations, ICLR 2019*. 2019.

[184] Mariya Toneva, Tom M Mitchell, and Leila Wehbe. "Combining computational controls with natural text reveals new aspects of meaning composition". In: *bioRxiv* (2020). DOI: 10.1101/2020.09.28.316935. eprint: https://www.biorxiv.org/content/10.1101/2020.09.28.316935v2.full.pdf.

[185] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. "An Empirical Study of Example Forgetting during Deep Neural Network Learning". In: *International Conference on Learning Representations (ICLR)*. 2019.

[186] Mariya Toneva, Otilia Stretcu, Barnabas Poczos, Leila Wehbe, and Tom M Mitchell. "Modeling Task Effects on Meaning Representation in the Brain via Zero-Shot MEG Prediction". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[187] Mariya Toneva and Leila Wehbe. "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)". In: *Advances in Neural Information Processing Systems*. 2019, pp. 14928–14938.

[188] Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. "How Reasonable are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and SWAG". In: 2018.

[189] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. "Exploring the brain network: a review on resting-state fMRI functional connectivity". In: *European neuropsychopharmacology* 20.8 (2010), pp. 519–534.

[190] David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil. "The WU-Minn Human Connectome Project: An overview". In: *NeuroImage* 80 (Oct. 2013), pp. 62–79. ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.2013.05.041. URL: https://www.sciencedirect.com/science/article/pii/S1053811913005351?via%3Dihub.

[191] Cyma Van Petten and Barbara J Luka. "Neural localization of semantic context effects in electromagnetic and hemodynamic studies". In: *Brain and language* 97.3 (2006), pp. 279–293.

[192] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[193] Maya Visser, Elizabeth Jefferies, and MA Lambon Ralph. "Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature". In: *Journal of cognitive neuroscience* 22.6 (2010), pp. 1083–1094.

[194] Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. "Glue: A multi-task benchmark and analysis platform for natural language understanding". In: *arXiv preprint arXiv:1804.07461* (2018).

[195] Aria Y Wang, Leila Wehbe, and Michael Tarr. "Neural taskonomy: Inferring the similarity of task-derived representations from brain activity". In: (2019).

[196] Xiaosha Wang, Yangwen Xu, Yuwei Wang, Yi Zeng, Jiacai Zhang, Zhenhua Ling, and Yanchao Bi. "Representational similarity analysis reveals task-dependent semantic influence of the visual word form area". In: *Scientific reports* 8.1 (2018), pp. 1–10.

[197] Leila Wehbe, Idan A Blank, Cory Shain, Richard Futrell, Roger Levy, Titus von der Malsburg, Nathaniel Smith, Edward Gibson, and Evelina Fedorenko. "Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network". In: *bioRxiv* (2020).

[198] 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". In: *PLoS ONE* 9.11 (Nov. 2014). Ed. by Kevin Paterson, e112575. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0112575. URL: https://dx.plos.org/10.1371/journal.pone.0112575.

[199] Leila Wehbe, Aaditya Ramdas, Rebecca C Steorts, Cosma Rohilla Shalizi, et al. "Regularized brain reading with shrinkage and smoothing". In: *Annals of Applied Statistics* 9.4 (2015), pp. 1997–2022.

[200] Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom M. 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)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 233–243. URL: http://www.aclweb.org/anthology/D14-1030.

[201] Sebastian Weichwald, Timm Meyer, Ozan Özdenizci, Bernhard Schölkopf, Tonio Ball, and Moritz Grosse-Wentrup. "Causal interpretation rules for encoding and decoding models in neuroimaging". In: *Neuroimage* 110 (2015), pp. 48–59.

[202] Gail Weiss, Yoav Goldberg, and Eran Yahav. "On the practical computational power of finite precision RNNs for language recognition". In: *arXiv preprint arXiv:1805.04908* (2018).

[203] Masha Westerlund and Liina Pylkkänen. "The role of the left anterior temporal lobe in semantic composition vs. semantic memory". In: *Neuropsychologia* 57 (2014), pp. 59–70.

[204] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. "HuggingFace's Transformers: State-of-the-art Natural Language Processing". In: *ArXiv* abs/1910.03771 (2019).

[205] Michael C-K Wu, Stephen V David, and Jack L Gallant. "Complete functional characterization of sensory neurons by system identification". In: *Annu. Rev. Neurosci.* 29 (2006), pp. 477–505.

[206] Yangwen Xu, Xiaosha Wang, Xiaoying Wang, Weiwei Men, Jia-Hong Gao, and Yanchao Bi. "Doctor, teacher, and stethoscope: neural representation of different types of semantic relations". In: *Journal of Neuroscience* 38.13 (2018), pp. 3303–3317.

[207] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. "Performance-optimized hierarchical models predict neural responses in higher visual cortex". In: *Proceedings of the national academy of sciences* 111.23 (2014), pp. 8619–8624.

[208] Michael M Yartsev. "The emperor's new wardrobe: rebalancing diversity of animal models in neuroscience research". In: *Science* 358.6362 (2017), pp. 466–469.

[209] Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. "Learning and evaluating general linguistic intelligence". In: *arXiv preprint arXiv:1901.11373* (2019).

[210] Linmin Zhang and Liina Pylkkänen. "The interplay of composition and concept specificity in the left anterior temporal lobe: An MEG study". In: *NeuroImage* 111 (2015), pp. 228–240.

[211] Xunjie Zhu, Tingfeng Li, and Gerard de Melo. "Exploring Semantic Properties of Sentence Embeddings". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2018, pp. 632–637.

[212] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 19–27.

[213] Jayden Ziegler and Liina Pylkkänen. "Scalar adjectives and the temporal unfolding of semantic composition: An MEG investigation". In: *Neuropsychologia* 89 (2016), pp. 161–171.