# Neuro-computational models of language comprehension : characterizing similarities and differences between language processing in brains and language models.

Subba Reddy Oota

THÈSE PRÉSENTÉE
POUR OBTENIR LE GRADE DE

**DOCTEUR DE**
**L'UNIVERSITÉ DE BORDEAUX**

ÉCOLE DOCTORALE MATHÉMATIQUES ET INFORMATIQUE

SPÉCIALITÉ
Informatique

Subba Reddy OOTA

---

# Modèles neurocomputationnels de la compréhension du langage : caractérisation des similarités et des différences entre le traitement cérébral du langage et les modèles de langage

# Neuro-Computational Models of Language Comprehension: characterizing similarities and differences between language processing in brains and language models

---

Sous la direction de: Xavier HINAUT
et de Frederic ALEXANDRE

Soutenue le April 30, 2024

Membres du Jury:

| | | | |
|---|---|---|---|
| M. Xavier HINAUT | Chercheur (HDR) | Inria Bordeaux | **Directeur de thèse** |
| M. Frederic ALEXANDRE | Directeur de recherche | Inria Bordeaux | **Directeur de thèse** |
| M. Christophe PALLIER | Directeur de recherche | CNRS | **Rapporteur** |
| M. Stefan FRANK | Professeur | Radbourd University | **Rapporteur** |
| Mme. Leila WEHBE | Maîtresse de conférences | CMU | **Examinatrice** |
| M. Fabien LOTTE | Directeur de recherche | Inria Bordeaux | **Examinateur** |
| M. Gael JOBARD | Maitre de conférence | Université de Bordeaux | **Invité** |

**Title of the thesis:** Neuro-Computational Models of Language Comprehension: characterizing similarities and differences between language processing in brains and language models

**Abstract**

This thesis explores the synergy between artificial intelligence (AI) and cognitive neuroscience to advance language processing capabilities. It builds on the insight that breakthroughs in AI, such as convolutional neural networks and mechanisms like experience replay [1], often draw inspiration from neuroscientific findings. This interconnection is beneficial in language, where a deeper comprehension of uniquely human cognitive abilities, such as processing complex linguistic structures, can pave the way for more sophisticated language processing systems. The emergence of rich naturalistic neuroimaging datasets (e.g., fMRI, MEG) alongside advanced language models opens new pathways for aligning computational language models with human brain activity. However, the challenge lies in discerning which model features best mirror the language comprehension processes in the brain, underscoring the importance of integrating biologically inspired mechanisms into computational models.

In response to this challenge, the thesis introduces a data-driven framework bridging the gap between neurolinguistic processing observed in the human brain and the computational mechanisms of natural language processing (NLP) systems. By establishing a direct link between advanced imaging techniques and NLP processes, it conceptualizes brain information processing as a dynamic interplay of three critical components: "what," "where," and "when", offering insights into how the brain interprets language during engagement with naturalistic narratives. This study provides compelling evidence that enhancing the alignment between brain activity and NLP systems offers mutual benefits to the fields of neurolinguistics and NLP. The research showcases how these computational models can emulate the brain's natural language processing capabilities by harnessing cutting-edge neural network technologies across various modalities—language, vision, and speech. Specifically, the thesis highlights how modern pretrained language models achieve closer brain alignment during narrative comprehension. It investigates the differential processing of language across brain regions, the timing of responses (Hemodynamic Response Function (HRF) delays), and the balance between syntactic and semantic information processing. Further, the exploration of how different linguistic features align with MEG brain responses over time and find that the alignment depends on the amount of past context, indicating that the brain encodes words slightly behind the current one, awaiting more future context. Furthermore, it highlights grounded language acquisition through noisy supervision and offers a biologically plausible architecture for investigating cross-situational learning, providing interpretability, generalizability, and computational efficiency in sequence-based models. Ultimately, this research contributes valuable insights into neurolinguistics, cognitive neuroscience, and NLP.

**Keywords:** brain encoding; natural language processing; fMRI; MEG; hemodynamic response function delays; language models; brain language processing; Recurrent neural network; Reservoir computing; Transformers.

[1] Human-level control through deep reinforcement learning [Mnih et al., 2015]

# ACKNOWLEDGMENTS

I am deeply indebted to a myriad of individuals who have contributed to the development of this thesis, both directly and indirectly. Completing this work would not have been possible without the invaluable guidance of my advisors, Dr. Xavier Hinaut and Prof. Alexandre Frederic. My journey began during my master's at IIIT-Hyderabad, where my initial foray into machine learning models and brain language processing set the direction for my PhD research under the tutelage of Xavier and Frederic. Their mentorship has been a privilege, with Xavier empowering me to pursue my scientific interests and exemplifying clear communication of ideas—a skill I aspire to master. His efforts in connecting me with numerous collaborators have been instrumental to this thesis. Frederic has been exceptionally patient and insightful, generously offering his time and ideas to facilitate my research growth.

I am also grateful for the support of the rest of my thesis committee–Prof. Christophe Pallier, Prof. Stefan Frank, Prof. Leila Wehbe, Prof. Fabien Lotte and Dr. Gael Jobard. Their insightful feedback has been instrumental in refining the main arguments of my thesis and has enhanced my ability to communicate our findings to a multidisciplinary audience. My PhD journey was significantly enriched by a research internship at the Max Planck Institute for Software Systems (MPI-SWS) under the guidance of Dr. Mariya Toneva. This opportunity was a cornerstone of my academic path, revealing how my scientific mindset could contribute to a deeper understanding of black-box natural language processing models and their intersection with brain language processing. This revelation propelled me towards new research directions. Furthermore, Mariya Toneva's connections with collaborators like Fatma Deniz and Emin Celik have been instrumental in providing new opportunities for my career.

A heartfelt thanks go to Dr. Manish Gupta and Prof. Bapi Raju, whose mentorship during my Masters paved the way for my journey into research and ultimately towards pursuing a PhD. Their guidance has been invaluable, shaping my understanding of how to formulate problems and write papers. Additionally, I would like to acknowledge the significant contributions of several faculty members and colleagues at Inria. Special thanks are extended to Chrystel Plumejeau and Anne Lise for their instrumental roles in providing substantial financial support for conferences, as well as for their welcoming administrative assistance throughout.

Collaborations with numerous talented researchers enriched my PhD journey. I am particularly grateful to Gael Jobard for his valuable ideas and insights that greatly enhanced Chapter 5, offering a systematic examination of the intricate processing in various language regions of the human brain at differing HRF delays. I also owe a debt of gratitude to Manish Gupta, Mariya Toneva, and Bapi Raju for their significant contributions to the survey on aligning deep neural networks with brain processing in Chapter 3. Additionally, I thank them for their efforts in conducting high-quality tutorials at various conference

venues, which laid the groundwork for the achievements in Chapter 3. My collaboration with Nathan Trouvain has been enriching. Nathan is not just a brilliant thinker but also a fantastic friend and one of my favorite companions for conference travel. Our conversations have profoundly influenced my research, including the work that forms the foundation of Chapter 6.

Relocating to numerous places, Bordeaux emerged as the first to truly resonate with me as home, significantly due to the incredible individuals I encountered there. I am grateful to both current and former members of the Mnemosyne group—Hugo Chateau-Laurent, Naomi Chaix-Eichel, Maeva Andriantsoamberomanga, Fjola Hyseni, Chloé Mercier, Axel Palaude, Nathan Trouvain, Snigdha Dagar, and Remya Shankar—for their camaraderie and contributions to a vibrant community. To my dear home friends Blanche Benedict and Vagini, sharing the triumphs and challenges of my PhD journey with you has been invaluable. Your support and friendship have enriched my experience in ways words cannot fully capture. I also want to express my gratitude to the many remarkable people whose paths crossed with mine—Aw Khai Loong, Gabriele Merlin, Ruchit Rawal, Ding Tong, RJ Antonello, Sukesh Adiga, Mounika Marreddy, Vijay Rowtula, Jashn Arora, Veeral Agarwal, Venkat Charan Chinni, Sirisha Vakada, Kushbu Pahwa, Pawan Kumar Neerudu, and many others. Your presence brought laughter and joy, lightening the load of the most demanding years. Thank you for the memories and moments that we shared.

Finally, I extend my heartfelt gratitude to my family. To my parents, Venkayemma and Rami Reddy, I am deeply thankful for instilling in me the value of education and the importance of staying connected to my roots. My gratitude also goes to my brother, Konda Reddy, and his family for supporting my parents back in our hometown. To my extended family, sister Mounika, bother-in-law Uday, and niece Ruhi, your guidance has been invaluable, and to Ramadevi and Sanjeeva Reddy, your warm hospitality and the delicious snack recipes have been a source of comfort and motivation throughout my research journey. To my partner, Nagamani, your unwavering support and sacrifices for our family's betterment have been the foundation of my strength and perseverance. Your patience, kindness, and ability to make tough times bearable and good times even more joyful have been my most outstanding support. Moreover, to my two beautiful children, Moksha and Yokshith, your love has been the brightest part of my journey. I eagerly anticipate the adventures that lie ahead for us. To the extended family of my partner, which includes my brother-in-law Mallikarjuna Reddy, sister-in-law Lakshmi, and mother-in-law Ravana, I extend my deepest gratitude for your unwavering support in caring for my children during emergency situations. I am also profoundly thankful to my childhood mentors, M. Sambi Reddy, and my relatives, Srinivasa Reddy, Sai Brunda, Rami Reddy, Lakshmi Reddy, Sasi Kanth Reddy, and Koti Reddy, along with my friends, Rupesh Sanagapalli and Sreekanth, for their invaluable support and guidance. Your assistance has been truly exceptional. I look forward to our next chapter.

# ACRONYMS

| | |
|---|---|
| AAC | Auditory Association Cortex |
| AC | Auditory Cortex |
| AG | Angular Gyrus |
| AI | Artificial Intelligence |
| ANN | Artificial Neural Networks |
| ASR | Automated Speech Recognition |
| ATL | Anterior Temporal Lobe |
| BCI | Brain-Computer Interface |
| BERT | Bidirectional Encoder Representation Transformers |
| BOLD | Blood Oxygen Level-Dependent |
| CCS | Computational Cognitive Science |
| CL | Continual Learning |
| CLIP | Contrastive Language-Image Pre-training |
| CNN | Convolutional Neural Networks |
| CSL | Cross-Situational Learning |
| CV | Computer Vision |
| CVD | Cross-View Decoding |
| DFL | Dorsal Frontal Lobe |
| DNN | Deep Neural Networks |
| EAC | Early Auditory |
| ECoG | Electro-Corticography |
| EEG | Electroencephalography |
| ELMo | Embeddings From Language Model |
| ER | Entity Recognition |
| ESN | Echo-State Networks |
| FDR | False Discovery Rate |
| FFT | Fast Fourier Transform |
| fMRI | functional Magnetic Resonance Imaging |
| fNIRS | Functional Near-Infrared Spectroscopy |
| GLUE | General Language Understanding Evaluation |
| GPT | Generalized Pretrained Transformer |
| HRF | Hemodynamic Response Function |
| IC | Intent Classification |
| IFG | Inferior Frontal Gyrus |
| LATL | Left Anterior Temporal Lobe |
| LED | Longformer Encoder Decoder |

| | |
|---|---|
| LM | Language Model |
| LSTM | Long Short-Term Memory Networks |
| LXMERT | Cross-Modality Encoder Representations from Transformers |
| ME | Micro-Electrode |
| MEA | Micro-Electrode Array |
| MEG | Magnetoencephalography |
| MFCC | Mel Frequency Cepstral Coefficients |
| MFG | Middle Frontal Gyrus |
| ML | Machine Learning |
| MLM | Masked Language Modeling |
| MRI | Magnetic Resonance Imaging |
| MTG | Middle Temporal Gyrus |
| MVD | Multi-View Decoding |
| NIRS | Near-Infrared Spectroscopy |
| NLP | Natural Language Processing |
| NLU | Natural Language Understanding |
| PCA | Principal Component Analysis |
| PCC | Pearson Correlation Coefficient |
| PMC | Posterior Medial Cortex |
| POS | Part-of-Speech |
| pSTG | Posterior Superior Temporal Gyrus |
| pSTS | Posterior Superior Temporal Sulcus |
| PTL | Posterior Temporal Lobe |
| PTM | Pre-trained Models |
| RNN | Recurrent Neural Networks |
| ROI | Region Of Interest |
| rTMS | repetitive Transcranial Magnetic Stimulation |
| SID | Speaker Identification |
| STG | Superior Temporal Gyrus |
| TPJ | Temporo-Parietal Junction |
| TPOJ | Temporo Parieto-Occipital junction |
| TR | Time Repetition |

# Contents

# List of Figures

# List of Tables

# RÉSUMÉ DE THÈSE

**Titre de la thèse:** Modèles neurocomputationnels de la compréhension du langage : caractérisation des similarités et des différences entre le traitement cérébral du langage et les modèles de langage

**Résumé** Cette thèse explore la synergie entre l'intelligence artificielle (IA) et la neuroscience cognitive pour faire progresser les capacités de traitement du langage. Elle s'appuie sur l'idée que les avancées en IA, telles que les réseaux neuronaux convolutionnels et des mécanismes comme le « replay d'expérience » [1], s'inspirent souvent des découvertes neuroscientifiques. Cette interconnexion est bénéfique dans le domaine du langage, où une compréhension plus profonde des capacités cognitives humaines uniques, telles que le traitement de structures linguistiques complexes, peut ouvrir la voie à des systèmes de traitement du langage plus sophistiqués. L'émergence de riches ensembles de données neuroimagerie naturalistes (par exemple, fMRI, MEG) aux côtés de modèles de langage avancés ouvre de nouvelles voies pour aligner les modèles de langage computationnels sur l'activité cérébrale humaine. Cependant, le défi réside dans le discernement des caractéristiques du modèle qui reflètent le mieux les processus de compréhension du langage dans le cerveau, soulignant ainsi l'importance d'intégrer des mécanismes inspirés de la biologie dans les modèles computationnels. En réponse à ce défi, la thèse introduit un cadre basé sur les données qui comble le fossé entre le traitement neurolinguistique observé dans le cerveau humain et les mécanismes computationnels des systèmes de traitement automatique du langage naturel (TALN). En établissant un lien direct entre les techniques d'imagerie avancées et les processus de TALN, elle conceptualise le traitement de l'information cérébrale comme une interaction dynamique de trois composantes critiques : le « quoi », le « où » et le « quand », offrant ainsi des perspectives sur la manière dont le cerveau interprète le langage lors de l'engagement avec des récits en conditions écologiques. L'étude fournit des preuves convaincantes que l'amélioration de l'alignement entre l'activité cérébrale et les systèmes de TALN offre des avantages mutuels aux domaines de la neurolinguistique et du TALN. La recherche montre comment ces modèles computationnels peuvent émuler les capacités de traitement du langage naturel du cerveau en exploitant les technologies de réseau neuronal de pointe dans diverses modalités - langage, vision et parole. Plus précisément, la thèse met en lumière comment les modèles de langage pré-entraînés modernes parviennent à un alignement plus étroit avec le cerveau lors de la compréhension de récits. Elle examine le traitement différentiel du langage à travers les régions cérébrales, le timing des réponses (délais La fonction de réponse hémodynamique (HRF)) et l'équilibre entre le traitement de l'information syntaxique et sémantique. En outre, elle explore comment différentes caractéristiques linguistiques s'alignent avec les réponses cérébrales MEG au fil du temps et con-

[1] Human-level control through deep reinforcement learning [Mnih et al., 2015]

state que cet alignement dépend de la quantité de contexte passé, indiquant que le cerveau code les mots légèrement en retard par rapport à celui actuel, en attendant plus de contexte futur. De plus, elle met en évidence la plausibilité biologique de l'apprentissage des états de réservoir de calcul, offrant ainsi une interprétabilité, une généralisabilité et une efficacité computationnelle dans les modèles basés sur des séquences. En fin de compte, cette recherche apporte des contributions précieuses à la neurolinguistique, à la neuroscience cognitive et au TALN.

**Mots-clés de la thèse:** Traitement cérébral du cerveau; Apprentissage développemental du langage; Neurosciences computationnelles; codage cérébral; traitement du langage naturel; Réseau de neurones récurrent; Calcul en réservoir; Transformateurs.

## ÉNONCÉ DE THÈSE ET PLAN

Cette thèse s'articule autour de l'énoncé suivant : Combler l'écart entre les modèles actuels de réseaux neuronaux profonds et la compréhension du langage dans le cerveau : 1) notre compréhension mécaniste de la compréhension du langage dans le cerveau à travers la plausibilité à long terme des modèles de langage, 2) une compréhension plus profonde du traitement du langage dans le cerveau par l'interprétation des fonctions de réponse hémodynamique grâce à la modélisation computationnelle, et 3) la performance au niveau des mots et du sémantique des modèles de traitement automatique du langage naturel grâce au transfert des insights du cerveau.

**Chapitre 3 :** Résume les derniers efforts sur la manière dont les réseaux neuronaux profonds commencent à résoudre des problèmes computationnels (encodage et décodage) à travers diverses modalités (langage, vision et parole) et illuminent ainsi les calculs que le cerveau accomplit sans effort. Nous discutons en particulier des représentations populaires des stimuli de langage, de vision et de parole dérivées des embeddings de mots statiques, des modèles basés sur des séquences, des transformateurs et des modèles basés sur des transformateurs multimodaux. Nous présentons ensuite un résumé des ensembles de données cérébrales naturalistes et des métriques d'évaluation populaires, passons en revue les architectures populaires d'encodage et de décodage basées sur l'apprentissage profond, et notons leurs avantages et leurs limitations dans le contexte de l'alignement cérébral.

Un long article de revue a été soumis au journal Transactions on Machine Learning Research (TMLR), actuellement en cours de révision.

**Chapitre 4 :** Résume l'efficacité de divers modèles de langage, notamment les réseaux de mémoire à court terme (LSTMs), ELMo et Longformer, dans la prédiction de l'activité cérébrale lorsque les sujets écoutent des histoires narratives. Bien que ces modèles aient réussi à prédire l'activation cérébrale basée sur le texte, ils ne peuvent toujours pas gérer les dépendances à long terme et fournir des aperçus des mécanismes neuronaux en jeu. Nos résultats suggèrent que les états de cellule LSTM s'alignent mieux avec les enregistrements cérébraux que les états cachés LSTM, indiquant leur capacité à capturer des informations à long terme. De plus, les représentations ELMo et Longformer prédisent l'activité cérébrale dans différentes régions du langage cérébral.

Un long article a été publié précédemment lors de la $44^e$ conférence annuelle de la Société des sciences cognitives (juillet 2022, Toronto, Canada).

**Chapitre 5 :** Enquête sur la manière dont différents retards de fonction de réponse hémodynamique dans la fonction de réponse du cerveau affectent l'alignement entre les représentations des modèles de langage et les enregistrements cérébraux obtenus. En même temps, les participants écoutent ou lisent une histoire. Nous explorons l'importance relative de l'information syntaxique (c'est-à-dire, les embeddings syntaxiques basés sur les arbres de constituants) par rapport à l'information sémantique en utilisant des modèles de langage open-source tels que BERT, GPT-2 et Llama-2. De plus, nous examinons différentes longueurs de contexte et révélons des différences dans la façon dont le cerveau traite le langage à travers différents retards.

**Chapitre 6 :** Résume l'utilisation de la magnétoencéphalographie (MEG), avec une résolution temporelle plus élevée que l'IRM fonctionnelle, nous permet de regarder de manière plus précise le timing du traitement des caractéristiques linguistiques. Inspirés par des études précédentes d'encodage IRM fonctionnel, nous étudions l'encodage cérébral MEG en utilisant des caractéristiques syntaxiques et sémantiques de base, avec différentes longueurs de contexte et directions (passé vs futur), pour un ensemble de données de 8 sujets écoutant des histoires. Nous avons montré que le modèle Bidirectional Transformer (BERT), contrairement à d'autres caractéristiques, conduit à une prédiction significative de l'activité cérébrale MEG dans les régions auditives et langagières entre 50-550 ms (250-750 ms avec un début de mot à 200 ms).

Un long article a été publié précédemment lors de la $24^e$ conférence INTERSPEECH (août 2023, Dublin, Irlande).

**Chapitre 7 :** Résume comment les enfants apprennent la langue et comment leur cerveau la traite, en appliquant des connaissances à l'apprentissage automatique et à la robotique. Nous nous concentrons principalement sur l'apprentissage trans-situationnel (CSL) en utilisant des phrases complètes pour comprendre le développement précoce du langage. La recherche compare différents modèles et représentations de mots, trouvant que les représentations BERT affinées fonctionnent le mieux. Ces modèles peuvent aider l'interaction humain-robot et améliorer notre compréhension de l'acquisition du langage chez les enfants.

Un rapport initial a été présenté lors de l'atelier Splu-RoboNLP à ACL en juillet 2021, puis étendu en un article de revue. Il est actuellement en cours de révisions mineures pour publication dans le journal Nature Scientific Reports.

# RÉSUMÉ DES CONTRIBUTIONS

Les contributions de chaque chapitre de cette thèse peuvent être résumées comme suit :

### AMÉLIORER L'INFÉRENCE SCIENTIFIQUE POUR LES MODÈLES D'ENCODAGE EN UTILISANT DES ENSEMBLES DE DONNÉES CÉRÉBRALES NATURALISTES ÉTENDUS ET LES PROGRÈS DE L'IA GÉNÉRATIVE

L'objectif central des neurosciences est de comprendre comment le cerveau représente l'information et la traite pour accomplir diverses tâches (visuelles, linguistiques, auditives, etc.). Les réseaux neuronaux profonds (DNN) offrent un moyen computationnel de capturer la complexité et la richesse sans précédent de l'activité cérébrale. L'encodage et le décodage, formulés comme des problèmes computationnels, résument de manière succincte ce puzzle. Le domaine évolue rapidement avec la disponibilité de vastes ensembles de données en neuroimagerie lorsque les participants traitent des stimuli dans des environnements naturels. Parallèlement, il existe des progrès considérables dans les réseaux neuronaux profonds (DNN) qui traitent efficacement et robustement des données multimodales. En s'inspirant de l'efficacité des récents modèles d'IA générative pour le traitement du langage naturel, la vision par ordinateur et la parole, nous passons en revue les architectures populaires d'encodage et de décodage basées sur l'apprentissage profond et notons leurs avantages et leurs limitations dans le contexte de l'alignement cérébral. Dans le chapitre 3, nous résumons divers modèles d'encodage sous forme d'arbre de classification taxonomique. Ces modèles s'adressent aux domaines de la vision, de l'auditif, du langage et des multimédias. Étant donné l'abondance de publications récentes dans ce domaine, le chapitre 3 vise à faciliter les contributions de la communauté des neurosciences cognitives computationnelles, contribuant ainsi à faire progresser le domaine de l'encodage et du décodage cérébral.

### DÉVOILER LE SUBSTRAT NEURONAL : MODÈLES DE LANGAGE ET DÉPENDANCES À LONG TERME DANS LA PRÉDICTION DE L'ACTIVATION CÉRÉBRALE

Plusieurs modèles de langage préentraînés basés sur des séquences et populaires se sont révélés efficaces pour la prédiction basée sur le texte des activations cérébrales [Jain and Huth, 2018, Toneva and Wehbe, 2019]. Cependant, ces modèles manquent toujours de plausibilité en termes de mémoire à long terme (c'est-à-dire, comment ils gèrent les dépendances à long terme et l'information contextuelle) et d'aperçus des mécanismes sous-jacents du substrat neuronal. De plus, les récents modèles Transformer préentraînés comme BERT et GPT-2 ne peuvent pas traiter les dépendances à long terme (la longueur de séquence est fixée à 512 mots) en raison de leur opération d'auto-attention. Pour surmonter cette limitation, récemment, Beltagy et al. [2020] a introduit *Longformer*, facilitant le traitement de documents de milliers de jetons ou plus et combinant une attention locale par fenêtre avec une attention globale. En tenant compte de ces défis, le chapitre 4 de cette thèse vise à éclairer la relation entre les activations de voxels fMRI et les représentations générées par divers modèles de langage. Nos résultats suggèrent que le développement de modèles de langage capables de gérer des informations contextuelles plus étendues et d'interpréter les représentations internes de ces modèles peut conduire à une compréhension plus approfondie de la manière dont les structures neuronales représentent l'information linguistique et maintiennent une mémoire narrative plus longue.

### DÉVOILER L'INTERACTION DES RETARDS DE RÉPONSE HÉMODYNAMIQUE ET DU TRAITEMENT DU LANGAGE DANS LE CERVEAU

L'augmentation de la disponibilité des ensembles de données fMRI issues de tâches écologiques et des modèles neuronaux à grande échelle peut permettre une meilleure compréhension de la réponse du cerveau aux stimuli naturels. Rien que ces dernières années, les chercheurs ont montré que les réponses cérébrales des personnes comprenant le langage peuvent être bien prédites par des modèles de langage basés sur le texte [Wehbe et al., 2014, Jain and Huth, 2018, Toneva and Wehbe, 2019, Deniz et al., 2019, Caucheteux and King, 2020, Schrimpf et al., 2021b, Caucheteux et al., 2021a, Toneva et al., 2022, Oota et al., 2022c, Antonello et al., 2021, Aw and Toneva, 2023, Merlin and Toneva, 2022]. Cependant, les études existantes sur l'alignement entre la compréhension du langage et le cerveau ont été observées à un retard constant de la fonction de réponse hémodynamique (HRF) (environ 7.5 à 8 secondes). Il y a encore une exploration en cours sur la manière dont le langage et les mécanismes de traitement du cerveau se synchronisent lorsque confrontés à différents retards de HRF [Jain and Huth, 2018, Jain et al., 2020, Toneva and Wehbe, 2019, Deniz et al., 2019, Toneva et al., 2022, Aw and Toneva, 2023, Oota et al., 2022c, 2023c]. De plus, les études existantes ont principalement construit des modèles d'encodage cérébral en considérant un retard de HRF fixe et en analysant comment différentes régions d'intérêt (ROIs) impliquées dans le traitement du langage influencent les aspects sémantiques et syntaxiques du traitement de l'information dans le cerveau [Jain and Huth, 2018, Jain et al., 2020, Toneva and Wehbe, 2019, Caucheteux et al., 2021a, Toneva et al., 2022, Merlin and Toneva, 2022, Aw and Toneva, 2023, Oota et al., 2022c, 2023c]. Dans cette thèse, l'interaction systématique entre les retards de HRF et le traitement du langage est un domaine d'investigation, visant à comprendre comment l'activité neurale liée aux tâches linguistiques s'aligne avec la réponse hémodynamique subséquente, et comment cet alignement peut différer selon les conditions variables des retards de HRF. Nos résultats suggèrent que la décomposition des représentations en différentes caractéristiques linguistiques permet une compréhension fine du traitement du langage par le cerveau à travers différents retards, ouvrant la voie à des approches plus personnalisées et efficaces dans les applications linguistiques et cliniques.

### EXPLORATION DU TIMING DU TRAITEMENT DES CARACTÉRISTIQUES LINGUISTIQUES DANS LE CERVEAU AVEC MEG

Au cours de la dernière décennie, les interfaces cerveau-ordinateur (BCI) ont contribué à des avancées significatives dans la compréhension du traitement du langage dans le cerveau en utilisant un paradigme computationnel populaire : l'encodage cérébral, le processus visant à mapper les caractéristiques des stimuli sur l'activité cérébrale. Il existe une vaste littérature sur l'encodage cérébral linguistique pour l'IRM fonctionnelle (IRMf) liée aux représentations syntaxiques et sémantiques. La magnétoencéphalographie (MEG), avec une résolution temporelle plus élevée que l'IRMf, nous permet d'examiner plus précisément le timing du traitement des caractéristiques linguistiques. Contrairement au décodage MEG, peu d'études sur l'encodage MEG en utilisant des stimuli naturels existent. Les études existantes sur l'écoute d'histoires se concentrent sur les phonèmes et les caractéristiques

de mots simples, en ignorant des caractéristiques plus abstraites telles que le contexte, les aspects syntaxiques et sémantiques. Pour comprendre quand le cerveau traite la structure linguistique dans les phrases, dans cette thèse, le chapitre 5 exploite les représentations textuelles en utilisant des caractéristiques syntaxiques de base et des caractéristiques sémantiques, avec différentes longueurs de contexte, directions (passé vs futur) et importance relative dans le contexte.

### Supervision Bruitée dans l'Acquisition de Langage Ancrée : Une Perspective de Modèle de Langage

L'acquisition de langage ancrée englobe le processus d'acquisition d'une langue, dans lequel les nourrissons apprennent en observant leur environnement, en interagissant avec les autres et en saisissant les concepts d'une langue dans le contexte du monde réel [Yu and Ballard, 2004a,b, 2007, Chen and Mooney, 2008, Thomason et al., 2018, Juven and Hinaut, 2020, Vanzo et al., 2020]. Cependant, l'acquisition du langage devient difficile. Un seul mot dans une énonciation peut avoir plusieurs significations potentielles, introduisant une grande incertitude. Les approches traditionnelles de l'ancrage du langage se concentrent principalement sur la mise en correspondance des commandes de langage naturel avec des représentations, impliquant souvent des séquences d'actions robotiques fondamentales [Chen and Mooney, 2011, Matuszek et al., 2013, Tellex et al., 2011]. De plus, les cadres robotiques actuels Taniguchi et al. [2017], Roesler et al. [2018] ne traitent pas la manière dont les enfants apprennent naturellement à comprendre des phrases complètes par l'apprentissage inter-situationnel sans indices spécifiques. Face à ces défis, le chapitre 7 de cette thèse se penche sur une enquête sur la façon dont les modèles de langage peuvent entreprendre l'acquisition de langage ancrée dans des conditions de supervision bruyante. Il explore également comment ces modèles peuvent rendre compte de la dynamique de l'apprentissage dans le cerveau.

# 1 Introduction

Human language is an incredibly complex ability, yet children can learn languages quickly (in a few years). We still need to gain more knowledge of such language learning mechanisms. Experiments in neuroscience and developmental psychology provide different hypotheses on potential mechanisms. However, it is difficult to grasp such mechanisms, partly because of the multiple modalities (audition, vision, ...) implied in such processes. Thus, modeling appears as an appealing and complementary tool to provide a deeper understanding of such language mechanisms and take apart the plausible ones from the implausible ones, finally providing a more transparent view for experimentalists. Furthermore, modeling has significantly impacted fields like artificial intelligence (AI) and machine learning (ML). For instance, the exploration of cell receptive fields and the way information is processed in the early visual system, as described by Hubel and Wiesel [1968], played a crucial role in the creation of deep networks, convolutional neural networks introduced by Fukushima and Miyake [1982]. These networks brought about a significant transformation in the field of computer vision. Similarly, the understanding that replaying past experiences in the hippocampus enhances memory consolidation, as proposed by McNaughton [1983], served as inspiration for the development of experience replay, as articulated by McClelland and Goddard [1996].

Deep Learning models have recently created a breakthrough in image and speech recognition and Natural Language Processing (NLP) methods. However, no equivalent breakthrough happened in understanding how brains perform similar functions. This breakthrough did not happen because Deep Learning does not reproduce learning mechanisms or brain dynamics. Thus, we still need critical neuronal mechanisms to model language comprehension and production functions properly.

Developing new mechanistic models of brain activity can help better understand how the brain works. This is particularly true for language, where a better understanding of the cognitive mechanisms involved could lead to improved treatment for developmental language disorders in children and rehabilitation methods for brain injuries that lead to aphasia. The development of theoretical models for the neural dynamics underlying brain functions, including learning, is essential to (1) better understand the general functioning of the brain and (2) explore new paths that are not accessible using purely experimental (neurobiological) methods.

Many functions (perception, attention, memory, language, ...) are studied in cognitive neuroscience using different methods, including neuroimaging, electrophysiology, behavioral testing, and computational modeling (e.g., brain encoding and decoding). Brain encoding is the process of learning the mapping from the stimuli representation to the neural brain activation. Recent brain encoding studies highlight the potential for natural language

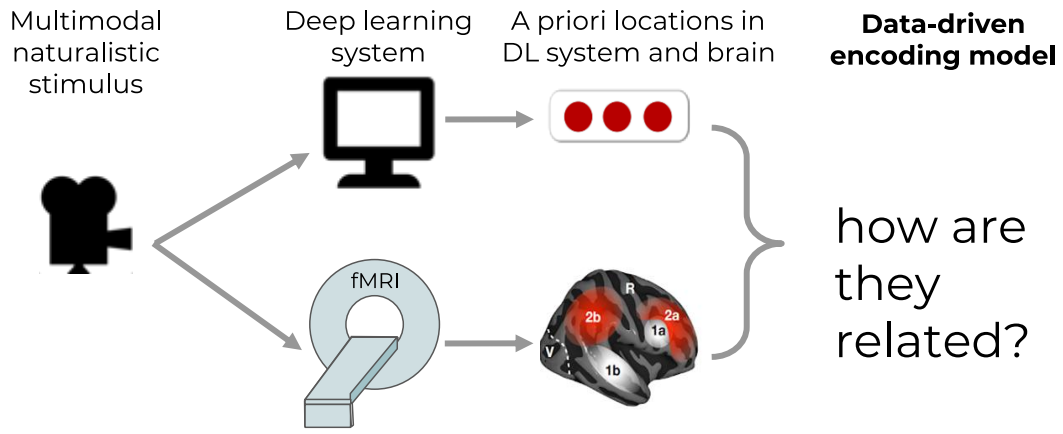



Figure 1.1: Alignment between deep learning systems and human brains. *This Figure is adapted from Toneva and Wehbe [2019].*

models to improve our understanding of language processing. Simultaneously, naturalistic fMRI/MEG/EEG datasets are becoming increasingly available and present even further avenues for understanding the alignment between brains and models. However, with the many available models and datasets, it can be challenging to know what aspects of the models are essential for studying language comprehension in brain alignment. Thus, we need language models using biologically plausible mechanisms from which plausible representations can emerge better to understand the insights of the brain in language processing.

In this thesis, we introduce a data-driven framework that addresses these limitations by establishing a direct link between information processing in the human brain, as measured by techniques such as fMRI (functional Magnetic Resonance Imaging) and MEG (Magnetoencephalography), and the functioning of natural language processing (NLP) computer systems. Specifically, we explore the information processing in the brain as a multifaceted and dynamic process that can be distilled into three key components: "what," "where," and "when." We leverage these three components to understand how the brain processes language-related information, as observed through fMRI (i.e., where the activity is located) and MEG (i.e., when the brain processes words and their semantics) measurements. In conclusion, our research aims to establish strong evidence between information processing in the human brain and NLP systems and strives to pinpoint the "where" and "when" aspects of this intricate interplay. By doing so, we contribute to neurolinguistics and NLP and pave the way for a deeper understanding of the dynamic mechanisms underpinning language comprehension and communication.

## 1.1 NLP Models for Human Language Comprehension

When reading or listening to the sentence "The trophy does not fit into the brown suitcase because it is too big", despite the unclear reference of "it" to either the trophy or suitcase, we intuitively understand it refers to the trophy [Levesque et al., 2012]. Conversely, if the

sentence were, "The trophy does not fit into the brown suitcase because it is too small", we would deduce that "it" signifies the suitcase. This raises questions about how our brains interpret these sentences and assign them real-world meanings. To explore this, we must first address basic inquiries regarding processing information in the brain - specifically, where and when this occurs and how the brain combines this information from various places and moments.

Through the use of neuroimaging tools that monitor brain activity during language comprehension, neuroscientists have advanced our understanding of the 'what', 'where', and 'when' aspects of this process. Research indicates that the meaning of individual words is represented across the cortex in a way that is broadly similar among different individuals (as shown in studies by [Mitchell et al., 2008, Wehbe et al., 2014, Huth et al., 2016]. Additionally, a specific group of brain regions, known as the "language network", has been identified as crucial for understanding language [Fedorenko et al., 2020]. The timing of word processing has also been pinpointed, with evidence suggesting that the meaning of a word begins to be processed between 200 and 600 milliseconds after the word is read [Salmelin, 2007]. Despite these advancements, the mechanisms by which the brain integrates information from various areas and across different time intervals during language comprehension remain a mystery.

## 1.2 Thesis Statement and Outline

This thesis is centered around the following statement: Bridging the gap between current deep neural network models and language comprehension in the brain: 1) our mechanistic understanding of language comprehension in the brain through long-term memory plausibility of language models (i.e. how they deal with long-term dependencies and contextual information), (2) a deeper understanding of language processing in the brain by the interpretation of hemodynamic response functions through computational modeling, and 3) the word and semantic-level performance of natural language processing models through the transfer of insight from the brain.

**Chapter 3:** Summarize the latest efforts in how deep neural networks begin to solve computational problems (encoding and decoding) across various modalities (language, vision, and speech) and thereby illuminate the computations that the brain accomplishes effortlessly. Specifically, we discuss popular representations of language, vision, and speech stimuli derived from static word embeddings, sequence, transformer-based, and multi-modal transformer-based models. We then present a summary of the naturalistic brain datasets and famous evaluation metrics, review the popular deep learning-based encoding and decoding architectures, and note their benefits and limitations in the context of brain alignment.

A long review paper is submitted to Transactions on Machine Learning Research (TMLR) journal, currently it is under review.

**Chapter 4:** Summarize the effectiveness of various language models, including Long Short-Term Memory Networks (LSTMs), ELMo, and Longformer, in predicting brain activity when subjects listen to narrative stories. While these models have succeeded in text-driven brain activation prediction, they cannot still handle long-term dependencies and pro-

vide insights into the neural mechanisms at play. Our findings suggest that LSTM cell states align better with brain recordings than LSTM hidden states, indicating their ability to capture long-term information. Additionally, ELMo and Longformer representations predict brain activity across different brain language regions.

A long paper was previously published at $44^{th}$ Annual Meeting of the Cognitive Science Society conference (July 2022, Toronto, Canada).

**Chapter 5:** Investigates how different hemodynamic response function delays in the brain's response function affect the alignment between language model representations and brain recordings obtained. At the same time, participants listened to or read a story. We explore the relative importance of syntactic information (i.e., syntactic embeddings based on constituency trees) versus semantic information using open-source language models such as BERT, GPT-2, and Llama-2. Further, we examine various context lengths and reveal differences in how the brain processes language across different delays.

**Chapter 6:** Summarize the use of Magnetoencephalography (MEG), with higher temporal resolution than fMRI, enables us to look more precisely at the timing of linguistic feature processing. Inspired by previous fMRI encoding studies, we study MEG brain encoding using basic syntactic and semantic features, with various context lengths and directions (past vs. future), for a dataset of 8 subjects listening to stories. We showed that Bidirectional Transformer (BERT) model, contrary to other features, lead to a significant prediction in MEG brain activity across auditory and language regions between 50-550ms (250ms to 750ms with word onset at 200ms).

A long paper was previously published at $24^{th}$ INTERSPEECH conference (August 2023, Dublin, Ireland).

**Chapter 7:** Summarize how children learn the language and how their brains process it, applying insights to machine learning and robotics. We mainly focus on cross-situational learning (CSL) using complete sentences to understand early language development. The research compares different models and word representations, finding that fine-tuned BERT representations perform best. These models can aid human-robot interaction and enhance our understanding of language acquisition in children.

An initial report was presented at the Splu-RoboNLP workshop at ACL in July 2021, later extended into a journal article. It is now undergoing minor revisions for publication in the Nature Scientific Reports journal.

## 1.3 Summary of Contributions

The contributions of each chapter in this dissertation can be summarized as follows:

**Chapter 3:**

- We discuss popular representations of language, vision, and speech stimuli derived from static word embeddings, sequence, transformer, and multi-modal transformer-based models.
- We present a summary of the naturalistic brain datasets such as Moth-radio-hour, Narratives, Little Prince, Harry Potter, Natural Scenes Dataset, Things, Short Movie

Clips, etc,. Further, famous evaluation metrics such as 2V2 accuracy, Pearson Correlation, and normalized predictivity are discussed.

- We review popular deep learning-based encoding and decoding architectures and note their benefits and limitations in the context of brain alignment. We summarize various encoding models in the form of a taxonomic survey tree. Similarly, we synthesize the literature related to decoding models into a survey tree and compare the vision, auditory, and linguistic stimuli reconstructed using various decoding models. Finally, we conclude with a summary and discussion about future trends.

**Chapter 4:**

- Given that a language model pre-trained on corpora by handling long-term dependencies, we propose the problem of finding which of these are the most predictive of fMRI brain activity for listening tasks.

- The investigation of the long-term context of language model results reveals that ELMo and Longformer representations display better brain alignment during narrative story listening. The layer-wise encoding performance results across brain language ROIs reveal that the intermediate layers have better brain alignment.

- We also find that the internal memory representations of LSTM (cell state and hidden state) derive interesting insights that the cell state representations yield better performance than hidden state representations.

**Chapter 5:**

- We examine how the intricate processing of diverse language regions at varying HRF delays in the human brain corresponds with transformer-based language models. For various HRF delays and context lengths, we analyze the impact on the alignment between brain recordings and language model representations.

- Further, We explore the relative importance of syntactic information (i.e.syntactic embeddings based on constituency trees) versus semantic information, using open-source language models. Using different HRF delays, we find that bilateral temporal lobes and frontal regions process syntactic information at early delays, while the angular gyrus processes semantic information in higher HRF delays.

- We further investigate different context lengths and find that more extended context may significantly increase HRF delays.

**Chapter 6:**

- We study how the brain encodes semantic and fine-grained syntactic features of words using MEG recordings. We address some critical questions: (1) How much context is maintained through time to process words? (2) Is the direction of context important (past context vs. future context)?

- We explore (a) basic syntactic features, (b) GloVe embeddings, and (c) semantic BERT embeddings for MEG brain encoding. We find that BERT representations predict MEG significantly but not other syntactic features or word embeddings (e.g., GloVe).

- We report that past context has greater predictive power than future context. $R^2$ scores are proportional to context length when dealing with past context.

**Chapter 7:**

- Grounded language acquisition is the process of learning a language - how infants can learn language by observing their environments, interacting with others, and understanding the concepts of a language as it relates to the world. We take the language acquisition perspective to machine learning and robotics, where part of the problem is understanding how language models can perform grounded language acquisition through noisy supervision and discussing how they can account for brain learning dynamics.

- Our experimental results demonstrate that fine-tuned BERT representations are more efficient and better at capturing the complex relations between words than other word representations.

- We find that biologically plausible ESNs have a better trade-off on all three grounded language datasets with better prediction error and low latency.

## 1.4 Additional Work

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.

**Joint processing of linguistic properties in brains and language models** Language models effectively predict brain recordings of subjects experiencing complex language stimuli. For a deeper understanding of this alignment, it is essential to understand the correspondence between the human brain's detailed processing of linguistic information versus language models. We investigate this correspondence via a direct approach, eliminating information related to specific linguistic properties in the language model representations and observing how this intervention affects the alignment with fMRI brain recordings obtained while participants listened to a story. We investigate a range of linguistic properties (surface, syntactic, and semantic) and find that eliminating each one significantly decreases brain alignment. Specifically, we find that syntactic properties (i.e., Top Constituents and Tree Depth) have the most significant effect on the trend of brain alignment across model layers. These findings provide clear evidence for the role of specific linguistic information in the alignment between brain and language models and open new avenues for mapping the joint information processing in both systems.

**Subba Reddy Oota**, Manish Gupta, and Mariya Toneva. "Joint processing of linguistic properties in brains and language models". In: Advances in Neural Information Processing Systems. NeurIPS-2023.

**How does the brain process syntactic structure while listening?** Syntactic parsing is the task of assigning a syntactic structure to a sentence. There are two popular syntactic parsing methods: constituency and dependency parsing. Recent works have used syntactic embeddings based on constituency trees, incremental top-down parsing, and other word syntactic features for brain activity prediction given the text stimuli to study how the syntax structure is represented in the brain's language network. However, the effectiveness of dependency parse trees or the relative predictive power of the various syntax parsers across brain areas, especially for the listening task, still needs to be explored. In this study, we investigate the predictive power of the brain encoding models in three settings: (i) individual performance of the constituency and dependency syntactic parsing based embedding methods, (ii) efficacy of these syntactic parsing based embedding methods when controlling for essential syntactic signals, (iii) relative effectiveness of each of the syntactic embedding methods when controlling for the other. Further, we explore the relative importance of syntactic information (from these syntactic embedding methods) versus semantic information using BERT embeddings. We find that constituency parsers help explain temporal lobe and middle-frontal gyrus activations. In contrast, dependency parsers better encode syntactic structure in the angular gyrus and posterior cingulate cortex. Although semantic signals from BERT are more effective than any of the syntactic features or embedding methods, syntactic embedding methods explain additional variance for a few brain regions.

**Subba Reddy Oota**, Mounika Marreddy, Manish Gupta, and Raju S. Bapi. "How does the brain process syntactic structure while listening?" In: ACL. Findings 2023.

**What aspects of NLP models and brain datasets affect brain-NLP alignment?** Recent brain encoding studies highlight the potential for natural language processing models to improve our understanding of language processing in the brain. Recent brain encoding studies highlight the potential for natural language processing models to improve our understanding of language processing in the brain. Simultaneously, naturalistic fMRI datasets are becoming increasingly available and present further avenues for understanding the alignment between brains and models. However, with the many available models and datasets, it can be challenging to know what aspects of them are essential to consider. In this work, we systematically study the brain alignment across five naturalistic fMRI datasets, two stimulus modalities (reading vs. listening), and different Transformer text and speech models. All text-based language models are significantly better at predicting brain responses than all speech models for both modalities. Further, bidirectional language models better predict fMRI responses and generalize across datasets and modalities.

**Subba Reddy Oota**, and Mariya Toneva. "What aspects of NLP models and brain datasets affect brain-NLP alignment?" In: CCN. 2023.

**Neural Language Taskonomy: Which NLP Tasks are the most Predictive of fMRI Brain Activity?**

Several popular Transformer based language models are successful for text-driven brain encoding. However, existing literature leverages only pre-trained text Transformer models and has yet to explore the efficacy of task-specific learned Transformer representations. In

this work, we explore transfer learning from representations learned for ten popular natural language processing tasks (two syntactic and eight semantic) for predicting brain responses from two diverse datasets: Pereira (subjects reading sentences from paragraphs) and Narratives (subjects listening to the spoken stories). Encoding models based on task features predict activity in different regions across the whole brain. Features from coreference resolution, NER, and shallow syntax parsing explain more significant variance for the reading activity. On the other hand, tasks such as paraphrase generation, summarization, and natural language inference for the listening activity show better encoding performance. Experiments across all 10 task representations provide the following cognitive insights: (i) language left hemisphere has higher predictive brain activity versus language right hemisphere, (ii) posterior medial cortex, temporo-parieto-occipital junction, dorsal frontal lobe have higher correlation versus early auditory and auditory association cortex, (iii) syntactic and semantic tasks display an excellent predictive performance across brain regions for reading and listening stimuli resp.

**Subba Reddy Oota**, Jashn Arora, Veeral Agarwal, Manish Gupta, and Raju S. Bapi. "Neural Language Taskonomy: Which NLP Tasks are the most Predictive of fMRI Brain Activity?" In: NAACL. 2022.

**Speech language models lack important brain-relevant semantics** Despite known differences between reading and listening in the brain, recent work has shown that text-based language models predict both text-evoked and speech-evoked brain activity to an impressive degree. This poses the question of what types of information language models capture that are correlated with features truly predicted in the brain. We investigate this question via a direct approach, in which we eliminate information related to specific low-level stimulus features (textual, speech, and visual) in the language model representations and observe how this intervention affects the alignment with fMRI brain recordings acquired while participants read versus listened to the same naturalistic stories. We further contrast our findings with speech-based language models, which would be expected to predict speech-evoked brain activity better, provided they model language processing in the brain well. Using our direct approach, we find that text-based and speech-based language models align well with early sensory areas due to shared low-level features. Text-based models continue to align well with later language regions even after removing these features, while, surprisingly, speech-based models lose most of their alignment. These findings suggest that speech-based models can be further improved to reflect brain-like language processing better.

**Subba Reddy Oota**, Emin Celik, Fatma Deniz, and Mariya Toneva. "Speech language models lack important brain-relevant semantics". Arxiv Preprint

# 2 BackGround and Related Work

The increasing availability of naturalistic fMRI datasets and large-scale neural models can enable a better understanding of the brain's response to natural stimuli. Just in the last few years, researchers have shown that brain responses of people comprehending language can be predicted well by text-based language models [Wehbe et al., 2014, Jain and Huth, 2018, Toneva and Wehbe, 2019, Caucheteux and King, 2020, Schrimpf et al., 2021b]. Understanding the processes that are involved in language comprehension has interested many philosophers, linguists, psycholinguists, neurolinguists, and computer scientists. In this thesis, we aim to bridge the empirical methodologies for understanding language comprehension in the brain with language models designed to process language. This chapter summarizes the relevant neurolinguistic findings about language in the brain and discusses the classical and modern computational methods designed to process language.

## 2.1 Language in the Brain

The invention of non-invasive imaging modalities that 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 give details about the most popular non-invasive imaging modalities and summarize the previous findings about language in the brain that these brain imaging modalities have enabled.

### 2.1.1 Sampling Language in the Brain via Brain Imaging Recordings

The most common non-invasive brain imaging modalities are Electroencephalography (EEG), Magnetoencephalography (MEG) and functional Magnetic Resonance Imaging (fMRI). These modalities present distinct advantages and drawbacks in capturing brain activity, which we will provide a concise overview of below. Within the context of this thesis, we leverage brain recordings obtained from two specific brain imaging modalities, fMRI and MEG, which complement each other's strengths.

Functional Magnetic Resonance Imaging (fMRI) fMRI measures the Blood Oxygen Level-Dependent (BOLD) signal, which refers to the change in oxygen levels in the blood. When neurons in the brain are active, they consume more oxygen, increasing blood flow to that specific region. A Magnetic Resonance Imaging device can detect this change in blood flow because oxygenated and deoxygenated blood have different magnetic properties, leading to a variation in magnetic signal. By tracking changes in blood flow,

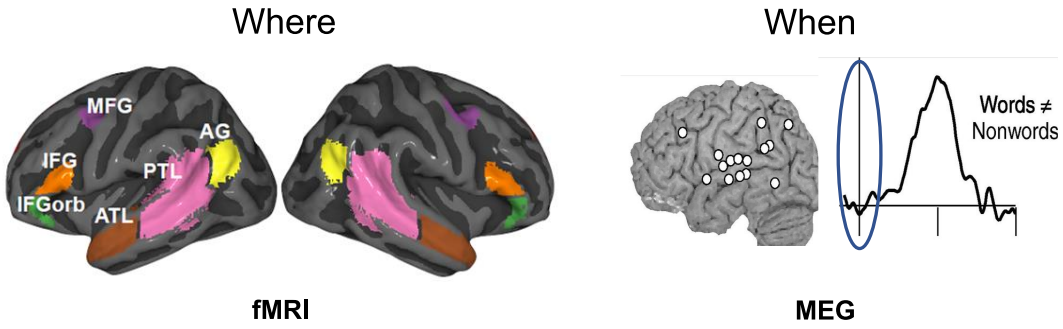



Figure 2.1: Non-invasive brain recordings: fMRI and MEG. This Figure is adapted from Toneva et al. [2022].

fMRI creates maps that show which parts of the brain are involved in specific tasks, such as sensory processing, motor function, language, visual, or memory. This change is called the hemodynamic response function (HRF), and it refers to the pattern of changes in blood flow, blood volume, and oxygenation that occurs in the brain in response to neuronal activity. The HRF describes the characteristic time course of this hemodynamic response – the rise, peak, and fall in blood oxygenation in a specific brain region following neural activity. For example, in the language comprehension task, once neurons in a brain area are active, the BOLD response takes about 12 seconds 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 mainly on the strength of the MRI magnet. A typical MRI with a 3T magnet results in a sample of the BOLD response in every 1 - 2mm $\times$ 1 - 2mm $\times$ 1 - 2mm volume pixel in the brain. A significant limitation of BOLD-based fMRI is that its measurements correspond to blood flow and not actual neuronal activity.

MAGNETOENCEPHALOGRAPHY (MEG) Magnetoencephalography (MEG) is another non-invasive, popular neuroimaging technique used to measure the magnetic fields produced by the electrical activity in the brain. When neurons in the brain are active, they generate electrical currents, and these currents produce small magnetic fields. MEG records these magnetic fields, providing information about the timing and location of brain activity. One of the critical strengths of MEG is its high temporal resolution. It can detect changes in brain activity on the order of milliseconds, allowing researchers to study the precise timing of neural events. MEG is often used in cognitive neuroscience and clinical applications to study various brain functions, such as sensory processing, motor control, language, and memory. It is beneficial in mapping brain activity associated with various cognitive tasks, providing valuable information about the dynamics of neural processes.

Neuroimaging techniques like fMRI (functional Magnetic Resonance Imaging) and MEG have superior temporal resolution, while fMRI typically has better spatial resolution.

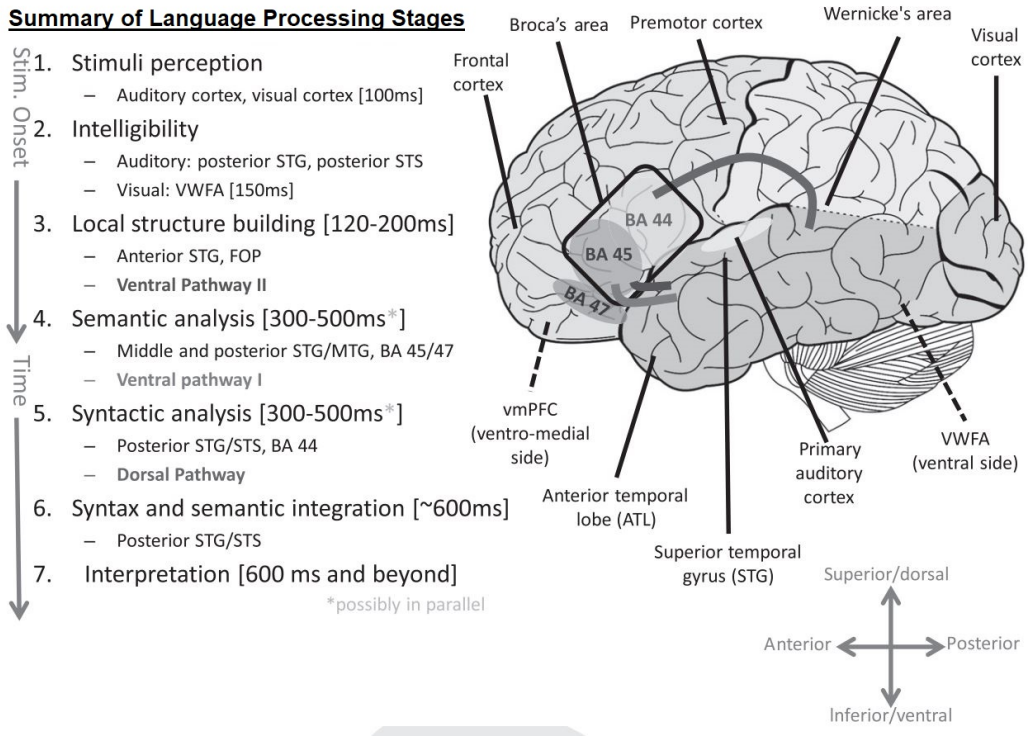



Figure 2.2: Summary of the stages of language processing in the brain. *This Figure is adapted from Friederici [2011].*

### 2.1.2 Individual Word Processing

Using high temporal resolution brain imaging recordings, researchers have started decoding the processing stages during language comprehension. These stages are outlined in Fig. 2.2. Upon encountering a word through reading or listening, the respective visual or auditory cortices are activated within 100 ms. Following this, at around 150 ms, the input undergoes further processing as sequences of letters or phonemes. This occurs in the visual word form area during reading or in the posterior superior temporal gyrus and sulcus (pSTG and pSTS) during listening. These areas show higher activity for language stimuli than other visual or auditory inputs [Salmelin, 2007, Friederici, 2011]. Studies comparing brain responses to actual and nonsensical words indicate that a word's meaning is processed between 200 and 600 milliseconds post-presentation, primarily involving the temporal cortex [Salmelin, 2007, Friederici, 2011]. The subsequent stages of processing, which are crucial for understanding sentences and longer language constructs, are further detailed in Chapter 2.1.3.

Word meaning Language processing theories identify the temporal lobe as key to the general retrieval and processing of words. However, these theories need more specifics on how the brain represents the meaning of particular words or categories of words, such as

tools, animals, and others. One hypothesis in neuroscience studies for representing concrete concepts (like "dog") in the brain is the Grounded Cognition Model, also known as the Embodied Cognition Model or the Simulation Model [Kemmerer, 2022]. This theory proposes that concrete concepts are represented through associated perceptual experiences [Barsalou, 1999, Barsalou et al., 2008]. Research based on this hypothesis indicates that the semantic attributes of concrete concepts are stored in a distributed yet organized way across the cortex. Specifically, a particular semantic attribute (for example, an auditory feature) is stored in the same cortical area responsible for related high-level sensory perception (like auditory perception). Empirical evidence supports this organization for semantic attributes linked to various senses, including color [Simmons et al., 2007], shape [Chao et al., 1999], motion [Damasio et al., 1996], and even olfaction and taste [Goldberg et al., 2006a,b]. Further backing for this hypothesis comes from computational models that predicted fMRI recordings based on the semantic properties of words [Mitchell et al., 2008]. This study found correlations between semantic properties and the functions of cortical regions where these properties predicted fMRI activity. For instance, the semantic property of the verb "push" significantly predicted activity in the motor cortex.

Integrating different semantic attributes into a higher-order representation is thought to be facilitated by the bilateral anterior temporal lobes (ATL) [Visser et al., 2010]. The ATL plays a crucial role in organizing these attributes, helping to differentiate between objects that fall within or outside the scope of a specific concept. This representation is then accessible to other brain areas for further processing.

Evidence supporting the ATL as a central hub for integrating semantic attributes comes from various studies. Clinical observations have shown a strong correlation between the progressive loss of object concept understanding and progressive atrophy in the ATL in patients [Bright et al., 2008]. Secondly, studies using repetitive Transcranial Magnetic Stimulation (rTMS), an invasive imaging technique, demonstrate that temporarily disrupting the ATL in healthy individuals impairs their ability to process object concepts [Pobric et al., 2007, Ralph et al., 2009]. These findings collectively highlight the ATL's pivotal role in semantic attribute integration.

### 2.1.3 Multi-word Composition

Comprehending a multi-word phrase involves more complex processing stages than those required for individual word processing [Bhattasali et al., 2019]. Theories from linguistics and cognitive psychology suggest that words are combined following various rule sets to create a composite meaning. This summary explores the different hypothesized types of composition and the evidence supporting these processes in the human brain.

Syntactic composition The grammar includes a set of rules for combining words. For instance, an adjective followed by a noun typically forms a noun phrase. However, syntactically correct combinations may not always make semantic sense, as in the example "colorless green ideas sleep furiously" [Chomsky, 1957].

One marker for syntactic and logico-semantic composition in the brain is the P600 response. This characteristic brain response is identified through electrophysiological record-

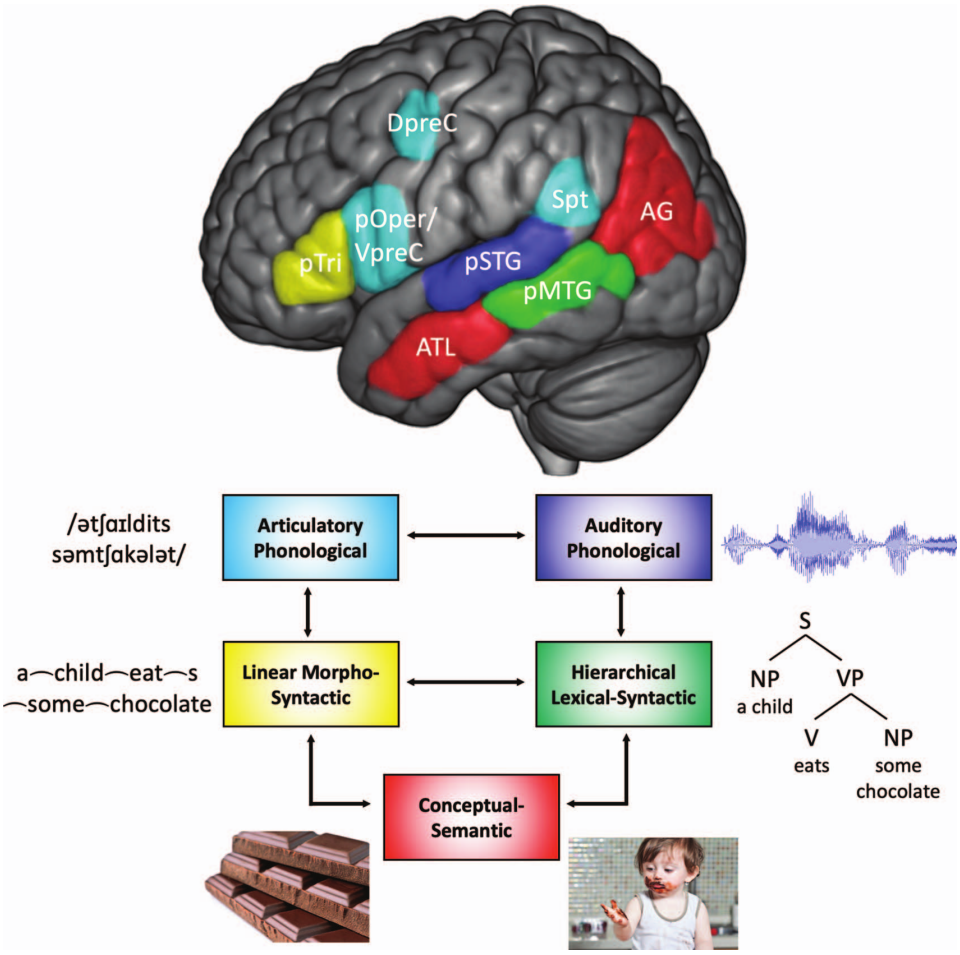



Figure 2.3: Cortical organization of syntax. *This Figure is adapted from Matchin and Hickok [2020].*

ings like EEG or MEG. The P600 manifests as a positive shift in the recorded signal, occurring between 500ms and 800ms after a stimulus is presented, peaking around 600ms. It is primarily associated with syntactic violations [Kemmerer, 2014, Coulson et al., 1998] and also with thematic role violations [Kutas et al., 2006, Kuperberg, 2007]. The generation of the P600 is believed to occur in the bilateral temporal lobes, in areas posterior to where the N400 response, which will be described subsequently, is generated [Service et al., 2007].

Fig.2.3 illustrates the neuroanatomy involved in the interface between the syntactic system and the phonological and semantic networks, as outlined by Matchin and Hickok [2020]. Auditory-based phonological features are processed in the pSTG region (Fig.2.3, indigo), while the representation of hierarchical syntactic information occurs in the cortical zone situated between the auditory-phonological and semantic zones. This cortical area corresponds to the pMTG (Fig.2.3, green), which is proposed as the hub for hierarchical lexical-syntactic processing.

LOGICO-SEMANTIC COMPOSITION Logico-semantic composition is another set of rules that focuses on forming predicate-argument structures [Pylkkänen, 2020]. Unlike syntactic rules, logico-semantic composition deals with the meanings and relationships between words, which are not always directly inferred from their syntactic arrangement [Partee and Borschev, 2003]. Pylkkänen [2020] demonstrates this difference using phrases with similar syntactic structures but different meanings: 'she liked my eye color' versus 'she guessed my eye color'. The disparity becomes clearer when 'eye color' is replaced with 'eyes'. While 'She liked my eyes' sounds correct, 'she guessed my eyes' does not. This inconsistency arises because verbs like 'guess' and 'like' carry different semantic requirements [Nathan, 2006]. 'Guess' typically needs a question as its argument. Nouns that denote relations, such as 'color', can be transformed into a question format, while nouns representing entities, like 'eyes', cannot. This illustrates how logico-semantic composition governs the compatibility and meaning of word combinations beyond their syntactic structure.

The N400, like the P600, is a prototypical brain response occurring between 200ms and 600ms following the presentation of a stimulus [Kutas and Federmeier, 2011]. It is associated with the semantic processing effort required to integrate a word into a specific context [Kutas et al., 2006]. The generation of the N400 is believed to primarily involve the temporal lobes on both sides of the brain, specifically the superior temporal gyrus (STG) and middle temporal gyrus (MTG). Research by Kutas and Federmeier [2011] and Kutas et al. [2006] has particularly highlighted these areas about the N400. Unlike the P600, the N400 is generally not influenced by syntax, as studies such as Allen et al. [2003] indicate. This response is more closely tied to the processing of meaning rather than language structure.

CONCEPTUAL COMPOSITION Syntactic and logico-semantic compositions do not encompass the conceptual content of combined words. This gap is addressed by a third type of composition: conceptual composition [Pylkkänen, 2020]. Conceptual composition has been a focus in cognitive psychology, particularly in adjective-noun and noun-noun compositions [Murphy, 1990, Smith and Osherson, 1984, Hampton, 2013]. The left anterior temporal lobe (LATL) is considered a key site for conceptual composition, with its activity extending beyond what can be explained by syntactic and logico-semantic compositions alone [Pylkkänen, 2020]. For instance, Pylkkänen [2020] found that no adjective-noun combination uniformly triggered LATL activation. Instead, it depended on the specific concepts described by the words and their meaning specificity [Westerlund and Pylkkänen, 2014, Zhang and Pylkkänen, 2015, Ziegler and Pylkkänen, 2016]. The underlying hypothesis is that conceptual composition in the LATL is driven by how integrating the first word modifies the feature space of the second word [Pylkkänen, 2020]. If the second word is general, its feature space is relatively sparse, allowing the first word to contribute significantly to the final composed meaning. Conversely, a particular first word can also add many features. From this, it follows that combinations with more specific first words and more general second words would produce pronounced conceptual composition effects in the LATL.

## 2.2 Language in AI Models

To understand brain language processing, older methods for *text-based stimulus representation* include text corpus co-occurrence counts, topic models, syntactic, and discourse features. Recently, semantic and experiential attribute models have been explored for text-based stimuli. Semantic representation models include distributed word embeddings, sentence representation models, recurrent neural networks (RNNs), and Transformer-based language models [Vaswani et al., 2017].

### 2.2.1 Natural Language Processing Systems

Over the past decade, neural networks have experienced a transformative evolution driven by the availability of larger datasets, increased computational power, and advanced optimization methods. In language processing, these advancements have enabled multi-layer neural networks to extract meaning from word sequences and execute a wide range of complex linguistic tasks. A pivotal development in Natural Language Processing (NLP) systems is their ability to learn statistics through a simple yet effective objective called language modeling. This language modeling objective, foundational even in early neural network research by Elman [1991], initially emerged within cognitive science before its integration into modern NLP frameworks. Language modeling involves training the system to predict the next word based on preceding context [Elman, 1991, Mikolov et al., 2013b, Graves and Graves, 2012]. Although it seems straightforward, language modeling has become a powerful technique in NLP, serving as a foundation for teaching networks a general understanding of language statistics during a pretraining phase. This pretraining phase is crucial as it typically precedes a secondary phase where the network is fine-tuned to perform specific tasks [Devlin et al., 2019, Radford et al., 2019]. The NLP system's parameters are refined to enhance task-specific performance in this fine-tuning stage. This dissertation utilizes publicly available NLP systems that have undergone extensive pretraining on large text corpora. We focus on four specific language models: ELMo [Peters et al., 2018], which is based on recurrent neural network architecture, BERT [Devlin et al., 2019], which employs encoder transformer-based architecture, GPT-2 [Radford et al., 2019], which employs decoder transformer-based architecture and Longformer [Beltagy et al., 2020], which employs encoder transformer-based architecture designed for longer context lengths. Further details about each system are provided in the subsequent sections.

### 2.2.2 Distributed Word Representations

Distributed word representations capture many precise syntactic and semantic word relationships in text classification problems.

- **Word2Vec Embeddings** Word2Vec model provides a non-deterministic way to determine the word representations [Mikolov et al., 2013b]. Here, the word "non-deterministic" primarily refers to the aspect of Word2vec training process where the initialization of weights and the sequence of training samples can affect the final word embeddings. This variability means that running the Word2Vec model multiple times

can lead to slightly different word embeddings each time, even if the training data remains the same. Further, It can learn similar word vectors for words in a similar context.

- **GloVe Embeddings** The input used in the GloVe model is a non-zero word-word co-occurrence matrix [Pennington et al., 2014], which adds the global context information by default, unlike the use of local context in Word2Vec [Mikolov et al., 2013b].

- **FastText Embeddings** Since the FastText model considers the bag of character n-grams to represent each word, it allows us to compute rare word representations [Joulin et al., 2017].

### 2.2.3 Pretrained Langauge Models

#### Embeddings from Language Models (ELMo)

ELMo, a recurrence-based NLP system, utilizes multiple layers of Long Short-Term Memory units (LSTMs) to process language [Peters et al., 2018]. In ELMo, for each word token t, an LSTM layer l generates a hidden representation $h_t^l$ using a series of update equations. These equations involve a combination of learned weights ($w_c$, $w_f$, $w_i$, $w_o$), biases ($b_c$, $b_f$, $b_i$, $b_o$), and gate states (forget gate $f_t$, output gate $o_t$, input gate $i_t$). The LSTM updates the cell state $c_t$ and hidden state $h_t^l$ as follows: For a word token t, an LSTM generates the corresponding hidden representation $h_t^l$ in layer l using the following update equations:

$$\tilde{c} = \tanh(w_c[h_{t-1}^l; h_t^{l-1}] + b_c)$$

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t$$

$$h_t^l = o_t \times \tanh(c_t)$$

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:

$$f_t = \sigma(w_f[h_{t-1}^l; h_t^{l-1}] + b_f)$$

$$i_t = \sigma(w_i[h_{t-1}^l; h_t^{l-1}] + b_i)$$

$$o_t = \sigma(w_o[h_{t-1}^l; h_t^{l-1}] + b_o),$$

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.

ELMo's architecture combines the internal representations from two independent LSTMs for each word token: a forward LSTM that processes previous words and a backward LSTM that considers future words. This dissertation uses a pre-trained version of ELMo with two hidden LSTM layers, as provided by Gardner et al. [2018]. Each independent LSTM (forward and backward) in this version has a 512-dimensional hidden state, and the system has

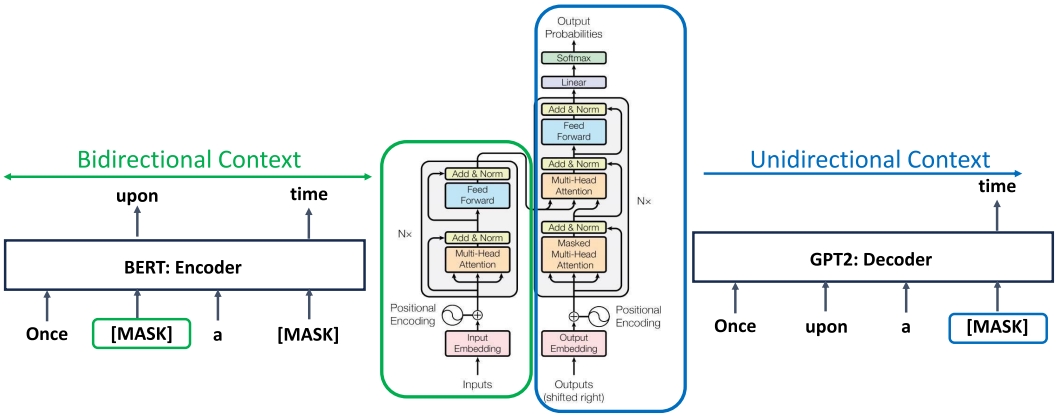



Figure 2.4: Transformer model architecture and its variants: BERT [Devlin et al., 2019] and GPT2 [Radford et al., 2019]. *The Transformer model architecture is adapted from Vaswani et al. [2017].*

13.6 million parameters. ELMo was trained on the One Billion Word Benchmark [Chelba et al., 2014], a dataset comprising approximately 800 million tokens of news crawl data from WMT 2011.

### Representations from Transformers

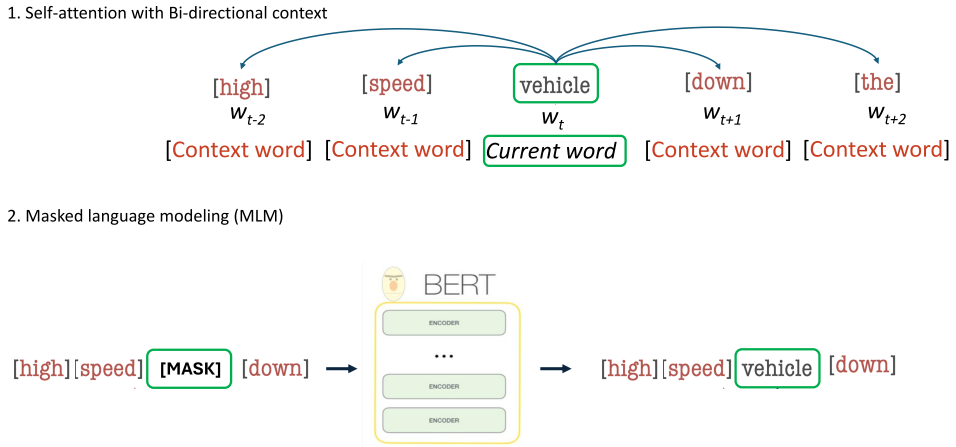



Figure 2.5: BERT model workflow. (1) Self-attention with Bi-direction context. (2) Word prediction with MLM modeling.

Transformer [Vaswani et al., 2017] is a prominent deep learning model that has been widely adopted in various fields, such as natural language processing (NLP), computer vision (CV), and speech processing. Transformer was initially proposed as a sequence-to-sequence model [Vaswani et al., 2017] for machine translation. Later works show that

Transformer-based pre-trained models (PTMs) can achieve state-of-the-art performances on various tasks.

BIDIRECTIONAL ENCODER REPRESENTATIONS FOR TRANSFORMERS (BERT) BERT is a pre-trained language model [Devlin et al., 2019] that provides bi-directional contextual information, while earlier methods have uni-directional context, as shown in Fig. 2.4. The bidirectional context means it considers both left and right-context words when encoding a word, as shown in Fig. 2.5. BERT leverages a multi-layer transformer architecture and the masked language modeling (MLM) objective to learn contextual representations of words in a bidirectional manner, making it highly effective for various NLP tasks. During training, a portion of the input tokens is randomly masked, and the model is trained to predict these masked tokens. BERT consists of multiple layers of transformer encoders. The following components can represent a single transformer layer: (1) Self-Attention Mechanism, (2) Multi-Head Attention, (3) Position-wise Feed-Forward Network, and (4) Layer Normalization. The BERT-base-uncased model consists of 12 transformer blocks, 768 hidden dimensions, 12 self-attention blocks, and 110 million parameters in total. On the other hand, the BERT-large-uncased model consists of 24 transformer blocks, 1024 hidden dimensions, and 16 self-attention blocks.

GENERALIZED PRETRAINED TRANSFORMER (GPT-2) GPT-2 is a unidirectional transformer model [Radford et al., 2019] designed for generative tasks, such as text generation. It is autoregressive, meaning it generates text one token at a time, conditioning on the previously generated tokens, as shown in Fig. 2.6. Similar to BERT, GPT-2 is a pre-trained model, which means it was initially trained on a large corpus of text data before being fine-tuned for specific tasks. This pre-training allows the model to understand and generate human-like text. GPT-2 comes in different sizes, with the most significant variant having 1.5 billion parameters. The model's size (regarding the number of parameters) directly correlates with its ability to understand and generate more complex text. Recently, OpenAI released more advanced models like GPT-3, GPT-3.5, and GPT-4, which further improved natural language understanding and generation capabilities. GPT-3+ models are much larger than GPT-2 (with 175 billion parameters in GPT-3) and are capable of even more sophisticated tasks.

LONGFORMER Transformer models like BERT [Devlin et al., 2019], RoBERTa [Liu et al., 2019], or GPT2 [Radford et al., 2019] are typically pretrained to process up to 512 tokens. This is problematic because real-world data can be arbitrarily long. As such, different models and strategies have been proposed to process longer sequences. One such model is Longformer [Beltagy et al., 2020], designed to process longer input sequences based on efficient self-attention that scales linearly with the length of the input sequence. Longformer also truncates the input, but it has the capacity to process up to 4,096 tokens rather than 512 tokens as in BERT or GPT2. Additionally, there are other Transformer models like BigBird and Longformer Encoder Decoder (LED) that can process even longer sequences beyond 4,096 tokens.

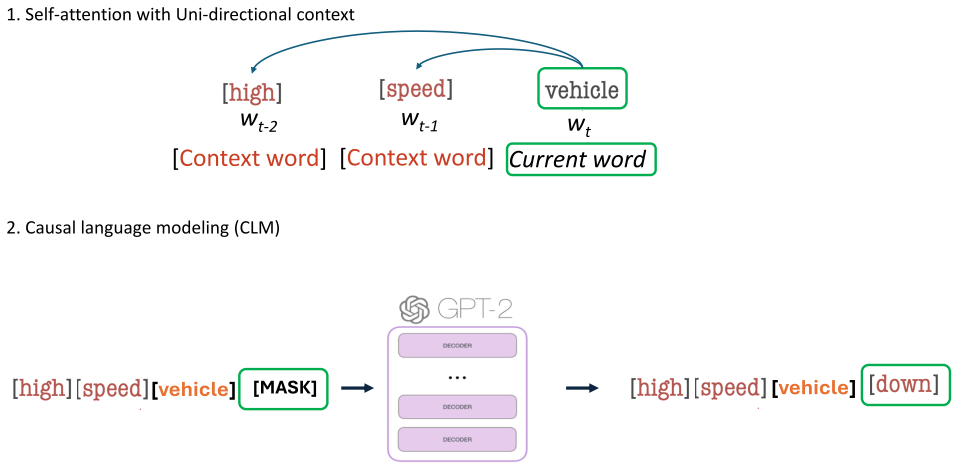



Figure 2.6: GPT-2 model workflow. (1) Masked Self-attention with Uni-direction context. (2) Word prediction with CLM modeling.

### 2.2.4 Linguistic Properties Captured by NLP Systems

Probing tasks [Adi et al., 2017, Hupkes et al., 2018, Jawahar et al., 2019, Mohebbi et al., 2021] help unpack the linguistic features possibly encoded in neural language models. These probing tasks are formulated as prediction tasks and focus on several aspects of sentence structure. To understand the degree to which various English language structures are encoded in Transformer-based encoders (BERT), Jawahar et al. [2019] focused on these probing tasks including surface, syntactic and semantic, and shown that BERT captures a rich hierarchy of linguistic information, with early layers encode surface information, intermediate layers encode syntactic information and higher layers encode semantic information, as shown in Fig. 2.7. Surface tasks probe for sentence length (SL) and word content (WC) for the presence of words in the sentence. Syntactic tasks test for sensitivity to word order (BShift) and the depth of the syntactic tree (TreeDepth). Semantic tasks check for the subject (respectively direct object) number in the main clause (SubjNum, respectively ObjNum).

In a similar study, Tenney et al. [2019] employed the edge probing tasks such as part-of-speech, constituents, dependencies, named entities, semantic roles, coreference, semantic proto-roles, and relation classification, defined by Tenney et al. [2018] to show the hierarchy of encoded knowledge through BERT layers, as shown in Fig. 2.8. Moreover, they observed that while most of the syntactic information can be localized in a few layers, semantic knowledge tends to spread across the entire network. Both studies were aimed at discovering the extent of linguistic information encoded across different layers.

| | Surface | | Syntactic | | | Semantic | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Layer** | **SentLen** (Surface) | **WC** (Surface) | **TreeDepth** (Syntactic) | **TopConst** (Syntactic) | **BShift** (Syntactic) | **Tense** (Semantic) | **SubjNum** (Semantic) | **ObjNum** (Semantic) | **SOMO** (Semantic) | **CoordInv** (Semantic) |
| 1 | 93.9 (2.0) | 24.9 (24.8) | 35.9 (6.1) | 63.6 (9.0) | 50.3 (0.3) | 82.2 (18.4) | 77.6 (10.2) | 76.7 (26.3) | 49.9 (-0.1) | 53.9 (3.9) |
| 2 | 95.9 (3.4) | 65.0 (64.8) | 40.6 (11.3) | 71.3 (16.1) | 55.8 (5.8) | 85.9 (23.5) | 82.5 (15.3) | 80.6 (17.1) | 53.8 (4.4) | 58.5 (8.5) |
| 3 | **96.2 (3.9)** | 66.5 (66.0) | 39.7 (10.4) | 71.5 (18.5) | 64.9 (14.9) | 86.6 (23.8) | 82.0 (14.6) | 80.3 (16.6) | 55.8 (5.9) | 59.3 (9.3) |
| 4 | 94.2 (2.3) | **69.8 (69.6)** | 39.4 (10.8) | 71.3 (18.3) | 74.4 (24.5) | 87.6 (25.2) | 81.9 (15.0) | 81.4 (19.1) | 59.0 (8.5) | 58.1 (8.1) |
| 5 | 92.0 (0.5) | 69.2 (69.0) | 40.6 (11.8) | 81.3 (30.8) | 81.4 (31.4) | 89.5 (26.7) | 85.8 (19.4) | 81.2 (18.6) | 60.2 (10.3) | 64.1 (14.1) |
| 6 | 88.4 (-3.0) | 63.5 (63.4) | **41.3 (13.0)** | 83.3 (36.6) | 82.9 (32.9) | 89.8 (27.6) | **88.1 (21.9)** | 82.0 (20.1) | 60.7 (10.2) | 71.1 (21.2) |
| 7 | 83.7 (-7.7) | 56.9 (56.7) | 40.1 (12.0) | **84.1 (39.5)** | 83.0 (32.9) | 89.9 (27.5) | 87.4 (22.2) | **82.2 (21.1)** | 61.6 (11.7) | 74.8 (24.9) |
| 8 | 82.9 (-8.1) | 51.1 (51.0) | 39.2 (10.3) | 84.0 (39.5) | 83.9 (33.9) | 89.9 (27.6) | 87.5 (22.2) | 81.2 (19.7) | 62.1 (12.2) | 76.4 (26.4) |
| 9 | 80.1 (-11.1) | 47.9 (47.8) | 38.5 (10.8) | 83.1 (39.8) | **87.0 (37.1)** | **90.0 (28.0)** | 87.6 (22.9) | 81.8 (20.5) | 63.4 (13.4) | **78.7 (28.9)** |
| 10 | 77.0 (-14.0) | 43.4 (43.2) | 38.1 (9.9) | 81.7 (39.8) | 86.7 (36.7) | 89.7 (27.6) | 87.1 (22.6) | 80.5 (19.9) | 63.3 (12.7) | 78.4 (28.1) |
| 11 | 73.9 (-17.0) | 42.8 (42.7) | 36.3 (7.9) | 80.3 (39.1) | 86.8 (36.8) | 89.9 (27.8) | 85.7 (21.9) | 78.9 (18.6) | 64.4 (14.5) | 77.6 (27.9) |
| 12 | 69.5 (-21.4) | 49.1 (49.0) | 34.7 (6.9) | 76.5 (37.2) | 86.4 (36.4) | 89.5 (27.7) | 84.0 (20.2) | 78.7 (18.4) | **65.2 (15.3)** | 74.9 (25.4) |

Figure 2.7: BERT composes a hierarchy of linguistic signals ranging from surface to semantic features [Jawahar et al., 2019]. *This Table is adapted from Jawahar et al. [2019].*

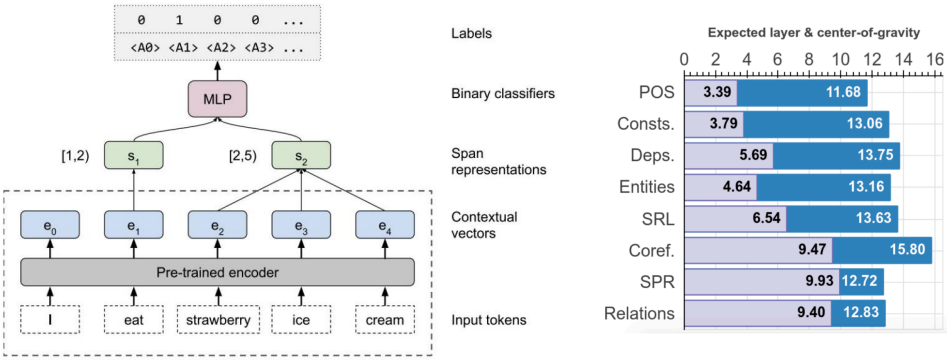



Figure 2.8: Edge Probing model architecture [Tenney et al., 2019]. Local syntax (word-level) captured at initial-middle layers and High-level semantics captured at later layers. *This Figure is adapted from Tenney et al. [2019].*

# 3 Deep Neural Networks and Brain Alignment (Review)

Can we obtain insights about the brain using AI models? How is the information in deep learning models related to brain recordings? Can we improve AI models with the help of brain recordings? Such questions can be tackled by studying brain recordings like functional magnetic resonance imaging (fMRI). As a first step, the neuroscience community has contributed several large cognitive neuroscience datasets related to passive reading/listening/viewing of concept words, narratives, pictures and movies. Encoding and decoding models using these datasets have also been proposed in the past two decades. These models serve as additional tools for basic research in cognitive science and neuroscience. Encoding models aim at generating fMRI brain representations given a stimulus automatically. They have several practical applications in evaluating and diagnosing neurological conditions and thus may also help design therapies for brain damage. Decoding models solve the inverse problem of reconstructing the stimuli given the fMRI. They are useful for designing brain-machine or brain-computer interfaces. Inspired by the effectiveness of deep learning models for natural language processing, computer vision, and speech, several neural encoding and decoding models have been recently proposed. In this survey, we will first discuss popular representations of language, vision and speech stimuli, and present a summary of neuroscience datasets. Further, we will review popular deep learning based encoding and decoding architectures and note their benefits and limitations. Finally, we will conclude with a brief summary and discussion about future trends. Given the large amount of recently published work in the computational cognitive neuroscience (CCN) community, we believe that this survey enables an entry point for DNN researchers to diversify into CCN research.

This chapter has been finalized based on our submission to the Transactions on Machine Learning Research (TMLR) journal, which is currently under review [Oota et al., 2023b].

*Subba Reddy Oota, Manish Gupta, Bapi Raju Surampudi, Gael Jobard, Mariya Toneva, Frederic Alexandre, Xavier Hinaut, Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey)*

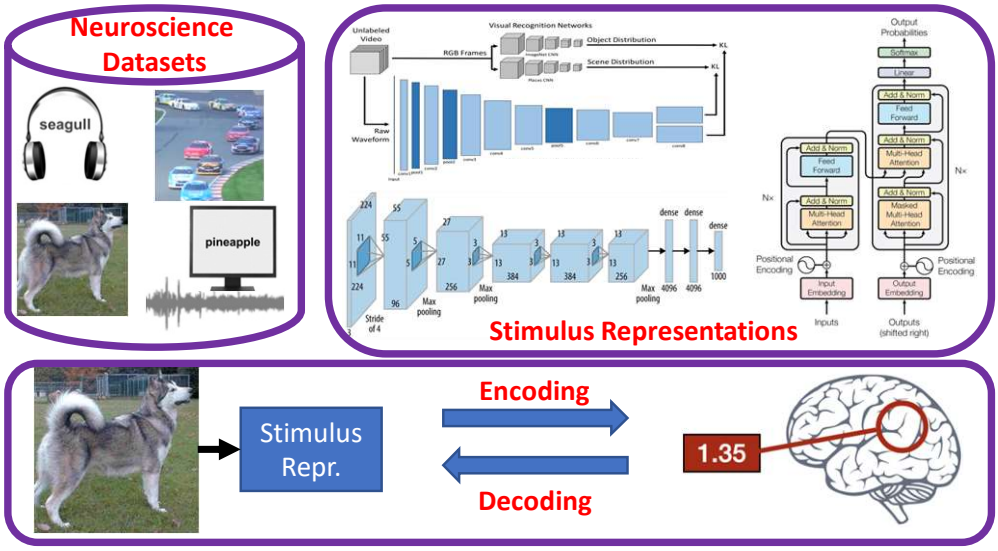



Figure 3.1: Brain Encoding and Decoding: Datasets & Stimulus Representations. *In this Figure, the encoding and decoding sub Figure is adapted from Ivanova et al. [2022].*

## 3.1 Introduction

The central aim of neuroscience is to unravel how the brain represents information and processes it to carry out various tasks (visual, linguistic, auditory, etc.). Two important models related to how brain represents information are, how external stimuli are represented in the form of neural responses (*the encoding model*) and how stimuli are recovered or reconstructed from the neuronal responses (*the decoding model*). The recent progress in deep neural networks in processing visual, auditory, linguistics, and multimodal stimuli makes one wonder if we could investigate these computational models and shed light on how the brain solves these problems. Thus, deep neural networks (DNN) may offer a computational medium to capture the unprecedented complexity and richness of brain activity, leading to accurate encoding and decoding solutions. Previous surveys, Cao et al. [2021] and Karamolegkou et al. [2023], have primarily focused on brain encoding and decoding studies for language stimuli. But recent attempts in cognitive neuroscience have focused on naturalistic and multimodal stimuli using DNNs. Hence, this survey systematically summarizes the latest encoding and decoding efforts on (i) how DNNs have begun to explain the underlying information processing in the brain for naturalistic stimuli of various modalities, (ii) the ways in which DNN models may be improved using the brain data, and (iii) the exploration of the shared underlying characteristics of both the systems.

The survey aims at introducing the problems in Computational Cognitive Neuroscience (CCN) problems to AI researchers familiar with recent advances in deep neural networks (DNNs). Therefore, in this survey we do not delve into architectural details and the learning procedures for DNNs but highlight how the advances in DNNs are used for addressing CCN

problems. This enables an entry point for DNN researchers to diversify into CCN research. The key takeaways of the survey are

1. Clear exposition of various opensource ecological stimuli datasets available and a curated GitHub repository for quick start of a study
2. An accessible taxonomy of models and approaches
3. A collection of open research problems in this fast-breaking research domain

**Brain encoding and decoding**: Two main tools studied in cognitive neuroscience are brain encoding and brain decoding, as shown in Figure 3.1. Encoding is learning the mapping $e$ from the stimuli $S$ to the neural activation $F$. The mapping can be learned using features engineering or deep neural networks. On the other hand, decoding constitutes learning mapping $d$, which predicts stimuli $S$ back from the brain activation $F$. However, in most cases, brain decoding aims to predict a stimulus representation $R$ rather than reconstructing $S$. In both cases, the first step is to learn a semantic representation $R$ of the stimuli $S$ at the train time. Next, a regression function $e : R \rightarrow F$ is trained for encoding. For decoding, a function $d : F \rightarrow R$ is trained. These functions $d$ and $e$ can then be used at test time to process new stimuli and brain activations. Ridge regression is the most popular choice for the functions $d$ and $e$.

To study the brain response to various modalities of stimuli, neuroscience researchers have curated several datasets. These datasets consist of stimuli and corresponding brain activity while participants interact with the stimuli and optionally perform tasks such as language comprehension, visual and auditory processing, etc. Next, we discuss various techniques for obtaining the brain recordings and methods for representing stimuli.

**Techniques for recording brain activations**: Popular techniques for recording brain activations can be broadly classified into invasive and non-invasive techniques, as shown in Figure 3.2. Invasive techniques include single Micro-Electrode (ME), Micro-Electrode array (MEA), and Electro-Corticography (ECoG). The non-invasive recording techniques include functional magnetic resonance imaging (fMRI), Magneto-encephalography (MEG), Electro-encephalography (EEG), and Near-Infrared Spectroscopy (NIRS). Apart from the dimension of invasiveness, these techniques differ in their spatial resolution of neural recording and temporal resolution. fMRI recording enables data acquisition at high spatial but low temporal resolution. Hence, they are suitable for examining which brain parts handle critical functions. A typical whole brain fMRI acquisition takes 1-4 seconds to complete a scan. This is far slower than the speed at which humans can process language. On the other hand, both MEG and EEG have high temporal but low spatial resolution. They can preserve rich syntactic information but cannot be used for source analysis [Hale et al., 2018]. fNIRS offers a compromise option. The time resolution is better than fMRI, and spatial resolution is better than EEG. However, this spatial and temporal resolution balance may not compensate for both loss and its restriction in terms of only recording cortical activity but not from nuclei deeper in the brain such as the basal ganglia, amygdala, hippocampus, etc.

**Stimulus Representations**: Neuroscience datasets contain stimuli across various modalities, including text, visual, audio, video, and other multimodal forms. Representations differ based on the modality.

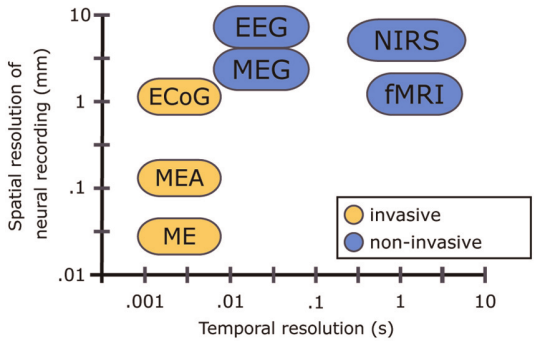



Figure 3.2: Overview of different brain–machine interfacing methods and their spatial and temporal resolution. Methods included: electroencephalography (EEG), magnetoencephalography (MEG), near-infrared spectroscopy (NIRS), functional magnetic resonance imaging (fMRI), electrocorticography (ECoG), microelectrode array (MEA) recordings and single microelectrode (ME) recordings. *This Figure is reproduced from Van Gerven et al. [2009].*

We briefly discuss the extraction of stimulus representations from DNN models according to the following criteria: (1) Traditional and advanced models for text-based stimulus representations. (2) Image-based representations from deep vision models. (3) Extraction of low-level speech to Transformer-based speech-based auditory representations. (4) Finally, for multimodal stimulus representations, we explore both early fusion and late fusion deep learning methods. Early fusion methods combine information across modalities at the initial processing stages, whereas late fusion combines it only at the end. Further details on different stimulus representation methods are discussed in Section 3.2.

**Naturalistic Neuroscience Datasets**: Several neuroscience datasets have been proposed across modalities (see Figure 3.3). These datasets differ in terms of the following criteria: (1) Method for recording activations: fMRI, EEG, MEG, etc. (2) Repetition time (TR), i.e., the sampling rate. (3) Characteristics of fixation points: location, color, shape. (4) Form of stimuli presentation: text, video, audio, images, or multimodality. (5) Task that participant performs during recording sessions: question answering, property generation, rating quality, etc. (6) Time given to participants for the task, e.g., 1 minute to list properties. (7) Demography of participants: males or females, sighted or blind, etc. (8) Number of times the response to stimuli was recorded. (9) Natural language associated with the stimuli. We discuss details of proposed datasets in Sec. 5.2.

**Evaluation of Brain Encoding and Decoding Methods**: 2V2 accuracy and Pearson Correlation are two famous metrics for evaluating brain encoding models. On the other hand, brain decoding models are evaluated using metrics such as pairwise accuracy, rank accuracy, $R^2$ score, and mean squared error. We discuss the detailed definitions of these metrics in Sec. 6.4.3.

**Computational Cognitive Neuroscience (CCN) Research goals**: CCN researchers have primarily focused on two main areas [Doerig et al., 2023].

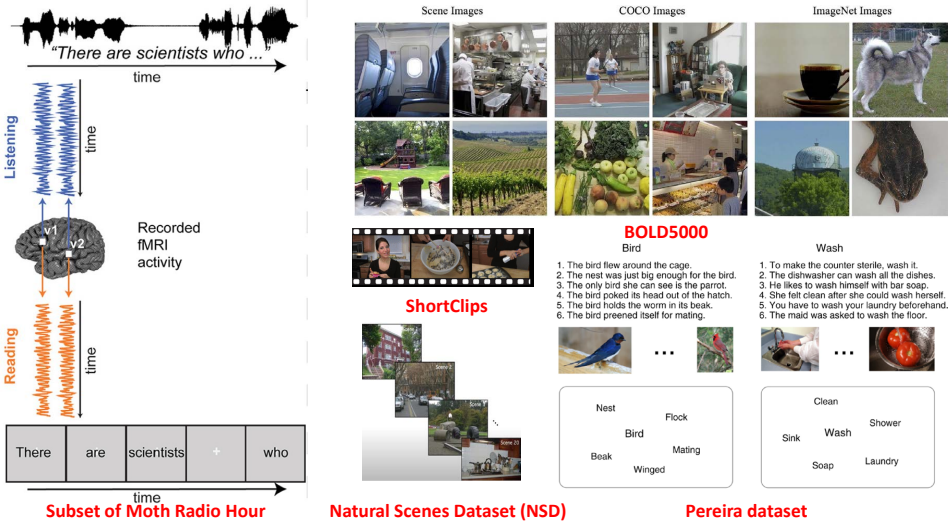



Figure 3.3: Representative Samples of Naturalistic Brain Datasets: (Left) Brain activity recorded when subjects are reading and listening to the same narrative [Deniz et al., 2019], and (Right) example naturalistic stimuli from various public repositories: BOLD5000 [Chang et al., 2019], ShortClips [Huth et al., 2022], Natural Scenes Dataset (NSD) [Allen et al., 2022] and Pereira dataset [Pereira et al., 2018].

1. Improving predictive accuracy. In this area, the work is around the following questions.
    - Compare feature sets: Which feature set provides the most faithful reflection of the neural representational space?
    - Test feature decodability: "Does neural data Y contain information about features X?"
    - Build accurate models of brain data: The aim is to enable the simulation of neuroscience experiments.
2. Interpretability. In this area, the work is around the following questions.
    - Examine individual features: Which contribute most to neural activity?
    - Test correspondences between representational spaces: "CNNs vs ventral visual stream" or "Two text representations".
    - Interpret feature sets: Do features X, generated by a known process, accurately describe the space of neural responses Y? Do voxels respond to a single feature or exhibit mixed selectivity?
    - How does the mapping relate to other brain function models or theories?

We discuss these questions in Sections 3.5 and 3.6.

Brain encoding literature [Mitchell et al., 2008, Wehbe et al., 2014, Huth et al., 2016] has focused on studying several important aspects: (1) Which models lead to better predictive accuracy across modalities? [Toneva and Wehbe, 2019, Deniz et al., 2019, Schrimpf

et al., 2021b] (2) How can we disentangle the contributions of syntax and semantics from language model representations to the alignment between brain recordings and language models? [Lopopolo et al., 2017, Reddy and Wehbe, 2021] (3) Why do some representations lead to better brain predictions? How are deep learning models and brains aligned regarding their information processing pipelines? [Merlin and Toneva, 2022, Aw and Toneva, 2023] (4) Does joint encoding of task and stimulus representations help? [Oota et al., 2023c]. We discuss these details of encoding methods in Sec. 3.5.

Brain decoding models aim to understand what a subject is thinking, seeing, and perceiving by analyzing neural recordings. Over the past decades, using non-invasive recordings, the brain-computer interface (BCI) has made significant progress in decoding stimuli (language/images/speech) from the brain. Like brain encoding literature, decoding literature studies a few essential aspects: (1) In the context of language, how we compose the linguistic meaning from different stimuli such as text, images, videos, or speech by analyzing the evoked brain activity [Pereira et al., 2016, 2018]. (2) Given brain activations corresponding to visual stimuli, how accurately can we decode a sentence representing the visual stimuli? [Nishimoto et al., 2011, Beliy et al., 2019] (3) How can we decode natural speech processing from non-invasive brain recordings using a single architecture and a data-driven approach? [Denk et al., 2023] (4) How accurately can we reconstruct perceived natural images or decode their semantic contents from non-invasive recording data using popular deep learning models? [Takagi and Nishimoto, 2023a]. We discuss these details of decoding methods in Sec. 3.6.

## 3.2 Stimulus Representations

This section discusses types of stimulus representations proposed in the literature across different modalities: text, visual, audio, video, and other multimodal stimuli.

**Text Stimulus Representations**: Older methods for text-based stimuli representation include text corpus co-occurrence counts [Mitchell et al., 2008, Pereira et al., 2013, Huth et al., 2016], topic models [Pereira et al., 2013], syntactic features and discourse features [Wehbe et al., 2014]. Recently, for text-based stimuli, both semantic models and experiential attribute models have been explored. Semantic representation models include word embedding methods [Pereira et al., 2018, Wang et al., 2020b, Pereira et al., 2016, Toneva and Wehbe, 2019, Anderson et al., 2017a, Oota et al., 2018], sentence representation models [Sun et al., 2020, 2019, Toneva and Wehbe, 2019], RNNs [Jain and Huth, 2018, Oota et al., 2019] and Transformer methods [Gauthier and Levy, 2019, Toneva and Wehbe, 2019, Schwartz et al., 2019, Schrimpf et al., 2021b, Antonello et al., 2021, Oota et al., 2022c, Aw and Toneva, 2023]. Popular word embedding methods include textual (i.e., Word2Vec [Mikolov et al., 2013a], fastText [Bojanowski et al., 2017], and GloVe [Pennington et al., 2014]), linguistic (i.e., dependency), conceptual (i.e., RWSGwn [Goikoetxea et al., 2015] and ConceptNet [Speer et al., 2017]), contextual (i.e., ELMo [Peters et al., 2018]). Famous sentence embedding models include average, max, concat of avg and max, SIF [Arora et al., 2017], SkipThoughts [Kiros et al., 2015], GenSen [Subramanian et al., 2018], InferSent [Conneau et al., 2017], ELMo, BERT [Devlin et al., 2019], RoBERTa [Liu

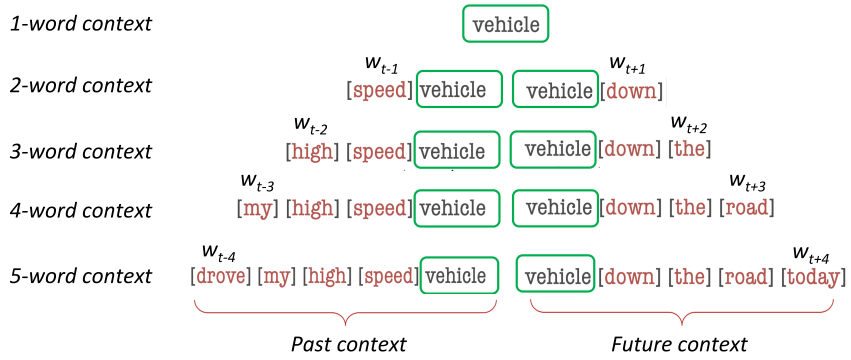



Figure 3.4: *Context representation of several orders:* Past / Future context is constructed by considering words preceding / succeeding the current word (see *Past / Future context* illustrated for current word *vehicle* for various orders).

et al., 2019], USE [Cer et al., 2018], QuickThoughts [Logeswaran and Lee, 2018] and GPT-2 [Radford et al., 2019]. Transformer-based methods include pretrained BERT with various NLU tasks, finetuned BERT, Transformer-XL [Dai et al., 2019], GPT-2, BART [Lewis et al., 2020], BigBird [Zaheer et al., 2020], Longformer [Beltagy et al., 2020], and LongT5 [Guo et al., 2022]. Experiential attribute models represent words in terms of human ratings of their degree of association with different attributes of experience, typically on a scale of 0-6 [Anderson et al., 2019, 2020, Berezutskaya et al., 2020, Just et al., 2010, Anderson et al., 2017b] or binary [Handjaras et al., 2016, Wang et al., 2017].

In the practice of employing word embeddings, encoding studies often utilize the average of word representations within a given context or derive complete sentence representations through sentence embedding models. More recently, brain encoding research has shifted towards the use of contextualized word representations, examining how the amount of context affects the brain predictivity [Jain and Huth, 2018, Toneva and Wehbe, 2019]. To obtain these contextualized word representations, Figure 3.4 illustrates how Past / Future context is constructed by considering words preceding / succeeding the current word. Given the constrained context length, each word is successively input to the network with at most $C$ previous tokens. For instance, given a story of $M$ words and considering the context length of 20, while the third word's vector is computed by inputting the network with ($w_1$, $w_2$, $w_3$), the last word's vectors $w_M$ is computed by inputting the network with ($w_{M-20}$, ..., $w_M$).

**Visual Stimulus Representations**: For visual stimuli, older methods used visual field filter bank [Thirion et al., 2006, Nishimoto et al., 2011] and Gabor wavelet pyramid [Kay et al., 2008, Naselaris et al., 2009]. As shown in Figure 3.5, recent methods use models like CNNs [Du et al., 2020, Beliy et al., 2019, Anderson et al., 2017a, Yamins et al., 2014, Nishida et al., 2020] and concept recognition models [Anderson et al., 2020].

**Audio Stimuli Representations**: For audio stimuli, phoneme rate and presence of phonemes have been leveraged [Huth et al., 2016]. Further, low-level speech features like filter banks (FBank), Mel Spectrogram, and MFCC from raw audio files, phonological features, articulation and power spectrum (PowSpec) feature vectors were used in Deniz et al. [2019].

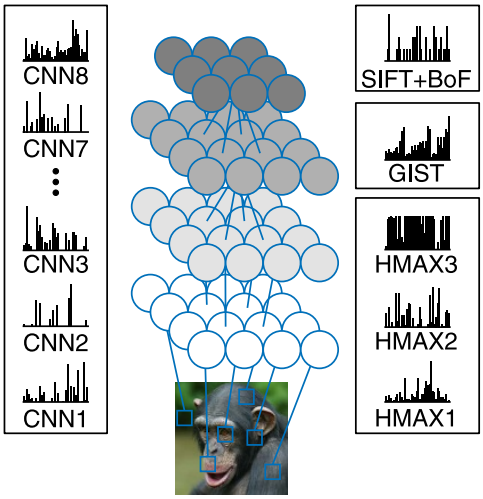



Figure 3.5: *Extraction of image representations:* Prior research has explored the impact of various layer-wise image representations from CNN models [Yamins et al., 2014, Horikawa and Kamitani, 2017], for both brain encoding and decoding models. The plot of the image feature extraction was derived from the study by Horikawa and Kamitani [2017].

Recently, Nishida et al. [2020] used features from an audio deep learning model called SoundNet for audio stimuli representation. To extract representations from Transformer-based speech models such as Wav2Vec2.0, HuBERT and Whisper, Vaidya et al. [2022], Antonello et al. [2024], Oota et al. [2023a] varied the length of the time windows from 16, 32, to 64 seconds, with strides ranging from 10 to 100 milliseconds, as illustrated in Figure 3.6. Moreover, these studies utilized an autoregressive approach to derive speech representations. This method involves considering the representations of the last frame within each window, allowing for the capture of temporal dynamics and contextual nuances in speech.

**Multimodal Stimulus Representations**: To jointly model the information from multimodal stimuli, recently, various multimodal representations have been used. These include processing videos using audio+image representations like VGG [Simonyan and Zisserman, 2015] and SoundNet [Aytar et al., 2016] in Nishida et al. [2020] or using image+text combination models like GloVe+VGG and ELMo+VGG in Wang et al. [2020b]. Recently, the usage of multimodal text+vision models like Contrastive Language-Image Pretraining (CLIP) [Radford et al., 2021], Learning Cross-Modality Encoder Representations from Transformers (LXMERT) [Tan and Bansal, 2019], and VisualBERT [Li et al., 2019] was proposed in Oota et al. [2022f].

## 3.3 Naturalistic Neuroscience Datasets

In this section, we discuss the popular text, visual, audio, video and other multimodal neuroscience datasets that have been proposed in the literature. Table 3.1 shows a detailed

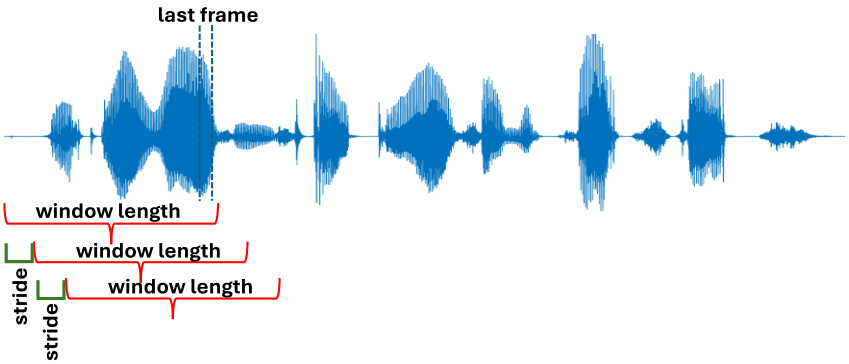



Figure 3.6: Representation of the last frame within each window allows for the capture of temporal dynamics and contextual nuances in the speech signal. The length of the time window is typically varied from 16, 32, to 64 secs, with strides ranging from 10 to 100 milliseconds.

overview of brain recording type, language, stimulus, number of subjects ($|S|$) and the task across datasets of different modalities. Figure 3.3 shows examples from a few datasets.

**Text Datasets**: These datasets are created by presenting words, sentences, passages, or chapters as stimuli. Some of the text datasets include Harry Potter Story [Wehbe et al., 2014], ZuCo EEG [Hollenstein et al., 2018] and datasets proposed in Handjaras et al. [2016], Anderson et al. [2017a, 2019], Wehbe et al. [2014]. In Handjaras et al. [2016], participants were asked to verbally enumerate in one minute the properties (features) that describe the entities the words refer to. There were four groups of participants: 5 sighted individuals were presented with a pictorial form of the nouns, five sighted individuals with a verbal-visual (i.e., written Italian words) form, 5 sighted individuals with a verbal auditory (i.e., spoken Italian words) form, and 5 congenitally blind with a verbal auditory form. Data proposed by Anderson et al. [2017a] contains 70 Italian words taken from seven taxonomic categories (abstract, attribute, communication, event/action, person/social role, location, object/tool) in the law and music domain. The word list contains concrete as well as abstract words. ZuCo dataset [Hollenstein et al., 2018] contains sentences for which EEG recordings were obtained for three tasks: regular reading of movie reviews, normal reading of Wikipedia sentences, and task-specific reading of Wikipedia sentences. For this dataset curation, sentences were presented to the subjects in a naturalistic reading scenario. A complete sentence is presented on the screen. Subjects read each sentence at their own speed, i.e., the reader determines how long each word is fixated on and which word to fixate on next.

**Visual Datasets**: Older visual datasets were based on binary visual patterns [Thirion et al., 2006]. Recent datasets contain natural images. Examples include Vim-1 [Kay et al., 2008], BOLD5000 [Chang et al., 2019], Algonauts [Cichy et al., 2019], NSD [Allen et al., 2022], Things-data [Hebart et al., 2023], and the dataset proposed in Horikawa and Kamitani [2017]. BOLD5000 includes $\sim$20 hours of MRI scans per each of the four participants. 4,916 unique images were used as stimuli from 3 image sources. Algonauts contains two sets of training data, each consisting of an image set and brain activity in RDM format (for fMRI and MEG). Training set 1 has 92 silhouette object images, and training set 2

| | Dataset | Authors | Type | Lang. | Stimulus | \|S\| | Task |
|---|---|---|---|---|---|---|---|
| Text | Harry Potter | Wehbe et al. [2014] | fMRI/MEG | English | Reading Chapter 9 of Harry Potter and the Sorcerer's Stone | 9 | Story understanding |
| | - | Handjaras et al. [2016] | fMRI | Italian | Verbal, pictorial or auditory presentation of 40 concrete nouns, four times | 20 | Property Generation |
| | - | Anderson et al. [2017a] | fMRI | Italian | Reading 70 concrete and abstract nouns from law/music, five times | 7 | Imagine a situation with noun |
| | ZuCo | Hollenstein et al. [2018] | EEG | English | Reading 1107 sentences with 21,629 words from movie reviews | 12 | Rate movie quality |
| | 240 Sentences with Content Words | Anderson et al. [2019] | fMRI | English | Reading 240 active voice sentences describing everyday situations | 14 | Passive reading |
| | BCCWJ-EEG | [Oseki and Asahara, 2020] | EEG | Japanese | Reading 20 newspaper articles for ∼30-40 minutes | 40 | Passive reading |
| | Subset Moth Radio Hour | Deniz et al. [2019] | fMRI | English | Reading 11 stories | 9 | Passive reading and Listening |
| Visual | - | Thirion et al. [2006] | fMRI | - | Viewing rotating wedges (8 times), expanding/contracting rings (8 times), rotating 36 Gabor filters (4 times), grid (36 times) | 9 | Passive viewing |
| | Vim-1 | Kay et al. [2008] | fMRI | - | Viewing sequences of 1870 natural photos | 2 | Passive viewing |
| | Generic Object Decoder | Horikawa and Kamitani [2017] | fMRI | - | Viewing 1,200 images from 150 object categories; 50 images from 50 object categories; imagery 10 times | 5 | Repetition detection |
| | BOLD5000 | Chang et al. [2019] | fMRI | - | Viewing 5254 images depicting real-world scenes | 4 | Passive viewing |
| | Algonauts | Cichy et al. [2019] | fMRI/MEG | - | Viewing 92 silhouette object images and 118 images of objects on natural background | 15 | Passive viewing |
| | NSD | [Allen et al., 2022] | fMRI | - | Viewing 73000 natural scenes | 8 | Passive viewing |
| | THINGS | [Hebart et al., 2023] | fMRI/MEG | - | Viewing 31188 natural images | 8 | Oddball Detection |
| Audio | - | Handjaras et al. [2016] | fMRI | Italian | Verbal, pictorial or auditory presentation of 40 concrete nouns, 4 times | 20 | Property Generation |
| | The Moth Radio Hour | Huth et al. [2016] | fMRI | English | Listening eleven 10-minute stories | 7 | Passive Listening |
| | Narrative Brain Dataset | Lopopolo et al. [2018] | fMRI | Dutch | Spoken presentation of short excerpts of three stories | 24 | Passive Listening |
| | - | Brennan and Hale [2019] | EEG | English | Listening Chapter one of Alice's Adventures in Wonderland (2,129 words in 84 sentences) as read by Kristen McQuillan | 33 | Question answering |
| | - | Anderson et al. [2020] | fMRI | English | Listening one of 20 scenario names, 5 times | 26 | Imagine personal experiencs |
| | Narratives | Nastase et al. [2020b] | fMRI | English | Listening 27 diverse naturalistic spoken stories. 891 functional scans | 345 | Passive Listening |
| | Natural Stories | Zhang et al. [2020a] | fMRI | English | Listening Moth-Radio-Hour naturalistic spoken stories. | 19 | Passive Listening |
| | The Little Prince | Li et al. [2021] | fMRI | English, Chinese, French | Listening audiobook for about 100 minutes. | 112 | Passive Listening |
| | MEG-MASC | Gwilliams et al. [2023a] | MEG | English | Listening two hours of naturalistic stories. 208 MEG sensors | 27 | Passive Listening |
| | Music Genre | Nakai et al. [2022] | fMRI | English | Listening 540 music pieces from 10 music genres | 5 | Passive Listening |
| Video | BBC's Doctor Who | Seeliger et al. [2019] | fMRI | English | Viewing spatiotemporal visual and auditory videos (30 episodes). 120.8 whole-brain volumes (∼23 h) of single-presentation data, and 1.2 volumes (11 min) of repeated narrative short episodes. 22 repetitions | 1 | Passive viewing |
| | Japanese Ads | Nishida et al. [2020] | fMRI | Japanese | Viewing 368 web and 2452 TV Japanese ad movies (15-30s). 7200 train and 1200 test fMRIs for web; fMRIs from 420 ads. | 52 | Passive viewing |
| | Pippi Langkous | Berezutskaya et al. [2020] | ECoG | Swedish/Dutch | Viewing 30 s excerpts of a feature film (in total, 6.5 min long), edited together for a coherent story | 37 | Passive viewing |
| | Algonauts | Cichy et al. [2021] | fMRI | English | Viewing 1000 short video clips (3 sec each) | 10 | Passive viewing |
| | Natural Short Clips | Huth et al. [2022] | fMRI | English | Watching natural short movie clips | 5 | Passive viewing |
| | Natural Short Clips | Lahner et al. [2023] | fMRI | English | Watching 1102 natural short video clips | 10 | Passive viewing |
| Other Multimodal | 60 Concrete Nouns | Mitchell et al. [2008] | fMRI | English | Viewing 60 different word-picture pairs from 12 categories, 6 times each | 9 | Passive viewing |
| | - | Sudre et al. [2012] | MEG | English | Reading 60 concrete nouns along with line drawings. 20 questions per noun lead to 1200 examples. | 9 | Question answering |
| | - | Zinszer et al. [2018] | fNIRS | English | 8 concrete nouns (audiovisual word and picture stimuli): bunny, bear, kitty, dog, mouth, foot, hand, and nose; 12 times repeated. | 24 | Passive viewing and listening |
| | Pereira | Pereira et al. [2018] | fMRI | English | Viewing 180 Words with Picture, Sentences, word clouds; reading 96 text passages; 72 passages. 3 times repeated. | 16 | Passive viewing and reading |
| | - | Cao et al. [2021] | fNIRS | Chinese | Viewing and listening 50 concrete nouns from 10 semantic categories. | 7 | Passive viewing and listening |
| | Neuromod | Boyle et al. [2020] | fMRI | English | Watching TV series and movies (Friends, Movie10) | 6 | Passive viewing and listening |

Table 3.1: Naturalistic Neuroscience Datasets. Publicly available datasets are linked to their sources in the Dataset column. In this table, |S| represents the number of participants in each dataset.

has 118 object images with natural backgrounds. Testing data consists of 78 images of objects on natural backgrounds. Most visual datasets involve passive viewing, but the dataset in Horikawa and Kamitani [2017] involved the participant doing the one-back repetition detection task.

**Audio Datasets**: Most of the proposed audio datasets are in English [Huth et al., 2016, Brennan and Hale, 2019, Anderson et al., 2020, Nastase et al., 2020b], while there is one Handjaras et al. [2016] on Italian, and another one Li et al. [2021] in Chinese and French. The participants were involved in a variety of tasks while their brain activations were measured: Property generation [Handjaras et al., 2016], passive listening [Huth et al., 2016, Nastase et al., 2020b], question answering [Brennan and Hale, 2019] and imagining themselves personally experiencing common scenarios [Anderson et al., 2020]. In the last one, participants underwent fMRI as they reimagined the scenarios (e.g., resting, reading, writing, bathing, etc.) when prompted by standardized cues. Narratives Nastase et al. [2020b] used 17 different stories as stimuli. Across subjects, it is 6.4 days worth of recordings.

**Video Datasets**: Recently, video neuroscience datasets have also been proposed. These include BBC's Doctor Who [Seeliger et al., 2019], Japanese Ads [Nishida et al., 2020], Pippi Langkous [Anderson et al., 2020] and Algonauts [Cichy et al., 2021]. Japanese Ads data contains data for two movies provided by NTT DATA Corp: web and TV ads. Four types of cognitive labels are also associated with the movie datasets: scene descriptions, impression ratings, ad effectiveness indices, and ad preference votes. Algonauts 2021 contains fMRIs from 10 human subjects that watched over 1,000 short (3 sec) video clips.

**Other Multimodal Datasets**: Finally, beyond the video datasets, datasets have also been proposed with other kinds of multimodality. These datasets are audiovisual ([Zinszer et al., 2018, Cao et al., 2021]), words associated with line drawings [Mitchell et al., 2008, Sudre et al., 2012], pictures along with sentences and word clouds [Pereira et al., 2018]. These datasets have been collected using a variety of methods like fMRIs [Mitchell et al., 2008, Pereira et al., 2018], MEG [Sudre et al., 2012] and fNIRS [Zinszer et al., 2018, Cao et al., 2021]. Specifically, in Sudre et al. [2012], subjects were asked to perform a question-answering (QA) task while their brain activity was recorded using MEG. Subjects were first presented with a question (e.g., "Is it manmade?"), followed by 60 concrete nouns and their line drawings in a random order. For all other datasets, subjects performed passive viewing and/or listening.

## 3.4 Evaluation Metrics

In this section, we discuss popular metrics for evaluation of brain encoding and decoding models.

### 3.4.1 Metrics for Brain Encoding Models

Two metrics are popularly used to evaluate brain encoding models: 2V2 accuracy [Toneva et al., 2020, Oota et al., 2022c] and Pearson Correlation [Jain and Huth, 2018]. They are

defined as follows. Given a subject and a brain region, let $N$ be the number of samples. Let $\{Y_i\}_{i=1}^N$ and $\{\hat{Y}_i\}_{i=1}^N$ denote the actual and predicted voxel value vectors for the $i^{th}$ sample. Thus, $Y \in R^{N \times V}$ and $\hat{Y} \in R^{N \times V}$ where $V$ is the number of voxels in that region.

**2V2 Classification Accuracy** This metric evaluates how close the brain activity prediction is from ground truth, such as Euclidean distance, cosine distance. This metric evaluates the fMRI predictions by using them in a classification task on held-out data in the cross-validation setting. The classification task is to trying to match the predicted left-out brain responses to their corresponding ground truth, as introduced in Mitchell et al. [2008], Wehbe et al. [2014], Toneva et al. [2020], Aw and Toneva [2023]. Having two sets of brain predictions $\hat{Y}_i$ and $\hat{Y}_j$, and corresponding ground truth $Y_i$ and $Y_j$, the 2V2 classification accuracy is computed as $\frac{1}{N_{C_2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} I[\{\text{cosD}(Y_i, \hat{Y}_i) + \text{cosD}(Y_j, \hat{Y}_j)\} < \{\text{cosD}(Y_i, \hat{Y}_j) + \text{cosD}(Y_j, \hat{Y}_i)\}]$ where cosD is the cosine distance function. $I[c]$ is an indicator function such that $I[c] = 1$ if $c$ is true, else it is 0. The higher the 2V2 accuracy, the better. Figure 3.7 (left) illustrates computation of 2V2 Accuracy for the case where sample $i$ and $j$ correspond to the brain activity of concepts "dog" and "house", respectively. This metric was proposed to boost the signal-to-noise ratio in estimating the brain alignment for single-trial data [Aw and Toneva, 2023]. Under this metric, chance performance is 50%.

**Pearson Correlation** This metric evaluates the similarity between the fMRI predictions ($\hat{Y}_i$) and the corresponding true fMRI data ($Y_i$) by computing the Pearson correlation for each voxel i. The Pearson correlation for voxel i is computed as PC$_i$=$corr[Y_i, \hat{Y}_i]$ where corr is the correlation function. The average Pearson correlation across all voxels is then computed as PCC=$\frac{1}{N} \sum_{i=1}^{n} corr[Y_i, \hat{Y}_i]$, where N denotes number of voxels. This metric is widely used in cognitive neuroscience [Jain and Huth, 2018, Toneva and Wehbe, 2019, Caucheteux et al., 2021a, Goldstein et al., 2022, Aw and Toneva, 2023, Oota et al., 2022c, 2023c].

**Noise Ceiling Estimate** To account for the intrinsic noise in biological measurements and obtain a more accurate estimate of the model's performance, Schrimpf et al. [2021b] proposed an approach to estimate the noise ceiling. This is achieved by estimating the amount of brain response in one subject that can be predicted using only the data from a combination of other subjects, using an encoding model. For instance, consider *Harry Potter* dataset with n=8 participants, the first step is to subsample—the data with $n$ participants into all possible combinations of $s$ participants for all $s \in [2, 8]$ (e.g. 2, 3, 4, 5, 6, 7, 8 for $n$=8). In the second step, for each subsample, select a random participant as the target that we attempt to predict from the remaining $s - 1$ participants (e.g., predict 1 subject from 1 (other) subject, 1 from 2 subjects, ..., 1 from 8, to obtain a mean score for each voxel in that subsample. In the third step, extrapolate to infinitely many humans and thus to obtain the highest possible (most conservative) estimate, as suggested by Schrimpf et al. [2021b], fit the equation $v = v_0 \times \left(1 - e^{-\frac{x}{\tau_0}}\right)$ where $x$ is each subsample's number of participants, $v$ is each subsample's correlation score and $v_0$ and $\tau_0$ are the fitted parameters. This fitting was performed for each voxel independently with 100 bootstraps each to estimate the variance where each bootstrap draws $x$ and $v$ with replacement. The final ceiling value was the median of the per-voxel ceilings $v_0$.

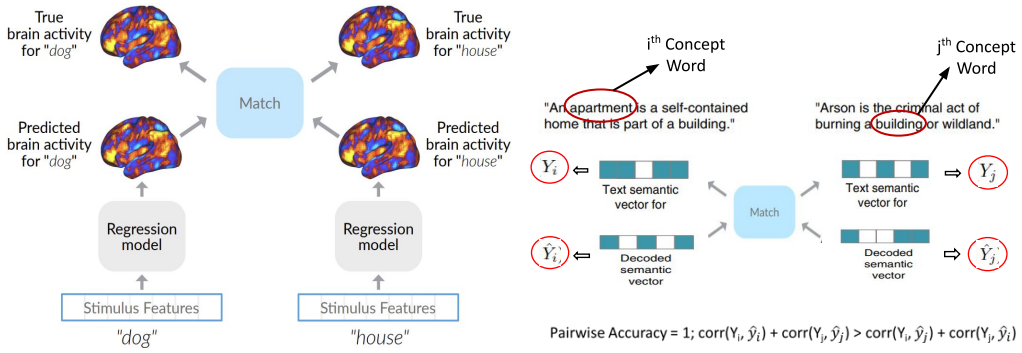



Figure 3.7: Evaluation Metrics for Brain Encoding and Decoding. (Left) 2V2 Accuracy [Toneva et al., 2020], (Right) Pairwise Accuracy [Pereira et al., 2018]. *The left Figure is adapted from Toneva et al. [2020] and the right Figure is adapted from Pereira et al. [2018].*

**Normalized Brain Alignment** The neural model predictivity values were normalized by their respective subject estimated noise ceiling values, as proposed by Schrimpf et al. [2021b]. The final measure of a model's performance ('normalized brain alignment' or 'score') on a dataset is thus Pearson's correlation between model predictions and neural recordings divided by the estimated ceiling and averaged across voxel locations and participants.

### 3.4.2 Metrics for Brain Decoding Models

Brain decoding methods are evaluated using popular metrics like pairwise and rank accuracy [Pereira et al., 2018, Sun et al., 2019, 2020, Oota et al., 2022d]. Other metrics used for brain decoding evaluation include $R^2$ score, mean squared error, and using Representational Similarity Matrix [Cichy et al., 2019, 2021].

**Pairwise Accuracy** is computed as follows. The first step is to predict all the test stimulus vector representations using a trained decoder model. Let S = [$S_0$, $S_1$,$\cdots$,$S_n$], $\hat{S}$ = [$\hat{S}_0$, $\hat{S}_1$,$\cdots$,$\hat{S}_n$] denote the "true" (stimuli-derived) and predicted stimulus representations for $n$ test instances resp. Given a pair $(i, j)$ such that $0 \leq i, j \leq n$, score is 1 if *corr*($S_i$,$\hat{S}_i$) + *corr*($S_j$,$\hat{S}_j$) > *corr*($S_i$,$\hat{S}_j$) + *corr*($S_j$,$\hat{S}_i$), else 0. Here, *corr* denotes the Pearson correlation. Figure 3.7 (right) illustrates the computation of Pairwise Accuracy for the case where sample $i$ and $j$ correspond to the brain activations for text stimuli "apartment" and "building" respectively. Final pairwise matching accuracy per participant is the average of scores across all pairs of test instances.

**Rank Accuracy** is computed as follows. We first compare each decoded vector to all the "true" stimuli-derived semantic vectors and rank them by their correlation. The classification performance reflects the rank $r$ of the stimuli-derived vector for the correct word, or picture stimuli: $1 - \frac{r-1}{\#instances-1}$. The final accuracy value for each participant is the average rank accuracy across all instances.

| Authors | Stimulus Representations | \|S\| | Dataset | Delays |
|---|---|---|---|---|
| Jain et al. [2020] | LSTM | 6 | Moth-Radio-Hour | 8secs (4 TRs) |
| Jain and Huth [2018] | LSTM | 6 | Moth-Radio-Hour | 8secs (4 TRs) |
| Caucheteux et al. [2021a] | GPT-2 | 345 | Narratives | 7.5secs (5 TRs) |
| Reddy and Wehbe [2021] | Syntax Parsers, BERT | 8 | Harry-Potter | 8secs (4 TRs) |
| Merlin and Toneva [2022] | GPT2 | 8 | Harry-Potter | 8secs (4 TRs) |
| Aw and Toneva [2023] | BART, LongT5, LED | 8 | Harry-Potter | 8secs (4TRs) |
| Antonello et al. [2021] | 100 Language Models | 7 | Moth-Radio-Hor | 8secs (4 TRs) |
| Oota et al. [2023c] | BERT and Probing Tasks | 18 | Narratives $21st$-Year | 9secs (6 TRs) |
| Oota et al. [2023g] | BERT, GPT-2, Wav2Vec2.0 | 6 | Moth-radio-hour | 12secs (6 TRs) |

Table 3.2: Summary of Brain Encoding Studies with constant HRF delays. Here, |S| denotes number of participants. These are studies on English text using fMRI activations.

## 3.5 Brain Encoding

Encoding is the learning of the mapping from the stimulus domain to the neural activation. The quest in brain encoding is for "reverse engineering" the algorithms that the brain uses for sensation, perception, and higher-level cognition. The foundational approach to constructing a brain encoder, illustrated in Figure 3.8, adopts a general brain alignment strategy previously implemented in several notable studies [Jain and Huth, 2018, Toneva and Wehbe, 2019, Aw and Toneva, 2023, Oota et al., 2023c]. This method focuses on predicting fMRI recordings at every voxel for each participant, utilizing DNN representations that mirror the participant's engagement in tasks such as reading or listening.

Building on this foundation, the recent advancements in neuroimaging technologies have enhanced our ability to closely approximate how the brain responds to different types of stimuli, thereby deepening our understanding of the brain's information processing mechanisms. Concurrently, advancements in deep neural network (DNN) models have led to the development of highly efficient models across different modalities, including language, vision, speech, and multimodal interactions. These models have set new benchmarks in performance for a wide range of applications. Leveraging cutting-edge neuroimaging techniques and DNN models, this section offers a comprehensive review of the task settings for brain encoding, latest achievements in understanding language processing, visual object recognition, auditory perception, and multimodal processing in the brain.

In the discussion on encoding task settings, we present stimulus downsampling, TR alignment and voxelwise encoding models. In linguistic brain encoding, we explore recent breakthroughs in applied Natural Language Processing (NLP) that facilitate the reverse engineering of the language function of the brain. In the realm of vision brain encoding, pioneering results have been achieved in reverse engineering the function of the ventral visual stream for object recognition, thanks to the advancements and impressive successes of deep Convolutional Neural Networks (CNNs) and Vision Transformers. Additionally, we present the latest insights into auditory and multimodal brain encoding. This systematic approach informs the organization of this section. Overall, Figure 3.9 classifies the encoding literature along various stimulus domains such as vision, auditory, multimodal, and language and the corresponding tasks in each domain. Finally, Table 3.3 summarizes various encoding models proposed in the literature related to textual, audio, visual, and multimodal stimuli.

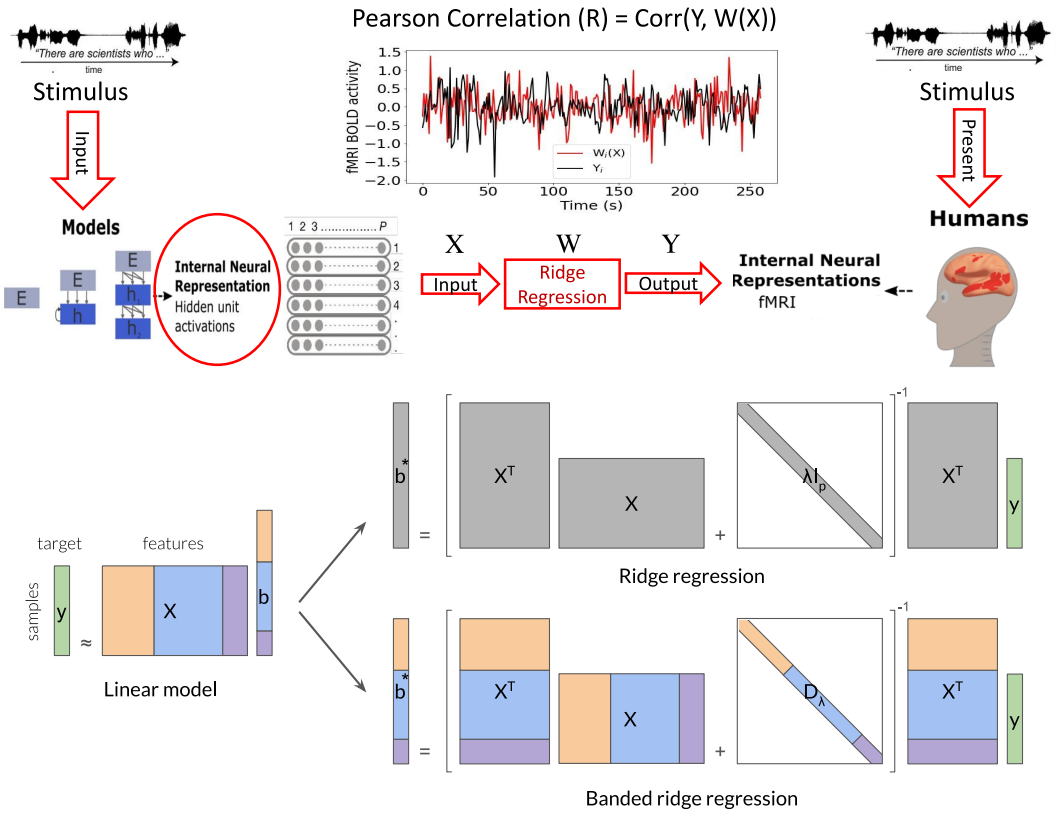



Figure 3.8: Scheme for Brain Encoding (top): this approach learns a function to predict the fMRI recordings at every voxel of each participant using the model representations that correspond to the same text read or listened by the participant. Ridge regression vs. Banded ridge regression (bottom), *adapted from la Tour et al. [2022]*. Each color (or band) represents a different feature space.

### 3.5.1 Encoding Task Settings

#### Stimulus Downsampling

In the context of narrative story reading or listening, the rate of fMRI data acquisition was lower than the rate at which the text stimulus was presented to the subjects, several words fall under the same TR in a single acquisition. Hence, previous studies match the stimulus acquisition rate to fMRI data recording by downsampling the stimulus features using a 3-lobed Lanczos filter [Huth et al., 2016, Jain and Huth, 2018, Toneva and Wehbe, 2019, Antonello et al., 2021, Oota et al., 2023c]. After downsampling, word-embeddings corresponding to each TR are obtained. For the naturalistic audio, Vaidya et al. [2022], Antonello et al. [2024] windowed the stimulus waveform with a sliding window of size 16 s and stride 100ms before feeding it into model. Further, the features are downsampled as previously described, using Lanczos interpolation, to match with sampling rate of fMRI recordings.

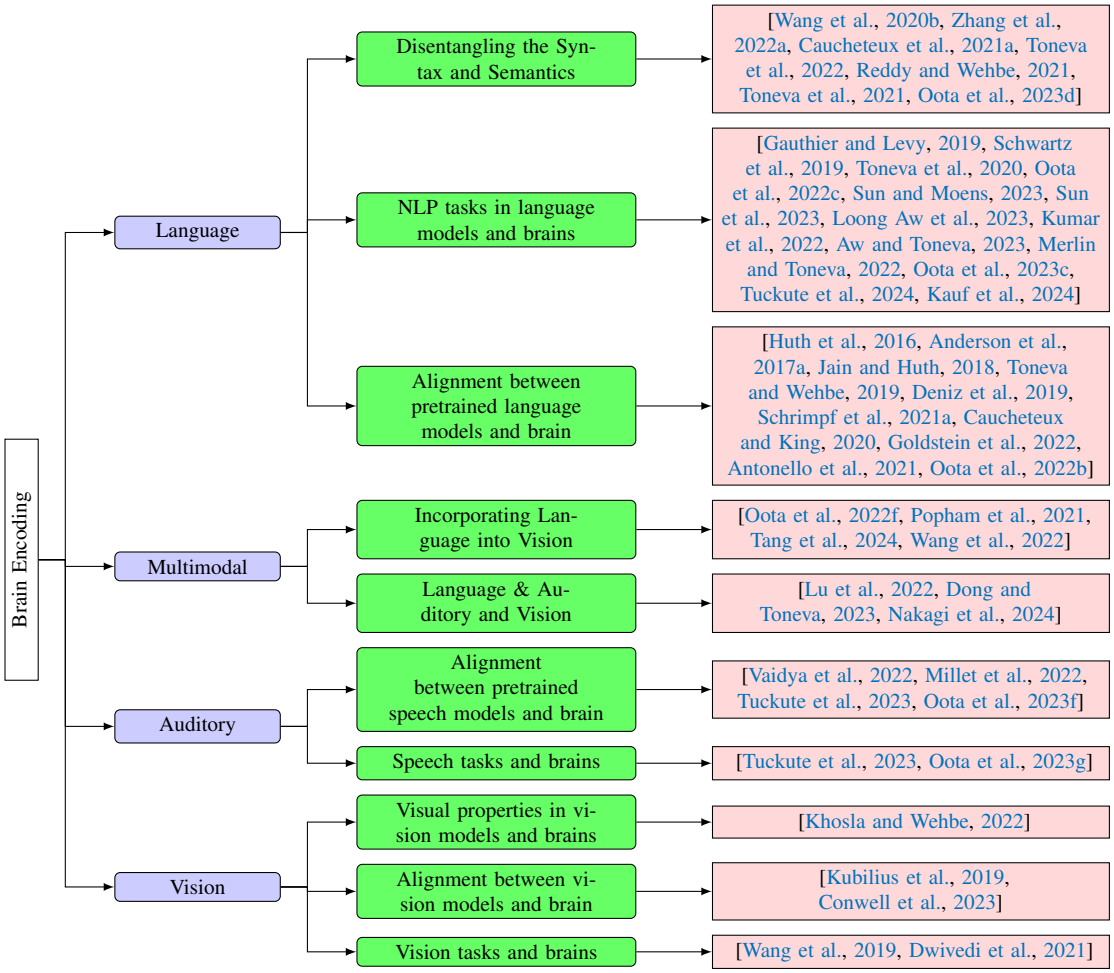



Figure 3.9: Categorization of Brain Encoding Studies

Similarly for the naturalistic videos, the rate of fMRI data acquisition (TR = 2 seconds) in the shortclips dataset [Huth et al., 2022] is lower than the rate at which the stimulus was presented to the subjects (15 frames per second), 30 frames of a video were viewed under the same TR for a single fMRI acquisition [Popham et al., 2021]. This helps synchronization between the stimulus presentation rate and fMRI data recording, which we then leverage to train our encoding models.

### fMRI Time Repetition (TR) Alignment

To account for the slowness of the hemodynamic response, in general, previous studies model the HRF using a finite response filter (FIR) per voxel and for each subject separately with delay of 8 to 12 secs [Jain and Huth, 2018, Toneva and Wehbe, 2019, Popham et al., 2021, Oota et al., 2023c, Antonello et al., 2024]. Table 3.2 summarizes current brain encoding studies with a fixed HRF delay.

| Stimuli | Authors | Dataset Type | Lang. | Stimulus Representations | \|S\| | Dataset |
|---|---|---|---|---|---|---|
| Text | Jain and Huth [2018] | fMRI | English | LSTM | 6 | Subset Moth Radio Hour |
| | Toneva and Wehbe [2019] | fMRI/ MEG | English | ELMo, BERT, Transformer-XL | 9 | Story understanding |
| | Toneva et al. [2020] | MEG | English | BERT | 9 | Question-Answering |
| | Schrimpf et al. [2021b] | fMRI/ECoG | English | 43 language models (e.g. GloVe, ELMo, BERT, GPT-2, XLNET) | 20 | Neural architecture of language |
| | Gauthier and Levy [2019] | fMRI | English | BERT, finetuned NLP tasks (Sentiment, Natural language inference), Scrambling language model | 7 | Imagine a situation with the noun |
| | Deniz et al. [2019] | fMRI | English | GloVe | 9 | Subset Moth Radio Hour |
| | Jain et al. [2020] | fMRI | English | LSTM | 6 | Subset Moth Radio Hour |
| | Caucheteux et al. [2021a] | fMRI | English | GPT-2, Basic syntax features | 345 | Narratives |
| | Antonello et al. [2021] | fMRI | English | GloVe, BERT, GPT-2, Machine Translation, POS tasks | 6 | Moth Radio Hour |
| | Reddy and Wehbe [2021] | fMRI | English | Constituency, Basic syntax features and BERT | 8 | Harry Potter |
| | Goldstein et al. [2022] | fMRI | English | GloVe, GPT-2 next word, pre-onset, post-onset word surprise | 8 | ECoG |
| | Oota et al. [2022c] | fMRI | English | BERT and GLUE tasks | 82 | Pereira & Narratives |
| | Oota et al. [2022b] | fMRI | English | ESN, LSTM, ELMo, Longformer | 82 | Narratives |
| | Merlin and Toneva [2022] | fMRI | English | BERT, Next word prediction, multi-word semantics, scrambling model | 8 | Harry Potter |
| | Toneva et al. [2022] | fMRI / MEG | English | ELMo, BERT, Context Residuals | 8 | Harry Potter |
| | Aw and Toneva [2023] | fMRI | English | BART, Longformer, Long-T5, BigBird, and corresponding Booksum models as well | 8 | Passive reading |
| | Zhang et al. [2022b] | fMRI | English, Chinese | Node Count | 19, 12 | Zhang |
| | Oota et al. [2023d] | fMRI | English | Constituency, Dependency trees, Basic syntax features and BERT | 82 | Narratives |
| | Oota et al. [2023e] | MEG | English | Basic syntax features, GloVe and BERT | 8 | MEG-MASC |
| | Tuckute et al. [2024] | fMRI | English | BERT-Large, GPT-2 XL | 12 | Reading Sentences |
| | Kauf et al. [2024] | fMRI | English | BERT-Large, GPT-2 XL | 12 | Pereira |
| | Singh et al. [2023] | fMRI | English | BERT-Large, GPT-2 XL, Text Perturbations | 5 | Pereira |
| Visual | Wang et al. [2019] | fMRI | - | 21 downstream vision tasks | 4 | BOLD 5000 |
| | Kubilius et al. [2019] | fMRI | - | CNN models AlexNet, ResNet, DenseNet | 7 | Algonauts |
| | Dwivedi et al. [2021] | fMRI | - | 21 downstream vision tasks | 4 | BOLD 5000 |
| | Khosla and Wehbe [2022] | fMRI | - | CNN models AlexNet | 4 | BOLD 5000 |
| | Conwell et al. [2023] | fMRI | - | CNN models AlexNet | 4 | BOLD 5000 |
| Audio | Millet et al. [2022] | fMRI | English | Wav2Vec2.0 | 345 | Narratives |
| | Vaidya et al. [2022] | fMRI | English | APC, AST, Wav2Vec2.0, and HuBERT | 7 | Moth Radio Hour |
| | Tuckute et al. [2023] | fMRI | English | 19 Speech Models (e.g. DeepSpeech, Wav2Vec2.0, VQ-VAE) | 19 | Passive listening |
| | Oota et al. [2023f] | fMRI | English | 5 basic and 25 deep learning based speech models (Tera, CPC, APC, Wav2Vec2.0, HuBERT, DistilHuBERT, Data2Vec | 6 | Moth Radio Hour |
| | Oota et al. [2023g] | fMRI | English | Wav2Vec2.0 and SUPERB tasks | 82 | Narratives |
| Multi Modal | Dong and Toneva [2023] | fMRI | English | Merlo Reseve | 5 | Neuromod |
| | Popham et al. [2021] | fMRI | English | 985D Semantic Vector | 5 | Moth Radio Hour & Short Movie Clips |
| | Oota et al. [2022f] | fMRI | English | CLIP, VisualBERT, LXMERT, CNNs and BERT | 5, 82 | Periera & Narratives |
| | Lu et al. [2022] | fMRI | English | BriVL | 5 | Pereira & Short Movie Clips |
| | Tang et al. [2024] | fMRI | English | BridgeTower | 5 | Moth Radio Hour & Short Movie Clips |

Table 3.3: Summary of Representative Brain Encoding Studies.

### 3.5.2 MEG PREPROCESSING AND ALIGNMENT

The minimal processing steps described in Gwilliams et al. [2023a] are as follows. On raw MEG data and for each subject separately, using *MNE-Python* defaults parameters, the following steps should be executed:

- bandpass filtered the MEG data between 0.5 and 30.0 Hz,
- temporally-decimated the data 10x
- segmented these continuous signals between -200 ms and 600 ms after word onset (note: this continuous signals varies for phoneme onset)
- applied a baseline correction between -200 ms and 0 ms, and
- clipped the MEG data between fifth and ninety-fifth percentile of the data across channels.

In contrast to the fMRI recordings, MEG recordings have much higher time resolution. Epoching and downsampling MEG data can result in aligned word-level or phoneme-level brain data [Gwilliams et al., 2023a, Toneva et al., 2020, Oota et al., 2023e].

### 3.5.3 Voxel-wise Encoding Model

The main goal of voxel-wise encoding model is to predict brain responses associated with each brain voxel given a stimulus. To estimate the brain alignment of a DNN model of a stimulus representations via training standard voxel-wise encoding models [Deniz et al., 2019, Toneva and Wehbe, 2019]. Specifically, for each voxel and participant, prior studies train fMRI encoding model using ridge regression to predict the fMRI recording associated with this voxel as a function of the stimulus representations obtained from DNN models [Mitchell et al., 2008, Wehbe et al., 2014, Huth et al., 2016]. To simultaneously accommodate different feature spaces, which may necessitate varying levels of regularization, Nunez-Elizalde et al. [2019] proposed voxel-wise encoding model that utilize an advanced form of ridge regression. This method, known as banded ridge regression, introduces individual regularization parameters for each feature space, as illustrated in Figure 3.8. Before doing the ridge regression or banded ridge regression, each feature channel was first z-scored separately for both training and testing. This was done to match the features to the fMRI responses, which were also z-scored for training and testing. Formally, at the time step (t), stimuli are encoded as $X_t \in \mathbb{R}^{N \times D}$ and brain region voxels $Y_t \in \mathbb{R}^{N \times V}$, where $N$ is the number of training examples, $D$ denotes the dimension of the concatenation of delayed TRs, and $V$ denotes the number of voxels. To find the optimal regularization parameter for each feature space, a range of regularization parameters that is explored using cross-validation.

### 3.5.4 Linguistic Encoding

#### Alignment between pretrained Language Models (LMs) and Brain

Previous works have investigated the alignment between pretrained language models and brain recordings of people comprehending language. Huth et al. [2016] have identified brain ROIs (Regions of Interest) that respond to words with a similar meaning and have thus built a "semantic atlas" of how the human brain organizes language. Many studies

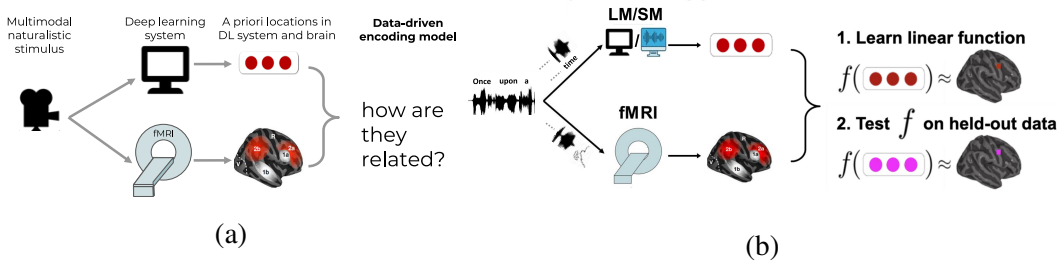



Figure 3.10: (a) Alignment of representations between deep learning systems and human brains [Toneva and Wehbe, 2019]. (b) For instance, a narrative story provided to both the Language model as well as human participants. For the Language model, we extract its representations for every word in the text. For the human participants, we record their brain activity using fMRI. Next, we train a linear function that uses the extracted Language model representations to predict human brain activity. Finally, we test this function on unseen data, and evaluate its accuracy as the amount of "brain alignment" [Toneva and Wehbe, 2019]. These two images are sourced from Cogsci-22 tutorial slides Oota et al. [2022e].

have shown accurate results in mapping brain activity using neural distributed word embeddings for linguistic stimuli [Anderson et al., 2017a, Pereira et al., 2018, Oota et al., 2018, Nishida and Nishimoto, 2018, Sun et al., 2019]. Unlike earlier models, where each word is represented as an an independent vector in an embedding space, Jain and Huth [2018] built encoding models using rich contextual representations derived from an LSTM language model in a story listening task. These contextual representations demonstrated dissociation in brain activation – auditory cortex (AC) and Broca's area in a shorter context, whereas left Temporo-Parietal Junction (TPJ) in a longer context. Hollenstein et al. [2019] presents the first multimodal framework for evaluating six types of word embeddings (Word2Vec, WordNet2Vec [Bartusiak et al., 2019], GloVe, fastText, ELMo, and BERT) on 15 datasets, including eye-tracking, EEG and fMRI signals recorded during language processing. With the recent advances in contextual representations in NLP, few studies incorporated them in relating sentence embeddings with brain activity patterns [Sun et al., 2020, Gauthier and Levy, 2019, Jat et al., 2020].

More recently, researchers have begun to study the alignment of language regions of the brain with the layers of language models (broadly following the method described in Figure 3.10) and found that the best alignment was achieved in the middle layers of these models [Jain and Huth, 2018, Toneva and Wehbe, 2019, Caucheteux and King, 2020], as shown in Figure 3.11. Toneva and Wehbe [2019] study how representations of various Transformer models differ across layer depth, context length, and attention type. The results demonstrated that across several larger NLP models, the middle layers of language models are well aligned with brain language regions. Schrimpf et al. [2021b] examined the relationship between 43 diverse state-of-the-art language models. They also studied the behavioral signatures of human language processing in self-paced reading times and a range of linguistic functions assessed via standard engineering tasks from NLP. They found that Transformer-based models perform better than RNNs or word-level embedding

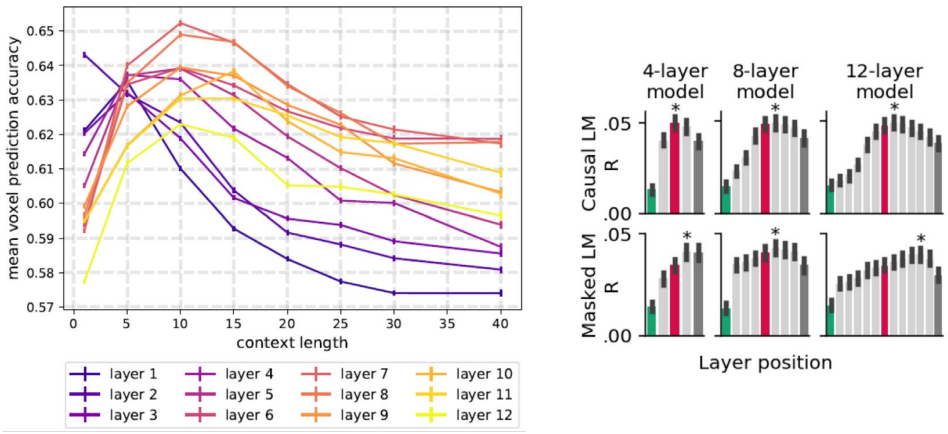



Figure 3.11: The strongest alignment with high-level language brain regions has consistently been observed in the middle layers. Left: Performance of BERT encoding model for all hidden layers as the amount of context provided to the network is increased [Toneva and Wehbe, 2019]. Right: fMRI encoding score (averaged across time and channels) of 6 representative transformers varying in tasks (CLM vs MLM) and depth (4-12 layers) [Caucheteux and King, 2020]. *The left Figure is adapted from Toneva and Wehbe [2019] and the right Figure is adapted from Caucheteux and King [2020].*

models. Larger-capacity models perform better than smaller models. Models initialized with random weights (prior to training) perform surprisingly similarly in neural predictivity compared to final trained models, suggesting that network architecture contributes as much or more than experience dependent learning to a model's match to the brain. Antonello et al. [2021] proposed a "language representation embedding space" and demonstrated the effectiveness of the features from this embedding in predicting fMRI responses to linguistic stimuli. Very recent work by Antonello et al. [2024] tested whether larger open-source models, such as those from the text-based model (OPT and LLaMA) families, are better at predicting brain responses recorded using fMRI. The results demonstrate that encoding performance improvements scale well with both model size and dataset size, and large datasets will no doubt be necessary in producing applicable encoding models.

### Disentangling the Syntax and Semantics

The representations of transformer models like BERT and GPT-2 have been shown to map onto brain activity during language comprehension linearly. Several studies have attempted to disentangle the contributions of different types of information from word representations to the alignment between brain recordings and language models [Lopopolo et al., 2017, Wang et al., 2020b, Caucheteux et al., 2021a, Reddy and Wehbe, 2021, Zhang et al., 2022a, Toneva et al., 2022, Oota et al., 2023d]. Wang et al. [2020b] proposed a two-channel variational autoencoder model to dissociate sentences into semantic and syntactic representations and separately associate them with brain imaging data to find feature-correlated brain regions. Similarly, Zhang et al. [2022a] separated different syntactic features from

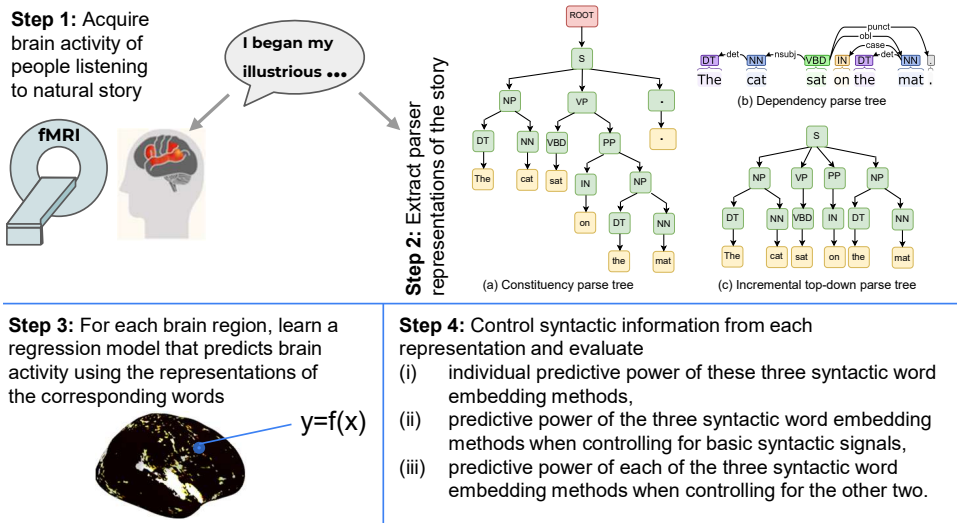



Figure 3.12: Four steps proposed in Oota et al. [2023d]: (1) fMRI acquisition, (2) Syntactic parsing, (3) Regression model training, and (4) Predictive power analysis of the three embeddings methods. This Figure is adapted from Oota et al. [2023d].

pretrained BERT representations, to explore the potential for distinct syntactic and semantic processing language regions in the brain. Compared to lexical word representations, word syntactic features (parts-of-speech, named entities) and word-relation features (semantic roles, dependencies) are distributed across brain networks instead of a local brain region. The previous two studies could not conclude whether all or any of these representations effectively drive the linear mapping between language models (LMs) and the brain. Toneva et al. [2022] presented an approach to disentangle supra-word meaning from lexical meaning in language models and showed that supra-word meaning is predictive of fMRI recordings in two language regions (anterior and posterior temporal lobes). Similar to the approach presented in Toneva et al. [2022], Oota et al. [2023e] disentangle both past and future context meaning from word meaning in language models and showed that past context is crucial in obtaining significant results while predicting MEG brain recordings. Caucheteux et al. [2021a] proposed a taxonomy to factorize the high-dimensional activations of language models into four combinatorial classes: lexical, compositional, syntactic, and semantic representations. They found that (1) Compositional representations recruit a more widespread cortical network than lexical ones and encompass the bilateral temporal, parietal, and prefrontal cortices. (2) Contrary to previous claims, syntax, and semantics are not associated with separated modules but appear to share a common and distributed neural substrate.

While previous works studied syntactic processing as captured through complexity measures (syntactic surprisal, node count, word length, and word frequency) [Zhang et al., 2020a, 2022a], very few have studied the syntactic representations [Caucheteux et al., 2021a, Reddy and Wehbe, 2021, Oota et al., 2023d]. Studying syntactic representations using fMRI is difficult because: (1) representing syntactic structure in an embedding space

is a non-trivial computational problem, and (2) the fMRI signal is noisy. To overcome these limitations, Reddy and Wehbe [2021] proposed syntactic structure embeddings that encode the syntactic information inherent in the natural text that subjects read in the scanner. The results reveal that syntactic structure-based features explain additional variance in the brain activity of various parts of the language system, even after controlling for complexity metrics that capture the processing load. While Reddy and Wehbe [2021] focused on constituency parsing, mainly including incremental top-down parsing, Oota et al. [2023d] leverage dependency information more systematically by learning the dependency representations using graph convolutional networks, using the four-step recipe as illustrated in Figure 3.12. The results reveal that constituency tree structure is better encoded in language regions such as bilateral temporal cortex (ATL and PTL) and MFG, while dependency structure is better encoded in AG and PCC language regions.

While previous studies focused on narrative English language stories and have shown that several brain regions are involved in building the hierarchical syntactic structure, a recent study in Zhang et al. [2022b] analyzes the neural basis of such structures between two diverse languages: Chinese and English. The results demonstrate that the brain may use different parsing strategies for different language structures to reduce the cognitive load.

#### NLP TASKS AND LINGUISTIC PROPERTIES IN LMS AND BRAINS

Understanding the reasons behind the observed similarities between language comprehension in language models and brains can lead to more insights into both systems. Further, the type of information in the finetuned language models that leads to high encoding accuracy needs to be clarified. It is unclear whether and how the two systems align in their information processing pipeline. Recent work Schwartz et al. [2019], Schrimpf et al. [2021b], Kumar et al. [2022], Goldstein et al. [2022], Aw and Toneva [2023], Merlin and Toneva [2022], Oota et al. [2022c, 2023c], Sun and Moens [2023], Sun et al. [2023], Loong Aw et al. [2023] addressed this question either by tuning the pretrained language model on downstream NLP tasks or inducing the brain relevant information into the language model. Several researchers have suggested that one contributor to the alignment is the LM's ability to predict the next word, with a positive relationship between next-word prediction ability and brain alignment across LMs [Schrimpf et al., 2021b, Goldstein et al., 2022]. However, more recent work shows no simple relationship exists, and language modeling loss is not a perfect predictor of brain alignment [Pasquiou et al., 2022, Antonello et al., 2021]. Schwartz et al. [2019] finetuned pretrained BERT model to predict brain activity and found that finetuned BERT has modified language representations to encode better the information relevant for predicting brain activity. Rather than finetuning the BERT model on brain data, Oota et al. [2022c] finetuned the BERT model on 10 GLUE (General Language Understanding Evaluation) [Wang et al., 2018] tasks to check whether task supervision leads to better encoding models to account for the brain's language representation. Oota et al. [2022c] found that using a finetuned BERT on downstream NLP tasks improved brain predictions. The results reveal that reading fMRI was best explained by Co-reference Resolution, NER (Named Entity Recognition), and shallow syntax parsing, and listening fMRI was best explained by paraphrasing, summarization, and NLI. Since full finetuning gen-

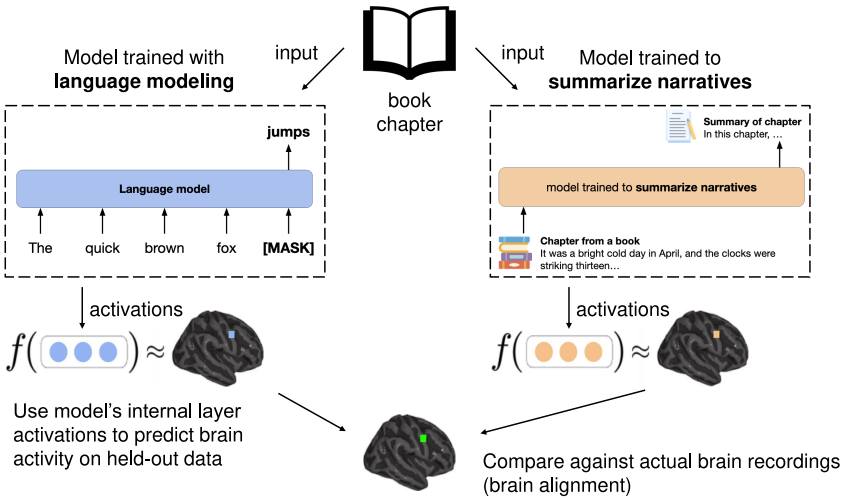



Figure 3.13: Comparison of brain recordings with language models trained on web corpora (Left) and language models trained on book stories (Right) [Aw and Toneva, 2023]. *This Figure is redrawn from Aw and Toneva [2023].*

erally updates the entire parameter space of the model, which has been proven to distort the pre trained features [Kumar et al., 2022], Sun and Moens [2023] explore prompt-tuning that generates representations that better account for the brain's language representations than finetuning. They find that prompt-tuning on tasks dealing with fine-grained concept meaning, including Word Sense Disambiguation and Co-reference Resolution, yields representations that are better at neural decoding than tuning on other tasks with both finetuning and prompt-tuning. Further, Sun et al. [2023] extended similar prompt-tuning to bridge the gap between the human brain and supervised DNN representations of the Chinese language. With the recent success of instruction-tuned large language models, Loong Aw et al. [2023] investigated the effect of instruction-tuning on large language models and alignment with the human brain's language representations. The results demonstrate that instruction-tuning large language models (LLMs) improves world knowledge representations and brain alignment. This suggests that mechanisms that encode world knowledge in LLMs also improve representational alignment to the human brain.

To investigate whether large language models with longer context are learning a deeper understanding of the text, Aw and Toneva [2023] used four pretrained large language models (BART, Longformer Encoder Decoder, BigBird, and LongT5) and also trained them to improve their narrative understanding, using the method detailed in Figure 3.13. They find that the improvements in brain alignment are larger for character names than for other discourse features, which indicates that these models are learning important narrative elements. However, it is not understood whether language models with the prediction of the next word are necessary for the observed brain alignment or simply sufficient, and whether there are other shared mechanisms or information that is similarly important. Merlin and Toneva [2022] proposed two perturbations to pretrained language models that, when used together, can control for the effects of next word prediction and word-level semantics on

the alignment with brain recordings. Specifically, they found improvements in alignment with brain recordings in two language processing regions–Inferior. Frontal Gyrus (IFG) and Angular Gyrus (AG)–are due to next-word prediction and word-level semantics. However, what linguistic information underlies the observed alignment between brains and language models was unclear. Recently, Oota et al. [2023c] tested the effect of a range of linguistic properties (surface, syntactic, and semantic) and found that eliminating each linguistic property significantly decreases brain alignment across all layers of BERT. Further, syntactic properties are more responsible and have the most significant effect on the trend of brain alignment across model layers. To further understand what aspects of linguistic stimuli contribute to ANN-to-brain similarity, Kauf et al. [2024] systematically manipulated the stimuli (i.e., perturbed sentences' word order, removed different subsets of words, or replaced sentences with other sentences of varying semantic similarity) and found that lexical semantic content rather than the sentence's syntactic form is primarily responsible for the DNN-to-brain similarity.

### 3.5.5 Auditory Encoding

To study auditory processing in the human brain, earlier studies focused on using hand-constructed features such as number of phonemes, MFCC (Mel Frequency Cepstral Coefficients), spectrotemporal modulations for auditory brain encoding [de Heer et al., 2017]. These basic acoustic features are part of a standard model of primary auditory cortex responses to sound encoding [Norman-Haignere and McDermott, 2018, Venezia et al., 2019, Mesgarani et al., 2014]. In several other studies, speech stimuli have predominantly been represented as text transcriptions [Huth et al., 2016], or basic features like phoneme rate and the sum of squared FFT (Fast Fourier Transform) coefficients have been employed when constructing encoding models [Pandey et al., 2022]. However, text transcription-based methods ignore the raw audio-sensory information completely. The basic speech feature engineering method misses the benefits of transfer learning from rigorously pretrained speech deep learning (DL) models. The benefits of using pretrained speech models include: (i) efficient contextual speech representations, (ii) enhanced accuracy and (iii) flexibility in fine-tuning.

#### Alignment between pretrained speech models and Brain

Recently, several researchers have used popular deep learning models such as APC [Chung et al., 2020], Wav2Vec2.0 [Baevski et al., 2020], HuBERT [Hsu et al., 2021], and Data2Vec [Baevski et al., 2022] for encoding speech stimuli. Millet et al. [2022] used a self-supervised learning model, Wav2Vec2.0, to learn latent representations of the speech waveform similar to those of the human brain. They find that the functional hierarchy of its transformer layers aligns with the cortical hierarchy of speech in the brain, and reveals the whole-brain organisation of speech processing with an unprecedented clarity. This means that the first transformer layers map onto the low-level auditory cortices (A1 and A2), the deeper layers map onto brain regions associated with higher-level processes (e.g. STS and IFG). Vaidya et al. [2022] present the first systematic study to bridge the gap between recent four self-supervised

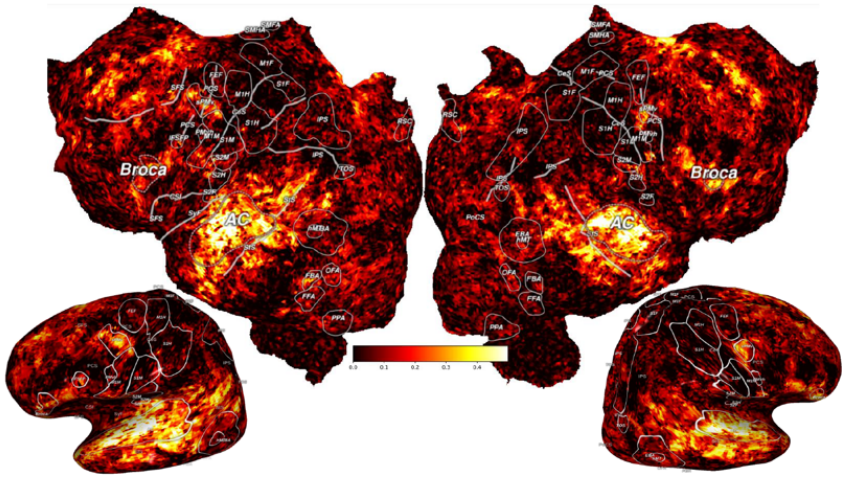



Figure 3.14: Brain prediction using self-supervised speech model: Data2Vec. The plot shows that speech-based models better predict early auditory cortex [Oota et al., 2023f].

speech representation methods (APC, Wav2Vec, Wav2Vec2.0, and HuBERT) and computational models of the human auditory system. Similar to Millet et al. [2022], they find that self-supervised speech models are the best models of auditory areas. Lower layers best modeled low-level areas, and upper-middle layers were most predictive of phonetic and semantic areas, while layer representations follow the accepted hierarchy of speech processing. Tuckute et al. [2023] analyzed 19 different speech models and find that some audio models derived in engineering contexts (model applications ranged from speech recognition and speech enhancement to audio captioning and audio source separation) produce poor predictions of auditory cortical responses, many task-optimized audio speech deep learning models outpredict a standard spectrotemporal model of the auditory cortex and exhibit hierarchical layer-region correspondence with auditory cortex. Further, Oota et al. [2023f] extended this analysis to more such deep learning based speech models (30 self-supervised speech models). They found that both language as well as auditory brain areas, are best aligned with intermediate layers in deep learning models. As shown in Figure 3.14, they also found that speech models better predict early auditory cortex than late language regions. Although pretrained speech models can understand broad aspects of speech in general, the implications of finetuning speech pretrained models for various speech-processing tasks for speech encoding in the brain, remains underexplored.

### Underlying speech properties in speech models and Brains

Understanding the reasons behind the observed similarities between speech processing in speech models and brains can lead to more insights into both systems. Recent work Oota et al. [2023g] has found that using a finetuned Wav2Vec2.0 leads to improved brain predictions. In particular, as shown in Figure 3.15, Oota et al. [2023g] build neural speech taskonomy models for brain encoding and aim to find speech-processing tasks that have the most explanatory capability of brain activation during naturalistic story listening exper-

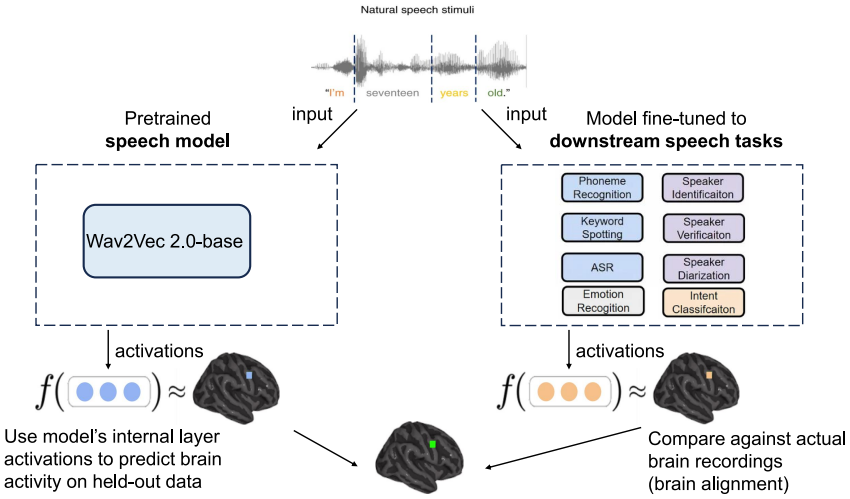



Figure 3.15: The pretrained Wav2Vec2.0 model and finetuned to eight different downstream speech tasks and their brain alignment [Oota et al., 2023g].

iments. They find that task-specific (Automated Speech Recognition (ASR), Entity Recognition (ER), Speaker Identification (SID) and Intent Classification (IC)) speech representations lead to a significant improvement in brain alignment compared to the pretrained Wav2Vec2.0 model for specific brain regions. Finetuning on ER, SID and IC leads to the best alignment for the early auditory cortex; finetuning on ASR provides the best encoding for the auditory associative cortex and language regions. Further, the layer-wise analysis of the effect of each speech task on the alignment with whole brain activity shows that the ASR task is better aligned in middle layers. A very recent study Oota et al. [2023a] reveals that in the context of brain listening, speech-based models outperform text-based language models in the auditory cortex. However, the alignment with the late language regions is significantly better for text-based than speech-based models in both during reading and listening. Specifically, low-level speech features such as phonological features explain the most variance for speech-based models in late language regions.

### 3.5.6 Visual Encoding

Similarly to language, in vision, early models focused on independent models of visual processing (object classification) using CNNs [Yamins et al., 2014]. Eickenberg et al. [2017] use CNNs as candidate models to model human brain activity during the viewing of natural images by constructing predictive models based on their different CNN layers and BOLD fMRI activations. They find that there are similarities between the computations of convolutional networks and cognitive vision at the beginning and at the end of the ventral stream object-recognition process. Cichy et al. [2016] further investigates the stages of human visual processing in both time (MEG recordings) and space (fMRI recordings). By comparing these findings with representations derived from deep neural networks (DNNs), the authors demonstrate that DNNs effectively encapsulate the sequential stages of human

visual processing. This encompasses the progression from early visual areas towards the specialized pathways of the dorsal and ventral streams, highlighting the DNN's capacity to mirror complex neural processes in both time and space. Despite the effectiveness of CNNs, it is difficult to draw specific inferences about neural information processing using CNN- derived representations from a generic object-classification CNN. Hence, Wang et al. [2019] built encoding models with individual feature spaces obtained from 21 computer vision tasks. One of the main findings is that features from 3D tasks, compared to those from 2D tasks, predict a distinct part of visual cortex. Recent efforts in visual encoding models, particularly self-supervised models (instance-prototype contrastive learning), operates by taking multiple samples over an image and projecting these through a deep convolutional neural network into a low-dimensional embeddings space [Konkle and Alvarez, 2022]. The results show that these self-supervised models achieve parity with the category-supervised models in accounting for the structure of brain responses. In a recent study by Matsuyama et al. [2023] on enhancing the precision of models for visual brain encoding, the research focused on two primary questions: (1) How does changing the size of the fMRI training dataset affect prediction accuracy? (2) How does the prediction accuracy across the visual cortex change with the size of the parameters in the vision models? The findings indicate that prediction accuracy improves with an increase in the training sample size, adhering to a scaling law. Similarly, an increase in the parameter size of the vision models also leads to improved prediction accuracy, following the same scaling law.

How can we push deeper CNN models to capture brain processing even more stringently? Continued architectural optimization on ImageNet alone no longer seems like a viable option. Instead of feed-forward deep CNN models, using shallow recurrence enabled better capture of temporal dynamics in the visual encoding models [Kubilius et al., 2019, Schrimpf et al., 2020]. Kubilius et al. [2019] proposed a shallow recurrent anatomical network, CORnet, that follows neuro-anatomy more closely than standard CNNs, and achieved the state-of-the-art results on the Brain-score benchmark [Schrimpf et al., 2020]. It has four computational areas, conceptualized as analogous to the ventral visual areas V1, V2, V4, and IT, and a linear category decoder that maps from the population of neurons in the model's last visual area to its behavioral choices.

### 3.5.7 Multimodal Brain Encoding

Recently Transformer-based multimodal models, which combine pairs of modalities such as language-vision, language-audio, and language-audio-vision, have emerged, offering rich aligned representations compared to single-modality models (i.e. text-only, audio-only or vision-only). Specifically, multimodal Transformers such as CLIP, LXMERT, VisaulBERT take both image and text stimuli as input and output a joint visio-linguistic representation. Since human brain perceives the environment using information from multiple modalities, examining the alignment between language and visual representations in the brain by training encoding models on fMRI responses, while extracting joint representations from multimodal models, can offer insights into the relationship between the two modalities. Oota et al. [2022f] experimented with multimodal models like CLIP, LXMERT, and VisualBERT and found VisualBERT better predict neural responses than vision-only models such as

CNNs and Image Transformers. Similarly, Wang et al. [2022] find that multimodal models like CLIP better predict neural responses in visual cortex as compared to previous vision-only models like CNNs. This is attributed to the fact that high-level human visual representations encompass semantics and the relational structure of the visual world, beyond object identity [Gauthier et al., 2003]. Dong and Toneva [2023] present a systematic approach to probe multi-modal video Transformer model by leveraging neuro-scientific evidence of multimodal information processing in the brain. The authors find that intermediate layers of a multimodal video transformer are better at predicting multimodal brain activity than other layers, indicating that the intermediate layers encode the most brain-related properties of the video stimuli. Recently, Tang et al. [2024] investigated a multimodal Transformer as the encoder architecture to extract the aligned concept representations for narrative stories and movies to model fMRI responses to naturalistic stories and movies, respectively. Since language and vision rely on similar concept representations, the authors perform a cross-modal experiment in which how well the language encoding models can predict movie-fMRI responses from narrative story features (story $\rightarrow$ movie) and how well the vision encoding models can predict narrative story-fMRI responses from movie features (movie $\rightarrow$ story). Overall, the authors find that cross-modality performance was higher for features extracted from multimodal transformers than for linearly aligned features extracted from unimodal transformers. A recent study by Nakagi et al. [2024], which used fMRI during the viewing of 8.3 hours of video content, and discovered distinct brain regions associated with different semantic levels, highlighting the significance of modeling various levels of semantic content simultaneously. The video material was meticulously annotated in five distinct semantic categories—speech, object, story, summary, and time/place—employing advanced large language models to derive latent representations. These representations were then used to predict fMRI brain activity across the various semantic categories. The authors discovered that the lack of unique variance for Summary and TimePlace is a notable insight, suggesting that merely incorporating these types of information into encoding analyses may not adequately capture higher-level semantic representations in the brain.

### 3.5.8 Key Takeaways

- **Alignment with Language Models:** Across several language models (like LSTMs and Transformers), middle layers of language models align well with brain language regions.
- **Semantic and Syntactic Processing:** Brain regions like the auditory cortex and Broca's area are involved in processing shorter contexts, while regions like the left temporo-parietal junction handle longer contexts.
- **Contextual Representations:** Contextual representations from language models improve the prediction of brain activity compared to traditional word embeddings.
- **Multimodal Integration:** Incorporating linguistic information with other modalities (like vision and auditory) can enhance understanding of how the brain processes complex stimuli.

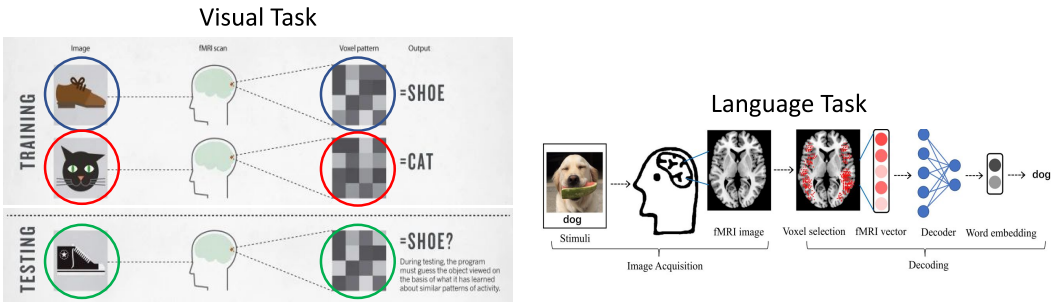



Figure 3.16: Scheme for Brain Decoding. Left: Image decoder [Smith, 2013], Right: Language Decoder [Wang et al., 2019]. *The left Figure is adapted from Smith [2013] and the right Figure is adapted from Wang et al. [2019].*

## 3.6 Brain Decoding

Decoding is the learning of the mapping from neural activations back to the stimulus domain. Figure 3.16 depicts the typical workflow for building an image/language decoder. Due to the inherent noise in brain recordings, obtaining reliable and robust representations or reconstructions of stimuli from these recordings continues to pose a significant challenge. Moreover, the recorded brain signals encompass not only the specific responses elicited by naturalistic stimuli but also include additional sources of noise arising from various cognitive, physiological processes, and scanner operations.

**Decoder Architectures**: In early decoding studies, the stimulus representation is decoded using typical ridge regression models trained on using the most informative voxels [Pereira et al., 2018, Sun et al., 2019, Oota et al., 2022d] or cortex specific voxels. In some cases, a fully connected layer [Beliy et al., 2019] or a multi-layered perceptron [Sun et al., 2019] has been used. In some studies, when decoding is modeled as multi-class classification, Gaussian Naïve Bayes [Singh et al., 2007, Just et al., 2010] and SVMs [Thirion et al., 2006] have also been used for decoding. However, these oversimplified methods often fall short of capturing the non-linear relationship between the stimulus and the neural responses. With the advent of recent generative AI models such as, large language models, multimodal models (CLIP), diffusion models (i.e. text-to-image, image-to-text, text-to-music, text-to-video, video-to-text), conditional generation of high-fidelity images, music and videos have become feasible. This exciting development leads to feasibility of reconstructing images, videos, speech, music and continuous language from brain activity. Figure 3.17 summarizes the literature related to various decoding solutions proposed in vision, auditory, and language domains. Table 3.4 aggregates the brain decoding literature along different stimulus domains such as textual, visual, and audio. The most common setting is to perform decoding to a vector representation using a stimuli of a single mode (visual, text or audio).

### 3.6.1 Linguistic Decoding

Initial brain decoding experiments studied the recovery of simple concrete nouns and verbs from fMRI brain activity [Nishimoto et al., 2011] where the subject watches either a picture

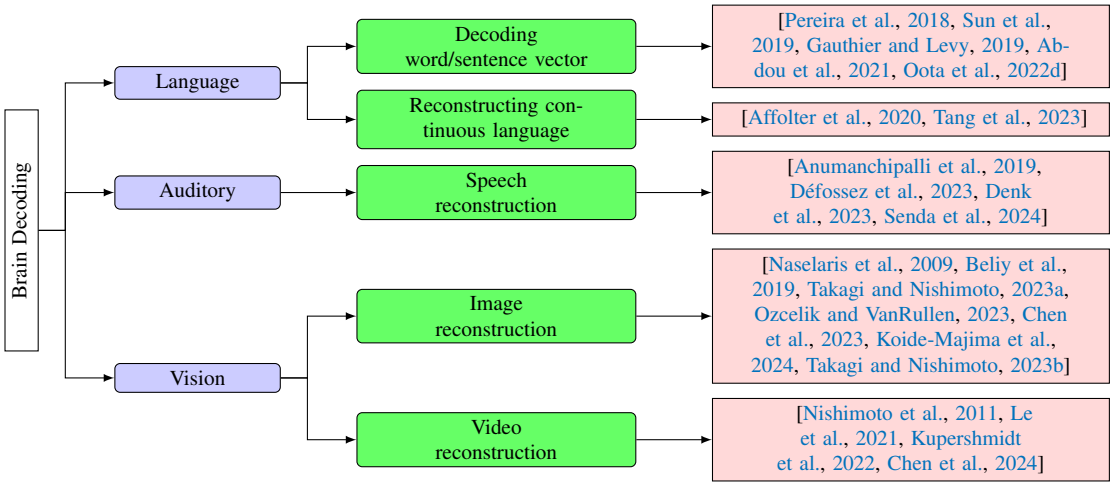



Figure 3.17: Categorization of Brain Decoding Studies.

| Stimuli | Authors | Dataset Type | Lang. | Stimulus Representations | \|S\| | Dataset |
|---|---|---|---|---|---|---|
| Text | Pereira et al. [2018] | fMRI | English | Word2Vec, GloVe, BERT | 17 | Pereira |
| | Wang et al. [2020b] | fMRI | English | BERT, RoBERTa | 6 | Pereira |
| | Oota et al. [2022d] | fMRI | English | GloVe, BERT, RoBERTa | 17 | Pereira |
| | Tang et al. [2023] | fMRI | English | GPT, finetuned GPT on Reddit comments and autobiographical stories | 7 | Moth Radio Hour |
| Visual | Beliy et al. [2019] | fMRI | | End-to-End Encoder-Decoder, Decoder-Encoder, AlexNet | 5 | Generic Object Decoding, ViM-1 |
| | Takagi and Nishimoto [2023a] | fMRI | | Latent Diffusion Model, CLIP | 4 | NSD |
| | Ozcelik and VanRullen [2023] | fMRI | | VDVAE, Latent Diffusion Model | 7 | NSD |
| | Chen et al. [2024] | fMRI | | Latent Diffusion Model, CLIP | 3 | HCP fMRI-Video-Dataset |
| Audio | Défossez et al. [2023] | MEG,EEG | English | MEL Spectrogram, Wav2Vec2.0 | 169 | MEG-MASC |
| | Gwilliams et al. [2023a] | MEG | English | Phonemes | 7 | MEG-MASC |
| | Denk et al. [2023] | fMRI | English | Music | 5 | MEG-MASC |

Table 3.4: Summary of Representative Brain Decoding Studies.

or a word. Sun et al. [2019] used several sentence representation models to associate brain activities with sentence stimulus, and found InferSent to perform the best. More work has focused on decoding the text passages instead of individual words [Wehbe et al., 2014]. Some studies have focused on multimodal stimuli based decoding where the goal is still to decode the text representation vector. For example, Pereira et al. [2018] trained the decoder on imaging data of individual concepts, and showed that it can decode semantic vector representations from imaging data of sentences about a wide variety of both concrete and abstract topics from two separate datasets. Further, Oota et al. [2022d] propose two novel brain decoding setups: (1) multi-view decoding (MVD) and (2) cross-view decoding (CVD). In MVD, the goal is to build an MV decoder that can take brain recordings for any view (picture, sentence, or word cloud) as input and predict the concept. In CVD, the goal is to train a model which takes brain recordings for one view as input and decodes a semantic vector representation of another view. Specifically, they study practically useful CVD tasks like image captioning, image tagging, keyword extraction, and sentence formation.

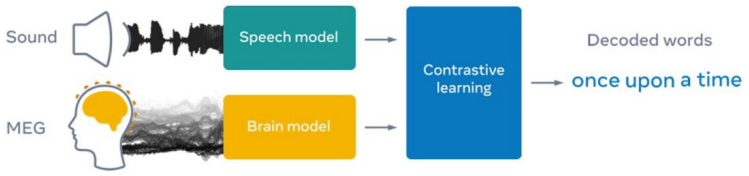



Figure 3.18: CLIP-MEG pipeline to align MEG activity onto pretrained speech embeddings [Défossez et al., 2023]. *The Figure is adapted from Défossez et al. [2023].*

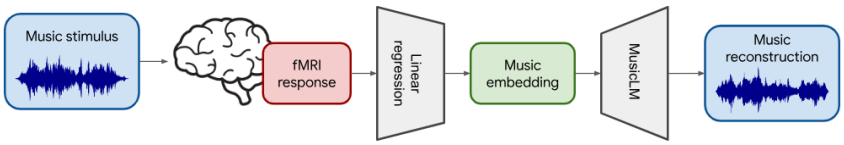



Figure 3.19: Brain2Music decoding pipeline [Denk et al., 2023]. *The Figure is adapted from Denk et al. [2023].*

To understand application of Transformer models for decoding better, Gauthier and Levy [2019] finetuned a pretrained BERT on a variety of Natural Language Understanding (NLU) tasks to find tasks that lead to improvements in brain-decoding performance. They find that tasks which produce syntax-light representations (representations extracted from a language model trained on randomly shuffled words from corpus samples, thereby eliminating all first-order cues to syntactic structure) yield significant improvements in brain decoding performance. This primary occurs because a significant portion (but not all) of the syntactic information initially represented in the baseline BERT model gets eliminated during training on the scrambled language modeling tasks.

With the recent development of large language models, rather than decoding stimuli vector representations, some studies have attempted to reconstruct words [Affolter et al., 2020], and continuous language [Tang et al., 2023] from fMRI brain activity.

### 3.6.2 Auditory Decoding

With the recent advancements of self-supervised speech models and generative AI models, recent studies have largely targeted reconstructing speech/music from brain recordings [Défossez et al., 2023, Denk et al., 2023, Senda et al., 2024]. As shown in Figure 3.18, [Défossez et al., 2023] proposed a CLIP-MEG pipeline to align MEG activity onto pretrained speech embeddings and generate speech from a stream of MEG signals. Unlike other methods which are experimented with on narrative speech, Denk et al. [2023] introduce a method for reconstructing music from fMRI brain activity, as shown in Figure 3.19.

### 3.6.3 Visual Decoding

A number of methods have been proposed for reconstructing a visual stimulus from brain recordings. Here, we initially address image reconstruction from brain recordings, followed by a discussion on video reconstruction.

### Image Reconstruction

Before the success of recent generative AI models, researchers have used deep-learning models and algorithms, including generative adversarial networks (GANs) and self-supervised learning models trained on a large number of naturalistic images [Du et al., 2020, Beliy et al., 2019, Fang et al., 2020, Gaziv et al., 2022, Lin et al., 2022]. For instance, Beliy et al. [2019] designed a separable autoencoder that enables self-supervised learning in fMRI and images to increase training data. Mind Reader [Lin et al., 2022] encoded fMRI signals into a pre-aligned vision-language latent space and used StyleGAN2 [Karras et al., 2020] for image generation. These methods generate more plausible and semantically meaningful images. Several other studies focused on reconstructing personal imagined experiences [Berezutskaya et al., 2020] or application-based decoding like using brain activity scanned during a picture-based mechanical engineering task to predict individuals' physics/engineering exam results [Cetron et al., 2019] and reflecting whether current thoughts are detailed, correspond to the past or future, are verbal or in images [Smallwood and Schooler, 2015].

With the recent success of CLIP and Diffusion models, deep generative models have been gaining attention to generate high-resolution images with high semantic fidelity [Takagi and Nishimoto, 2023b, Chen et al., 2023, Scotti et al., 2023, Benchetrit et al., 2023, Song et al., 2023]. Takagi and Nishimoto [2023b] proposed a method for image reconstruction from fMRI using Stable Diffusion [Rombach et al., 2022], as shown in Figure 3.20 (left). Their approach involves decoding brain activities to text descriptions and converting them to natural images using Stable Diffusion. Based on a similar philosophy, using a Stable Diffusion model as a generative prior and the pretrained fMRI features as conditions, Chen et al. [2023] reconstructed high-fidelity images with high semantic correspondence to the groundtruth stimuli, as shown in Figure 3.20 (right). Scotti et al. [2023] proposed a MindEye that can map fMRI brain activity to any high dimensional multimodal latent space, like CLIP image space, enabling image reconstruction using generative models that accept embeddings from this latent space. Different from previous studies, BrainCLIP framework was introduced by Liu et al. [2023] to align fMRI patterns with different modalities (especially from visual and textual modalities) through cross-modal contrastive loss. All these studies have been limited to 2D visual representations. A recent work Gao et al. [2023] aims to extend the scope of fMRI decoding to 3D representations. Specifically, Gao et al. [2023] introduce Recon3DMind, a groundbreaking task focused on reconstructing 3D visuals from fMRI signals.

Lastly, recent image reconstruction studies have focused on other non-invasive brain recordings such as MEG and EEG rather than fMRI signals. Benchetrit et al. [2023] proposed a CLIP-MEG pipeline to align MEG activity onto pretrained visual embeddings and generate images from a stream of MEG signals. Similarly, Song et al. [2023] proposed a CLIP-EEG pipeline to align these two modalities (image and EEG encoders to extract features from paired image stimuli and EEG responses) by constraining their similarity.

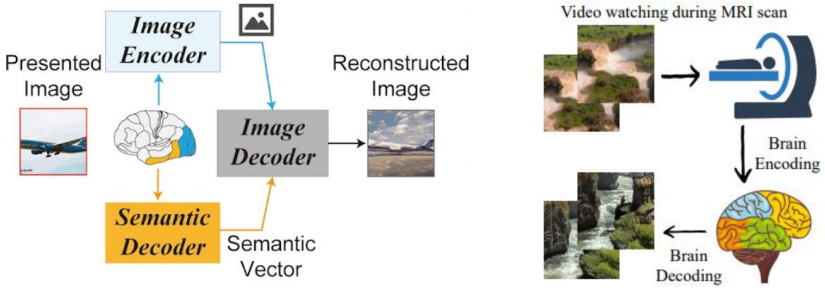



Figure 3.20: Image reconstruction from fMRI using Stable Diffusion. *The Left Figure is adapted from Takagi and Nishimoto [2023b] and the Right Figure is adapted from Chen et al. [2024].*

### Video Reconstruction

Unlike static natural images, human visual cortex can process a continuous, diverse flow of scenes, motions, and objects. To recover dynamic visual experience, the challenge lies in the nature of fMRI, which measures blood oxygenation level dependent (BOLD) signals and captures snapshots of brain activity every few seconds. Similar to image reconstruction works, Chen et al. [2024] present MinD-Video, a two-module pipeline (i.e. CLIP module followed by latent stable diffusion) designed to bridge the gap between image and video brain decoding.

## 3.7 Conclusion, Limitations, and Future Trends

In this paper, we surveyed important naturalistic brain datasets, stimulus representations, brain encoding and brain decoding methods across different modalities. A glimpse of how deep learning solutions throw light on putative brain computations is given. We hope that this systematic organization of recent ideas proposed in the field of cognitive computational neuroscience provides a comprehensive summary to the readers.

The insights from recent studies in brain encoding and decoding have far-reaching implications for the fields of AI engineering, neuroscience, and the interpretability of models—some with immediate effects, others with long-term impact.

**AI engineering:** The recent brain encoding studies most immediately fits in with the neuro-AI research direction that specifically investigates the relationship between representations in the brain and representations learned by powerful neural network models. This direction has gained recent traction, especially in the domain of language, vision, speech processing, thanks to advancements in language models [Schrimpf et al., 2021b, Goldstein et al., 2022], vision models [Schrimpf et al., 2020] and speech models [Tuckute et al., 2023, Oota et al., 2023f]. Furthermore, several recent works most immediately contributes to this line of research by understanding the reasons for the observed similarity in more depth [Merlin and Toneva, 2022, Oota et al., 2023c, Kauf et al., 2024, Sarch et al., 2024, Oota et al., 2023a]. Overall, these studies provide valuable insights for selecting features, enhancing transfer learning, and aiding in the creation of AI architectures that are cognitively plausible.

**Computational Modeling in Neuroscience:** Researchers have started viewing language models as useful *model organisms* for human language processing [Toneva, 2021] since they implement a language system in a way that may be very different from the human brain, but may nonetheless offer insights into the linguistic tasks and computational processes that are sufficient or insufficient to solve them [McCloskey, 1991, Baroni, 2020]. These brain encoding studies enables cognitive neuroscientists to have more control over using language models as model organisms of language processing. This approach can also be extended to visual and speech processing, where models in these domains serve as analogous organisms for investigation.

**Model Interpretability:** In the long-term, we aspire for these studies on brain encoding and decoding to enhance another research direction that utilizes brain signals to interpret the information processed by neural network models [Toneva and Wehbe, 2019, Aw and Toneva, 2023, Wang et al., 2019, Sarch et al., 2024]. Ultimately, our goal is to comprehend the essential and adequate underlying characteristics that result in a meaningful correlation between brain recordings and deep neural network models.

### 3.7.1 Future Trends

Some of the future areas of work in this field are as follows.

**Bridging the Gap: Enhancing Deep Neural Network Models for Deeper Insights into Auditory, Language and Visual Processing** While significant progress has been made in understanding text-based models, understanding the similarity in information processing between visual, speech and multimodal models versus natural brain systems remains an open area. For instance, Oota et al. [2023a] demonstrates that speech-based language models lack brain relevant semantics in language regions. Therefore, enhancing speech-based language models to align more closely with text-based models could provide valuable insights into language and auditory processing, given that speech is the most ancient form of human language. This suggests a promising direction for future research, aiming to bridge the gap between artificial intelligence models and the complex, multifaceted processes of human cognition.

**Advancing Multimodal Decoding: The Next Leap in Deep Learning Accuracy** Decoding actual multimodal stimuli has become increasingly feasible due to recent advancements in deep learning models dedicated to generation tasks. However, there is still a significant need for further research to enhance the accuracy of these models. This involves not only refining the algorithms and architectures used but also improving the quality and diversity of the datasets on which these models are trained. Advancements in computational power, algorithmic efficiency, and innovative training methodologies are critical for pushing the boundaries of what is possible in multimodal decoding, aiming to achieve more precise, reliable, and nuanced interpretations of complex stimuli.

**Mapping the Mind: The Effects of Brain Damage on Cognitive Capabilities** We need deeper understanding of the degree to which damage to different regions of the human brain could lead to the degradation of cognitive skills. This exploration requires detailed mapping of cognitive functions to specific brain areas, taking into account the brain's complex network of connections. Studies should investigate not only the immediate effects of brain

damage on cognitive skills but also the brain's capacity for reorganization and compensation over time. Ultimately, the goal is to translate these research findings into practical applications, such as more effective cognitive rehabilitation techniques and assistive technologies that can improve the quality of life for individuals with brain injuries.

**Towards Human-Like Understanding in ANNs: Integrating Self-Supervised Learning and Brain-Inspired Architectures** How can we train artificial neural networks in novel self-supervised ways such that they compose word meanings or comprehend images and speech like a human brain? Can we model the hierarchical and modular organization of the brain in neural network architectures?. This involves creating networks that reflect the brain's organization, from low-level feature detection to high-level semantic processing, allowing for the integration of information across different modalities. Moreover, how might we integrate dynamic learning strategies, such as curriculum learning, which progressively introduces more intricate tasks to the model? This method emulates the way humans naturally progress from understanding straightforward to more complex ideas over time.

**Bridging the Language Gap in Brain-NLP Research: The Need for Multilingual Exploration** Current brain-NLP research relies on brain recordings collected from individuals who speak English as their primary language. Additionally, these studies utilize experimental stimuli that are presented in the English language. As a result, all current neuro-AI studies predominantly leverages language models and neural models that have been trained extensively on English text data and brain responses elicited by text or speech in English. However, it is essential to acknowledge the potential variability in our study outcomes when extrapolated to languages other than English. The intricate interplay between language-specific nuances and neural responses may introduce distinctions in the results. Therefore, it becomes imperative for future research endeavors to delve into this aspect further and investigate how these factors might influence the generalizability of our findings across diverse linguistic contexts.

In addition to the current advancements, there are several potential avenues for future exploration in the intersection of neuroscience and artificial intelligence. One such direction involves leveraging enhanced understanding of neuroscience to propose modifications to existing artificial neural network architectures, with the aim of enhancing their robustness and accuracy. Furthermore, an intriguing area for further investigation lies in understanding the brain activity of multilingual, multi-scriptal individuals when processing stimuli in their second language (L2) or script. It remains unclear whether observed brain activity reflects the processing of L2 or the active suppression of their first language (L1) while focusing on L2. This ambiguity underscores the need for further research, particularly in the realm of multilingual multimodal stimuli, to elucidate the underlying mechanisms at play.

We hope that this survey motivates research along the above directions.

# 4 Long Short-Term Memory of Language Models for Predicting Brain Activation During Listening to Stories

Several popular sequence-based and pre-trained language models are successful for text-driven prediction of brain activations. However, these models still lack long-term memory plausibility (i.e., how they deal with long-term dependencies and contextual information) as well as insights into the underlying neural substrate mechanisms. This paper studies the influence of context representations of different language models, such as sequence-based models: Long Short-Term Memory Networks (LSTMs) and ELMo, and a pre-trained Transformer language model (Longformer). In particular, we study how the internal hidden representations align with the brain activity observed via fMRI when the subjects listen to several narrative stories. We use brain imaging recordings of subjects listening to narrative stories to interpret word and sequence embeddings. We further investigate how the representations of language model layers reveal better semantic context during listening. Experiments across all language model representations provide the following cognitive insights: (i) the representations of LSTM cell states are better aligned with brain recordings than LSTM hidden states, the cell state activity can represent more long-term information, (ii) the representations of ELMo and Longformer display an excellent predictive performance across brain regions for listening stimuli; (iii) Posterior Medial Cortex (PMC), Temporo-Parieto-Occipital junction (TPOJ), and Dorsal Frontal Lobe (DFL) have higher correlation versus Early Auditory (EAC) and Auditory Association Cortex (AAC).

This chapter has been finalized based on our previously published paper at $44^{th}$ Annual Meeting of the Cognitive Science Society conference (July 2022, Toronto, Canada). [Oota et al., 2022b].

## 4.1 Introduction

In the past decade, artificial neural networks have witnessed remarkable insights in the computational neuroscience community in understanding how the brain performs stimulus perception (1) given various forms of sensory inputs like visual processing in object recognition tasks [Yamins et al., 2014, Cadieu et al., 2014, Eickenberg et al., 2017], or (ii) by studying higher-level cognition like language processing [Gauthier and Levy, 2019, Schrimpf et al., 2021b, Schwartz et al., 2019]. This line of work, namely brain encoding, aims to construct neural brain activity given an input stimulus.

Sentence comprehension has been studied using fMRI for a while Constable et al. [2004]. Some studies have looked into the modeling of language comprehension, e.g., how sequence-based language models such as echo-state networks (ESN) [Hinaut and Dominey, 2013] or long short-term memory networks (LSTM) [Jain and Huth, 2018, Variengien and Hinaut, 2020] encode syntactic structures and contextual information. Moreover, Jain and Huth [2018] used LSTMs to get the context representation of sentences (with a next word prediction task) and then used this representation to predict fMRI data.

Some works studied the LSTM capacity of representing long-term information [Karpathy et al., 2015] and its ability to model working memory [O'Reilly and Frank, 2006]. However, there still needs to be more investigation of the long-term memory cognitive plausibility of LSTM and its link to fMRI data. In this paper, we open the black box of LSTM to look at particular LSTM activations: the cell and hidden states. This can give more insights into longer-term and shorter-term information. Indeed, the *cell state* mechanism has been introduced in the original LSTM paper [Hochreiter and Schmidhuber, 1997] to keep the error gradient of backpropagation constant over long-time scales. Thus, its activity can represent more long-term information than the *hidden state* of the LSTM. We also investigate how the pretrained bi-directional sequence embedding language model ELMo [Peters et al., 2018] handles the longer context and interprets the LSTM layers representations that better predict brain activity.

Recently, the researchers studied how the representations from Transformer [Vaswani et al., 2017] based language models such as BERT [Devlin et al., 2019] and RoBERTa [Liu et al., 2019] could directly predict fMRI data. Interestingly, such Transformer-based neural representations are very effective for brain encoding as well [Schrimpf et al., 2021b]. On the other hand, Gauthier and Levy [2019] fine-tune a pretrained BERT model on multiple natural language processing tasks to find tasks best correlated with high *decoding* performance. In recent works, Caucheteux et al. [2021a], Antonello et al. [2021] interpret the representations of the Transformer model (GPT-2 [Radford et al., 2019]) by disentangling the high-dimensional Transformer representations of language models into four combinatorial classes: lexical, compositional, syntactic, and semantic representations to explore which class is highly associated with language cortical ROIs. However, due to their self-attention operation, these models cannot handle the long-term dependencies (sequence length is fixed to 512 words). To overcome this limitation, recently, Beltagy et al. [2020] introduced *Longformer*, making it easy to process documents of thousands of tokens or longer and combining local windowed attention with global attention.

This paper reveals insights about the association between fMRI voxel activations and representations of diverse language models: LSTM, ELMo, and Longformer. The predictive power of language model specific representations with brain activation is ascertained by (1) using ridge regression on such representations and predicting activations and (2) computing famous metrics like 2V2 accuracy and Pearson correlation between actual and predicted activations.

Specifically, we make the following contributions in this paper. (1) Given a language model pretrained on corpora by handling long-term dependencies, we propose the problem of finding which of these are the most predictive of fMRI brain activity for listening tasks. (2) The investigation of the long-term context of language model results reveals that ELMo and Longformer representations display better correlation during narrative story listening. (3) We also investigate the internal memory representations of LSTM (cell state and hidden state) and derive interesting insights that the cell state representations yield better performance than hidden state representations.

## 4.2 Methodology

### 4.2.1 Brain Imaging Dataset

**Narratives-Pieman (Listening to Stories)** The "Narratives" collection aggregates a variety of fMRI datasets collected while human subjects listened to naturalistic spoken stories. The Narratives dataset that includes 345 subjects, 891 functional scans, and 27 diverse stories of varying duration totaling ∼4.6 hours of unique stimuli (∼43,000 words) was proposed in Nastase et al. [2020b]. Similar to earlier works Caucheteux et al. [2021b], we analyze data from 82 subjects listening to the story titled 'PieMan' with 259 TRs (repetition time)[1]. A TR is the length of time between corresponding consecutive points in fMRI: here it is 1.5 sec. We list number of voxels per ROI (Region of Interest) in this dataset in Table 4.1. We use the multi-modal parcellation of the human cerebral cortex (Glasser Atlas: consists of 180 ROIs in each hemisphere) to display the brain maps [Glasser et al., 2016], since the Narratives dataset contains annotations tied to this atlas. The data covers ten brain ROIs, i.e., Left hemisphere (L), and Right hemisphere (R) for each of the following: (i) early auditory cortex (EAC: A1, LBelt, MBelt, PBelt, and R1) which plays a key role for sound perception since it represents one of the first cortical processing stations for sounds; (ii) auditory association cortex (AAC: A4, A5, STSdp, STSda, STSvp, STSva, STGa, and TA2) which is concerned with the memory and classification of sounds; (iii) posterior medial cortex (PMC: POS1, POS2, v23ab, d23ab, 31pv, 31pd, 7m) which has been implicated in tasks as diverse as attention, memory, spatial navigation, emotion, self-relevance detection, and reward evaluation; (iv) the temporo parieto occipital junction (TPOJ: TPOJ1, TPOJ2, TPOJ3, STV, PSL) which is a complex brain territory heavily involved in several high-level neurological functions, such as language, visuo-spatial recognition, writing, reading, symbol processing, calculation, self-processing, working memory, musical memory, and face and object recognition; and (v) the dorsal frontal lobe (DFL: L_55b, SFL, L_44, L_45, IFJA,

[1] 282 TRs (before preprocessing) and 259 TRs (after preprocessing).

IFSP) which covers the aspects of pragmatic processing such as discourse management, integration of prosody, interpretation of nonliteral meanings, inference making, ambiguity resolution, and error repair. These five brain ROIs (EAC, AAC, TPOJ, DFL, and PMC) span a cortical hierarchy supporting language and narrative comprehension [Huth et al., 2016, Baldassano et al., 2017].

| ROIs→ | EAC | | AAC | | PMC | | TPOJ | | DFL | |
|---|---|---|---|---|---|---|---|---|---|---|
| | LH | RH | LH | RH | LH | RH | LH | RH | LH | RH |
| # Voxels | 808 | 638 | 1420 | 1493 | 1198 | 1204 | 847 | 1188 | 1061 | 875 |

Table 4.1: # Voxels in each ROI in the Narratives Dataset. LH - Left Hemisphere. RH - Right Hemisphere. Pieman has 82 subjects.

## 4.3 Language Models

To explore how and where contextual word features are represented in the brain when listening to stories, we extract internal hidden representations from four language models: GloVe, Random LSTM & LSTM, ELMo (obtaining context-dependent word embeddings), and popular pretrained Transformer-based language model (Longformer). This approach aims to better understand the contribution of different stimulus features to the brain alignment with language models. Our main objective is to compare the correlation between each model dense hidden representations and human cognitive process. In this paper, we train fMRI encoding models using Ridge regression on stimuli representations obtained using these four language models. The main goal of each fMRI encoder model is to predict brain responses associated with each brain region given stimuli. In all cases, we train a ridge regression model per subject separately. Following the literature on brain encoding [Caucheteux et al., 2021b, Toneva et al., 2020], we choose to use a simple ridge regression model instead of more popular regression models like Bootstrap [Tikhonov et al., 1977] or Banded models [la Tour et al., 2022]. For instance, Deniz et al. [2019] used a banded ridge regression-based model that combines all feature groups into one encoding model to map brain activity. We plan to explore more such models as part of future work in brain encoding. Here, our main is objective to investigate the influence of context representations of different language models and their alignment with language regions of the brain.

### 4.3.1 LSTM

First, we train an LSTM [Hochreiter and Schmidhuber, 1997] network to predict the probability of the next word as a function of the history of previous words. The weights of LSTMs are learned using the error back-propagation through time (BPTT). To make the association between encoded stimuli from LSTM's internal components and fMRI brain activity, we do the folllowing: (i) At the time step t, we use vector $a_t$ to represent the internal neurons of encoded stimuli in LSTM. In this paper, $a_t$ may be hidden state vector ($h_t$) and cell state vector ($c_t$). (ii) In order to map the stimuli encoded vector $a_t$ of LSTM and brain activity

at the $t$-th time step ($Y_t$), we define a simple linear model, ridge regression, to predict the brain activity ($\hat{Y}_t$) from $a_t$, as discussed in the ridge regression section.

### 4.3.2 Pretrained text Transformer: Longformer

Longformer [Beltagy et al., 2020] builds on BERT's language masking strategy and supports long document generative sequence-to-sequence tasks. We use the pretrained Longformer model with a local attention mechanism, where the default window size is set to 5. To obtain the stimuli representation, we follow previous work to extract the hidden-state representations from each layer of these language models, given a fixed-length input length [Toneva and Wehbe, 2019]. To extract the stimulus features from these pretrained models, we constrained the tokenizer to use a maximum context of 5 words. Given the constrained context length, each word is successively input to the network with at most $C$ previous tokens. For instance, given a story of $M$ words and considering the context length of 5, while the third word's vector is computed by presenting ($w_1$, $w_2$, $w_3$) as input to the network, the last word's vector $w_M$ is computed by presenting the network with ($w_{M-20}$, ..., $w_M$). The pretrained Transformer model outputs token representations at different layers. We use the #tokens $\times$ 768 dimension vector obtained from each hidden layer to obtain word-level representations from each pretrained Transformer language model.

Additionally, we varied the context lengths (10, 20, 50, 100, 500, and 1000) and measured their brain alignment. This approach reveals how brain alignment improves when longer input contexts are provided.

### 4.3.3 Linear Probing of Language Models

Here, we direct extract representations Random LSTM, LSTM, ELMo, and Longformer networks (i.e. no finetuning) to directly predict brain activities, for two reasons. First, the dimension of the fMRI voxels varies among different subjects and across different ROIs. Therefore, it is not convenient to design a universal neural network architecture for generating outputs of different dimensions. Second, the goal of this research is not to improve the performance of language models in predicting fMRI. We want to explore linear mappings between particular features of language models states and neural activities in the auditory and language brain ROIs. Namely, we look at (i) the characteristics of hidden state vectors ($h_t$) and artificial memory vectors ($c_t$) in both LSTMs and Random LSTMs, and (ii) local context vectors obtained from performance-optimized deep neural network models (ELMo and Longformer). Therefore, we avoid any possible supervision from the fMRI data when training LSTM and Random LSTM language models.

### 4.3.4 Comparison to Other Language Models

We compare the LSTM model with the Longformer and several other pretrained language models: Random LSTM, ELMo, and GloVe. To enable a fair comparison of encoding model performance, we use the same context length for extracting word representations from these language models.

**LSTM Training:** We experimented with one layer of LSTM to perform the next word prediction. In our next word prediction, we first split each story in half (of 27 stories); we designate the first half as the training set and the second half as the test set. The model is implemented in Keras with TensorFlow backend [Abadi et al., 2016] with cross-entropy as loss, Adam optimizer [Kingma, 2014], the number epochs set to 100, the batch size is of 64, applied dropout with a keep-probability of 0.2, and tried LSTM with hidden state size is set to 100, the dimensionality of word embeddings is set to 100. The other hyper-parameters are learning rate (0.01), and maximum sequence length is set to 5.
**Random LSTM:** We use a random LSTM model where the LSTM weights are randomly initialized and kept frozen. We use the output and cell state vectors at each time step to perform fMRI encoding. The configuration details of Random LSTM are the same as that original LSTM model.
**ELMo:** ELMo (Embeddings from Language Models) is a successful NLP framework developed by the AllenNLP [Peters et al., 2018] group. Unlike earlier embeddings, ELMo embeddings represent words in a contextual fashion using a bidirectional LSTM model. We perform the downsampling of word embeddings (see Section 4.3.5) to obtain the TR-level representations.
**GloVe:** based word vectors (each word is a 300-dimension vector) [Pennington et al., 2014], and the downsample of word embeddings (see Section 4.3.5) results in context vector in each time step.

### 4.3.5 Downsampling

Since the rate of fMRI data acquisition (TR = 1.5sec) was lower than the rate at which the text stimulus was presented to the subjects, several words fall under the same TR in a single acquisition. Hence, we match the stimulus acquisition rate to fMRI data recording by downsampling the stimulus features using a 3-lobed Lanczos filter. After downsampling, we obtain word-embeddings corresponding to each TR.

### 4.3.6 TR Alignment

To account for the slowness of the hemodynamic response, we model the HRF using a finite response filter (FIR) per voxel and for each subject separately with various temporal delays. For instance, in Narratives listening, a temporal delay of 1 TR corresponds to 1.5 secs , and 5 TRs translate to a delay of 7.5 secs. Overall, the FIR filters were implemented by concatenating feature vectors that had been delayed by various delays.

## 4.4 Experimental Setup

### 4.4.1 Voxelwise Encoding Model

We trained a ridge regression based encoding model to predict the fMRI brain activity associated with the semantic vector representation obtained from each language model: randLSTM (hidden state, cell state), LSTM (hidden state, cell state), GloVe, ELMo, and Long-

former. Each voxel value is predicted using a separate ridge regression model. Formally, at the time step (t), we encode the stimuli as $X_t \in \mathbb{R}^{N \times D}$ and brain region voxels $Y_t \in \mathbb{R}^{N \times V}$, where $N$ denotes the number of training examples, $D$ denotes the dimension of input stimuli representation after TR alignment, and $V$ denotes the number of voxels in a particular region.
**Hyper-parameter Setting**: We used sklearn's ridge-regression with default parameters, 5-fold cross-validation, Stochastic-Average-Gradient Descent Optimizer, Huggingface for Longformer, MSE loss function, and L2-decay ($\lambda$):1.0. We used Word-Piece tokenizer for the Longformer model and Spacy-tokenizer for the GloVe and ELMo models.

### 4.4.2 Model Prediction Across Whole Brain

To determine the significant voxel predictions across the whole-brain, we ran the permutation tests where we shuffled the true responses 5000 times, computed the Pearson correlation scores, and finally obtained the FDR corrected p-values for the whole brain results using both Longformer and ELMo. We set the correlation score of voxels to zero if the p-value of the correlation obtained from the permutation test is above the significance threshold (p$>$ 0.05, FDR corrected).

### 4.4.3 Evaluation Metrics

We evaluate our models using popular brain encoding evaluation metrics described in the following. Given a subject and a brain region, let $N$ be the number of samples. Let $\{Y_i\}_{i=1}^{N}$ and $\{\hat{Y}_i\}_{i=1}^{N}$ denote the actual and predicted voxel value vectors for the $i^{th}$ sample. Thus, $Y \in R^{N \times V}$ and $\hat{Y} \in R^{N \times V}$ where $V$ is the number of voxels in that region.
**2V2 Accuracy** is computed as follows.

$$\text{2V2Acc} = \frac{1}{N_{C_2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} I[\{cosD(Y_i, \hat{Y}_i) + cosD(Y_j, \hat{Y}_j)\} < \{cosD(Y_i, \hat{Y}_j) + cosD(Y_j, \hat{Y}_i)\}]$$

where $cosD$ is the cosine distance function. $I[c]$ is an indicator function such that $I[c] = 1$ if $c$ is true, else it is 0. The higher the 2V2 accuracy, the better.
**Pearson Correlation (PC)** is computed as PC=$\frac{1}{N} \sum_{i=1}^{n} corr[Y_i, \hat{Y}_i]$ where corr is the correlation function.

## 4.5 Results

In order to assess the performance of the fMRI encoder models learned using the representations from a variety of language models, we computed the 2V2 accuracy and Pearson correlation coefficient between the predicted and true responses across various ROIs for the listening (Narratives-Pieman) dataset (Fig. 4.1).

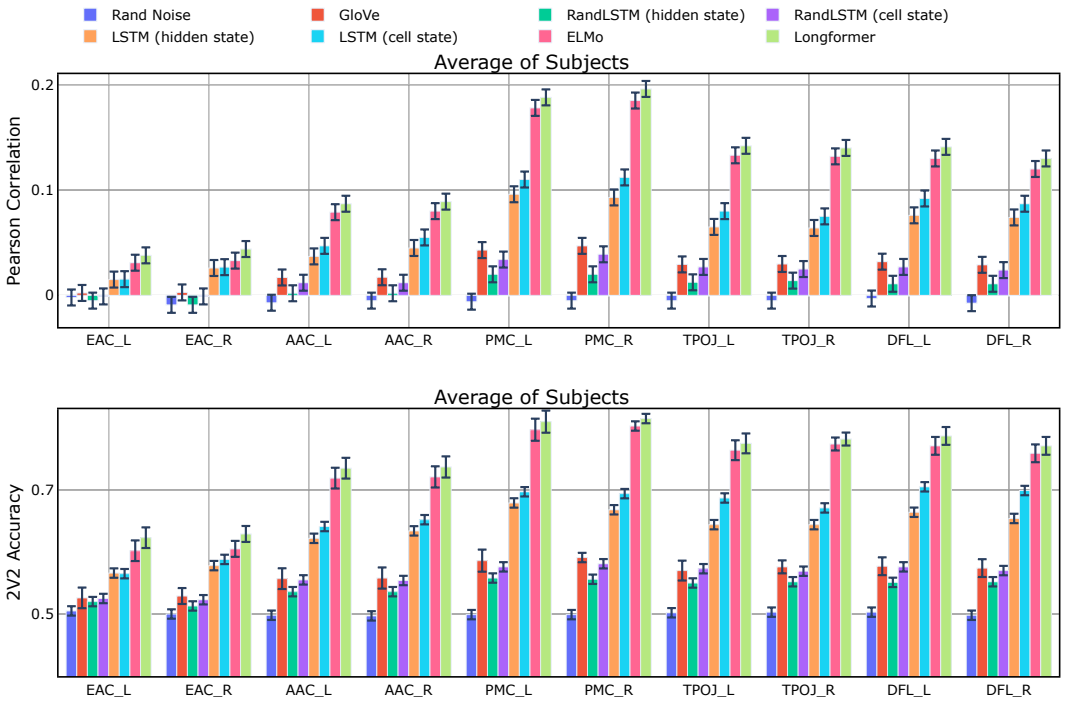



Figure 4.1: Pearson correlation coefficient (top figure) and 2V2 Accuracy (bottom figure) between predicted and true responses across different brain regions using a variety of language models (for Narratives-Pieman dataset). Results are averaged across all participants. ELMo and Longformer are the best. *Rand Noise* stands for a "Random noise vector". *(hidden state)* stands for the "short-term memory of internal state of the LSTM".

### 4.5.1 Encoding performance of Language Models

From Fig. 4.1, we observe that the profiles of performance show low scores in the early auditory cortex (EAC) and auditory association cortex (AAC); average scores in TPOJ and DFL; and superior scores in PMC. This aligns with the known language hierarchy for spoken language understanding [Huth et al., 2016, Baldassano et al., 2017, Nastase et al., 2020a]. Language models ELMo, and Longformer yield better performance in predicting the brain responses than the LSTM model across all the ROIs. These Pearson correlation ($\rho$) results are comparatively better to those obtained using the pretrained GPT2 model in Caucheteux et al. [2021a] ($\rho$ ranging from $0.02 - 0.06$). As shown in Fig. 4.1, our method obtains more than 3 times higher correlations ($\rho$ ranging from $0.02 - 0.19$)[2]. The main reason is that the Longformer is designed to process documents of thousands of tokens or longer sequences while GPT-2 models are unable to handle the long-term dependencies (sequence length is fixed to 512 words). Also, the narrative dataset consists of longer documents (more than

[2]we do not apply on the same number of subjects and/or same amount of stories than in Caucheteux et al. [2021a]. However, we tested with few other stories such as Lucy and Slumlord, and our results (higher correlations) show similar trends.

| Models compared | EAC | AAC | PMC | TPOJ | DFL |
|---|---|---|---|---|---|
| Longformer vs ELMo | 0.521 | 0.271 | 0.168 | 0.054 | 0.356 |
| LSTM (Cell state vs hidden state) | 0.991 | 0.177 | 0.0357* | 0.0038* | 0.158 |
| Longformer vs LSTM (cell state) | 0.048* | 0.0008* | 0.00002* | 0.00003* | 0.0015* |
| ELMo vs LSTM (cell state) | 0.372 | 0.003* | 0.00004* | 0.00006* | 0.0049* |

Table 4.2: p-values obtained using *post hoc* pairwise comparisons for the three best models (+ LSTM hidden state).

2000 words in one story); the traditional transformer models consider the context up to 512 words, whereas Longformer handles even longer documents.

Further, from Fig. 4.1, we see that the bilateral posterior medial cortex (PMC) associated with higher language function exhibits a higher correlation among all the brain ROIs. ROIs, including bilateral TPOJ and bilateral DFL, yield higher correlations with the ELMo and Longformer, which is in line with the language processing hierarchy in the human brain. Finally, across all regions, Rand Noise vector and Rand LSTM models have worse correlation compared to LSTM and other language models. In summary, different and distinct language model features seem to be related to the encoding performance in listening tasks.

In order to estimate the statistical significance of the performance differences, we performed one-way ANOVA on the mean correlation values for the subjects across the language models (GloVe, LSTM (cell state), LSTM (hidden state), ELMo, and Longformer) for the five brain ROIs. The main effect of the ANOVA test was significant for all the ROIs with $p \leq 10^{-2}$ with confidence 95%. Further, *post hoc* pairwise comparisons [Ruxton and Beauchamp, 2008] confirmed the visual observations that on both 2V2 accuracy and Pearson correlation measures, tasks such as ELMo and Longformer performed significantly better compared to other models, as shown in Table 4.2.

### 4.5.2 LSTM: Effects of Hidden State vs Cell State Vectors

In order to explore how LSTM hidden units learn to encode the long-term and short-term memory information and the interaction between the two types of working memories, we compare the encoding performance between representations of hidden state and cell state vectors. Fig. 4.1 showcases the fMRI encoding performance of both RandLSTM and LSTM models where the cell state representations (long term-memory vector) yield better performance than hidden state representations (short-term memory). This supports the cognitive plausibility of the LSTM cell architecture. Besides, the performance of GloVe and RandLSTM models have significantly equal performance, indicating that semantic context is missing in these models.

### 4.5.3 Which ELMo layers perform better encoding?

We investigate how the performance of ELMo changes at different layers (Embedding layer, LSTM layer-1, and LSTM layer-2), as they are provided in different contexts. The results

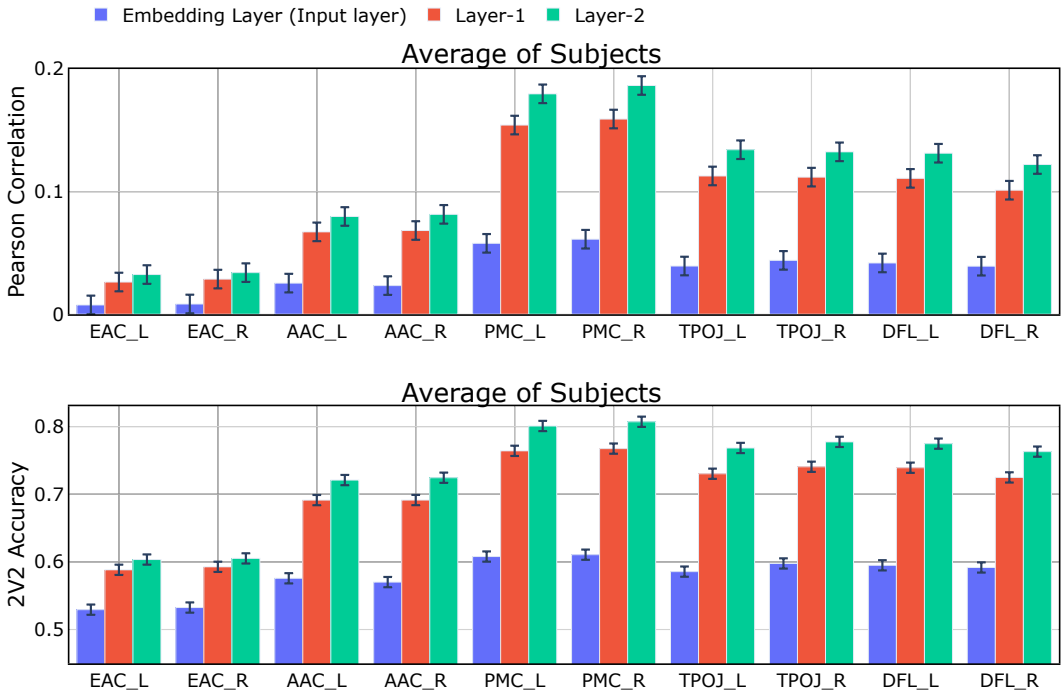



Figure 4.2: ELMo layers: Pearson correlation coefficient (top) and 2V2 Accuracy (bottom) between predicted and true responses across different brain regions using layers of ELMo model. Results are averaged across all participants.

are shown in Fig. 4.2. From Fig. 4.2, we observe that layer-2 displays better 2v2 accuracy and Pearson correlation score compared to other layers. We further observe that the layer 1 show a sharp increase in performance compared to embedding layer in the context of narrative story listening.

### 4.5.4 Which Longformer layers perform better encoding?

Given the hierarchical processing of language information across the Transformer layers, we further examine how these Transformer layers encode fMRI brain activity using encoder layers of Longformer. We present the layer-wise encoding performance results across brain ROIs in Fig. 4.3. We observe that in all the layers, intermediate layers (6 to 8) perform the best for narrative listening, followed by a decrease in performance. Overall, for the Longformer model, the best alignment with fMRI is observed in the middle layers, as noted in prior studies. [Jain and Huth, 2018, Toneva and Wehbe, 2019].

### 4.5.5 Cognitive Insights

We further analyse in more detail the prediction performance of the encoder model trained on sub ROIs for the ELMo and Longformer in Fig. 4.4. In the *EAC*, the sub ROI *pbelt*

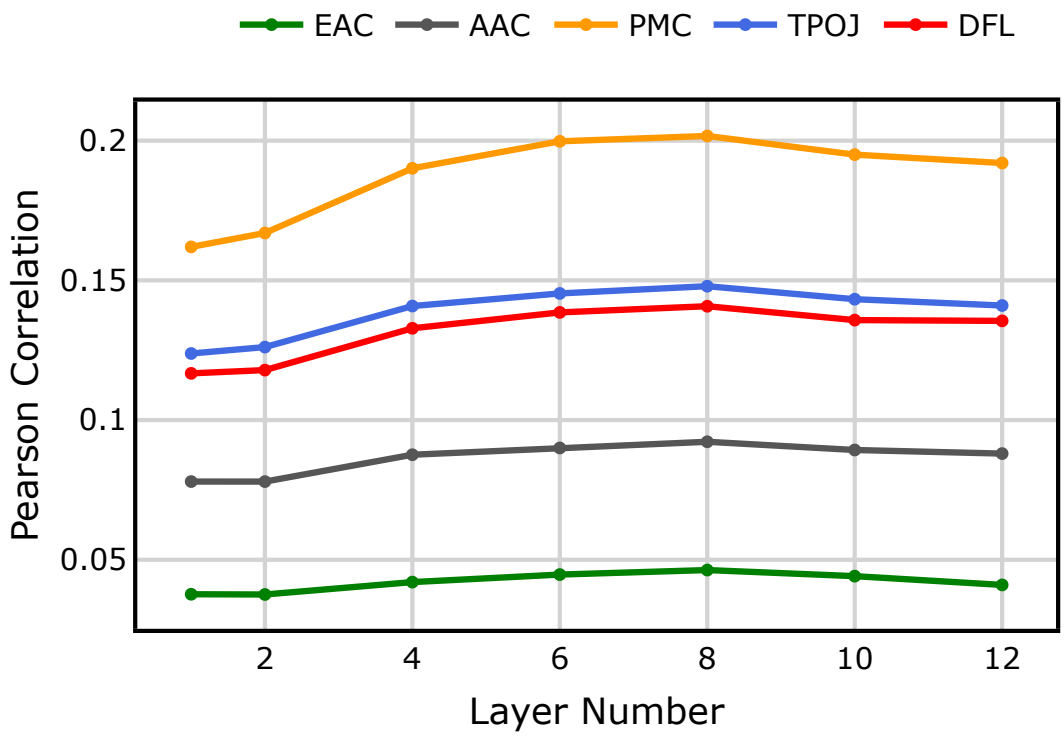



Figure 4.3: Longformer layers: Pearson correlation coefficient between predicted and true responses across different brain regions (different color lines) using Longformer. Results are averaged across all participants. Middle layers (6 and 8) show best correlation.

*(parabelt)* display higher Pearson correlation among other sub ROIs, and it is adjacent to the lateral belt on the exposed surface of the superior temporal gyrus (*STG*). Also, the *pbelt* area represent the next level in the auditory hierarchy and mainly concerned with memory or decision-making. Similarly in the *AAC*, the sub ROIs such as *A4*, *A5*, *stsvp*, and *stsda* yield better correlation, and these ROIs shares the primary medial and posterior borders with *TPOJ* [Trumpp et al., 2013]. Further, there is evidence that these sub ROIs of the AAC process perceptual and conceptual acoustic sounds during auditory stories and social interaction tasks [Glasser et al., 2016]. It can be observed that sub ROIs such as *Pos1* and *Pos2* have a higher Pearson correlation than other sub ROIs of the *PMC* region. Both *sfl* and *l55b* display a higher correlation among all the sub ROIs for the *DFL* ROI. However, all the sub ROIs in the *TPOJ* yield higher correlation, as shown in Fig. 4.4. The control and attention ROIs in the posterior cingulate cortex (for ex., *POS1* in *PMC*), together with the superior frontal language region (*sfl* in *DFL*) and *TPOJ*, are part of the language network associated with narrative comprehension [Nastase et al., 2020a]: it is encouraging to see that both ELMo and Longformer also relate to semantic analysis of the ongoing narrative because they obtain best performance, showing that capturing longer-term context is important.

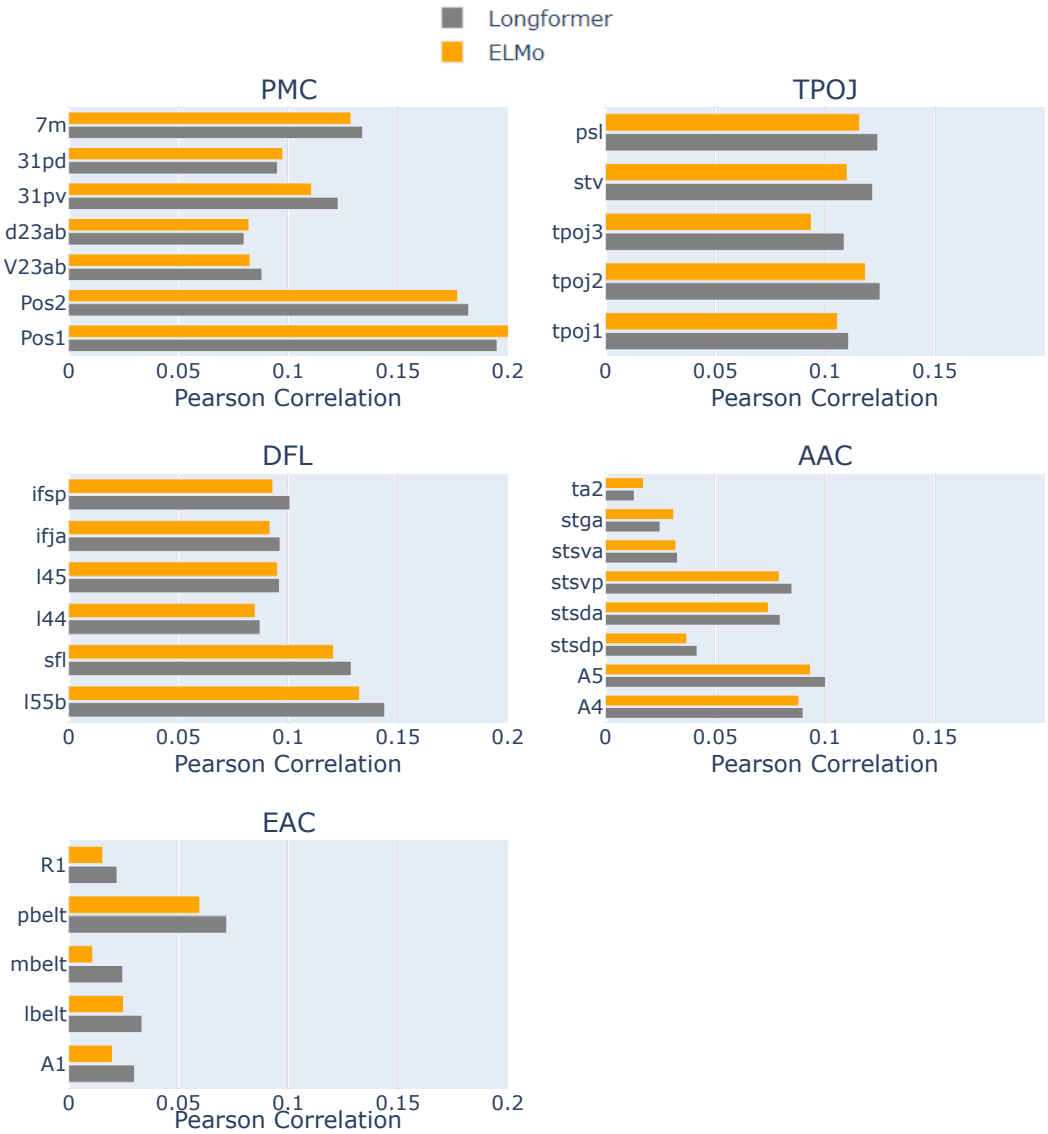



Figure 4.4: Pearson correlation coefficient between predicted and true responses across different sub ROIs of the Language Network using ELMo and Longformer. Results are averaged across all participants.

### 4.5.6 Longformer: Effect of Context Lengths

Fig. 4.5 displays the average Pearson correlation across several language ROIs by varying the context lengths from 5 to 1000 and observing their brain alignment. We make the following observations from Fig. 4.5: (i) Brain alignment improves with the increase in context lengths, specifically, we observe higher correlation when we provide longer input contexts (50-100). (ii) There is a decrease in brain alignment for the higher context length of 1000. This implies that the brain processes longer context information effectively up to

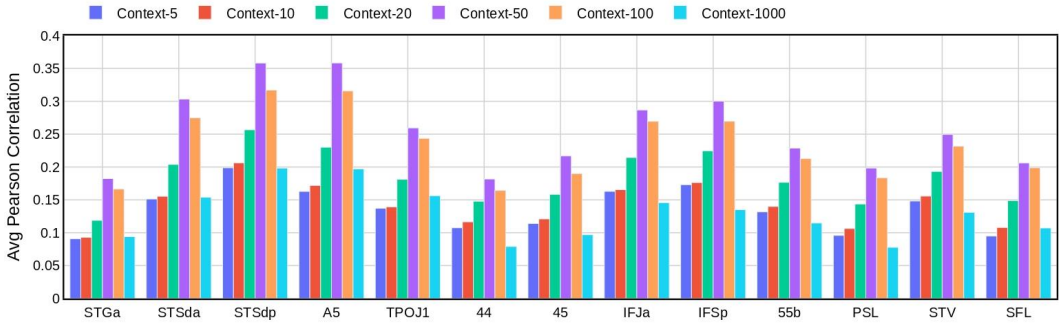



Figure 4.5: Average Pearson correlation across several language regions of interest (ROIs) for varying context lengths (5 to 1000) using the Longformer model.

context lengths of 100 but cannot process longer semantic information efficiently beyond that.

### 4.5.7 Brain maps for whole brain predictions

The whole-brain prediction Pearson correlation for all the voxels using ELMo, Longformer and LSTM representations is shown in Fig. 4.6. In the **listening task**, we observe from Fig. 4.6 that Longformer displays higher correlation values for many voxels than ELMo. From Fig. 4.6, we see that ROIs such as EAC and AAC have a lower percentage of voxels with a higher correlation compared to PMC and TPOJ brain ROIs (higher percentage of voxels with higher correlation).

## 4.6 Discussion & Conclusion

This paper studied the long-term memory plausibility of language models for brain encoding. We observe that building individual encoding models and interpreting the internal representations among models can provide a more in-depth understanding of the neural representation of language information. Our experiments on the Narrative listening stories dataset lead to several interesting findings.

(1) Pretrained language models, where the contextual word representations (such as in ELMo and Longformer) are used, are better predictors of voxel activations across language regions than static or sequential models. (2) In LSTM, the cell state representations (long term memory vector) yield better encoding performance than hidden state representations; thus, internal dynamics of LSTMs seem to have more cognitively plausible activations than classically studied LSTM activations. (3) We used different layers of ELMo and Longformer, where higher layers display better correlation for ELMo while intermediate layers show superior performance for Longformer. (4) The control and attention ROIs in the posterior cingulate cortex, together with the superior frontal language region (sfl in DFL) and TPOJ, are part of the language network associated with narrative comprehension. (5) Although text-based language models are pretrained on text data, the representations of these

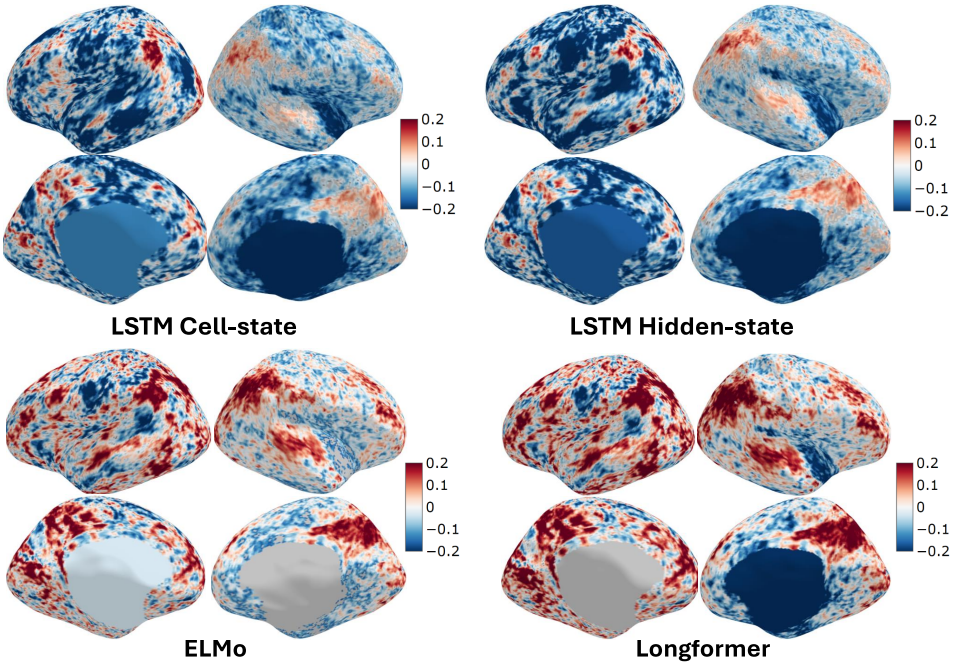



Figure 4.6: BrainMaps: Whole-brain prediction correlation using representations of ELMo, Longformer, and LSTM models, averaged across participants of Narratives-Pieman dataset.

models result in higher degree of correlation in the early auditory and auditory association cortex.

Our study resulted in three main conclusions: (1) we use human brain recordings to evaluate how well representations from language models (static vs. recurrent vs. pretrained) can predict representations of the human brain during language comprehension. (2) Richer representations learned from language models, designed to integrate longer contexts, have improved alignment with human brain activity. (3) Pretrained language models significantly predict brain language regions that are thought to underlie language comprehension.

Overall, advancements in language models, particularly in handling longer-context lengths, are crucial for better prediction of brain activation patterns and narrative memory retention.

### 4.6.1 Limitations

We believe that using popular regression models like Bootstrap [Tikhonov et al., 1977] or Banded models [la Tour et al., 2022] instead of simple ridge regression could lead to further exciting insights, such as linking internal model mechanisms more directly to brain activations. However, achieving this would require more computing resources, as these models combine all features to build a single encoding model.

In this paper, we only tested the contextual representations extracted from one longer-context model, Longformer. More popular models such as Long-T5, LED, BART, and Memory Transformer could be useful for a deeper understanding of their brain alignment.

Additionally, the importance of word order in longer models is underexplored. Hence, conducting perturbations on input stimuli and verifying whether longer models still capture efficient representations and their brain alignment across language regions is necessary.

# 5 Optimal Hemodynamic Response Function delays are different for syntax and semantics: a language model study of naturalistic story listening

Recent brain encoding studies highlight the potential for natural language processing models to improve our understanding of language processing in the brain. Simultaneously, naturalistic fMRI datasets are becoming increasingly available and present even further avenues for understanding the alignment between brains and language models. However, current brain encoding studies on language use constant hemodynamic response function (HRF) delay, literature studies on schizophrenia disorders report that it was worth to look at different HRF delays. This poses a question of how different language regions truly process syntax and semantics. Hence, for a deeper understanding of this brain alignment, it is important to understand the correspondence between the detailed processing of different language regions at different HRF delays in both the human brain and language models. In this work, we present a systematic study of the brain alignment across 8 HRF delays (ranging from 1.5 secs to 12 secs) related to the language model representations and observe how these delays affects the alignment with fMRI brain recordings obtained while participants listened to or read a story. In particular, we explore the relative importance of syntactic information (i.e. syntactic embeddings based on constituency trees) versus semantic information, using open-source, text-based language models such as BERT, GPT-2 and Llama-2, along with the speech-based language model Wav2vec2.0. Using different HRF delays, we find that early processing of syntactic information in frontal and temporal lobes, while semantic processing occurs in later delays in the angular gyrus. We further investigate different context lengths and find that longer context may play a significant role in higher HRF delays. Text-based models align strongly with language regions, whereas speech-based models align with early auditory cortex. These findings suggest that the decomposition of representations into different linguistic features enables a fine-grained understanding of brain language processing across various delays, cautioning against solely relying on speech-based models for late processing regions, paving the way for more personalized and effective approaches in both linguistic and clinical applications.

This chapter has been finalized based on our ongoing work, which we plan to submit to a journal.

## 5.1 Introduction

The increasing availability of naturalistic fMRI datasets and large-scale neural models can enable a better understanding of the brain's response to natural stimuli. Just in the last few years, researchers have shown that brain responses of people comprehending language can be predicted well by text-based language models [Wehbe et al., 2014, Jain and Huth, 2018, Toneva and Wehbe, 2019, Deniz et al., 2019, Caucheteux and King, 2020, Schrimpf et al., 2021b, Caucheteux et al., 2021a, Toneva et al., 2022, Oota et al., 2022c, Antonello et al., 2021, Aw and Toneva, 2023, Merlin and Toneva, 2022] and speech-based language models [Millet et al., 2022, Vaidya et al., 2022, Oota et al., 2023f,a]. Understanding the reasons behind the observed similarities between language comprehension in machines and brains can lead to more insight into both systems.

In the context of non-invasive fMRI (functional Magnetic Resonance Imaging) brain recordings, language processing within the brain is impacted by delays in the Hemodynamic Response Function (HRF). This refers to the time lag between the neural fMRI brain activity associated with language comprehension and the subsequent hemodynamic response, which is the blood flow change accompanying neural activity. The exact timing of this response can vary across individuals and brain regions.

While existing studies on the alignment between language comprehension and the brain have been observed at constant hemodynamic response function (HRF) delay (around 7.5 to 8 seconds), there is still ongoing exploration into how language and the brain's processing mechanisms synchronize when faced with different HRF delays [Jain and Huth, 2018, Jain et al., 2020, Toneva and Wehbe, 2019, Deniz et al., 2019, Toneva et al., 2022, Aw and Toneva, 2023, Oota et al., 2022c, 2023c]. Further, the existing studies have mainly built brain encoding models by considering a fixed HRF delay and analyzing how different regions of interest (ROIs) involved in language processing influence the semantic and syntactic aspects of information processing in the brain [Jain and Huth, 2018, Jain et al., 2020, Toneva and Wehbe, 2019, Caucheteux et al., 2021a, Toneva et al., 2022, Merlin and Toneva, 2022, Aw and Toneva, 2023, Oota et al., 2022c, 2023c]. Table 5.1 summarizes current brain encoding studies with a fixed HRF delay. Therefore, the interplay between HRF delays and language processing is an area of investigation, aiming to comprehend how neural activity related to language tasks aligns with the subsequent hemodynamic response and how this alignment may differ under varying conditions of HRF delays.

More recently, researchers have begun to study the alignment of these brain language regions with the language models, using a multi-timescale modeling approach employing LSTM to study how the human brain process neural language contains information at multiple timescales, ranging from phonemes to narratives [Jain et al., 2020]. In particular, the authors learn language model-derived representations at different timescales to perform brain encoding. However, the conclusions drawn in the Jain et al. [2020] study are constrained by the existence of much more effective language representations than those offered by LSTM networks. Also, efforts have been made towards ascertaining the language structures embedded exclusively within widely used Transformer-based language models, namely BERT and GPT-2 [Conneau et al., 2018, Rogers et al., 2020, Jawahar et al., 2019, Mohebbi et al., 2021]. This exploration indicates that Transformer-based language models learn effective

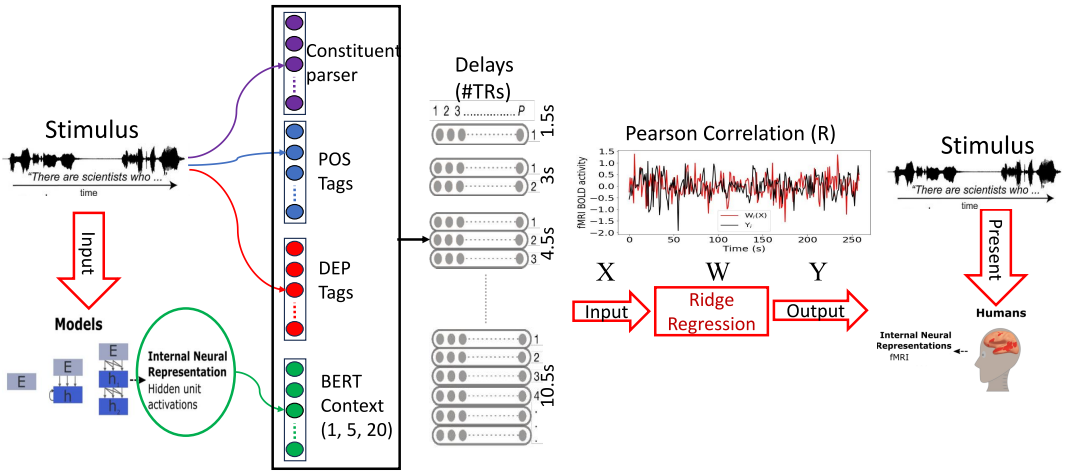



Figure 5.1: Brain Encoding Schema: Methodology for studying the alignment of neural language models with fMRI brain activity. In this study, we test the effect of basic syntax features, syntactic embeddings from constituent parse trees and contextual representations from pretrained language models on the alignment between these features and brain recordings across different HRF delays.

representations. Taken together, these findings open up the question of how the effectiveness of contextual representations influences the observed convergence between brains and language models across distinct HRF delays.

Prior brain encoding studies have revealed differences in how the brain processes language comprehension during reading and listening to a naturalistic stimulus [Jain and Huth, 2018, Toneva and Wehbe, 2019]. Particularly, these studies have focused on interpreting various contextual word representations from neural networks using brain recordings. These findings demonstrated that specific language regions in the medial parietal cortex, prefrontal cortex, and inferior temporal cortex are biased toward encoding contextualized representations. In contrast, certain areas in the superior temporal cortex and the temporoparietal junction do not prefer contextualized information. However, it remains unclear to what extent brain language regions, with various contextual word representations, contribute to encoding language structure across distinct HRF delays.

In the realm of syntactic processing, Matchin and Hickok [2020] illustrate the phonological and semantic networks that interact with syntactic systems, including word-level syntactic relationships, hierarchical parse trees, and complex syntax. They demonstrate the transformation of sequences of auditory phonological representations into hierarchical structures, encompassing both entity knowledge and event knowledge associated with different language regions. Recent research has explored how brain process syntax structure for linguistic stimuli by generating pure syntactic embeddings from parse trees (constituent and dependency parsers) [Reddy and Wehbe, 2021, Oota et al., 2023d, Zhang et al., 2022a], or disentangle the language model representations by controlling the syntactic information [Caucheteux et al., 2021a]. These findings reveal that both syntax and semantic

| Stimuli | Authors | Type | Lang. | Stimulus Representations | \|S\| | Dataset | Delays |
|---|---|---|---|---|---|---|---|
| Text | Jain et al. [2020] | fMRI | English | LSTM | 6 | Moth-Radio-Hour | 8secs (4 TRs) |
| | Jain and Huth [2018] | fMRI | English | LSTM | 6 | Moth-Radio-Hour | 8secs (4 TRs) |
| | Caucheteux et al. [2021a] | fMRI | English | GPT-2 | 345 | Narratives | 7.5secs (5 TRs) |
| | Reddy and Wehbe [2021] | fMRI | English | Syntax Parsers, BERT | 8 | Harry-Potter | 8secs (4 TRs) |
| | Merlin and Toneva [2022] | fMRI | English | GPT2 | 8 | Harry-Potter | 8secs (4 TRs) |
| | Aw and Toneva [2023] | fMRI | English | BART, LongT5, LED | 8 | Harry-Potter | 8secs (4TRs) |
| | Antonello et al. [2021] | fMRI | English | 100 Language Models | 7 | Moth-Radio-Hor | 8secs (4 TRs) |
| | Oota et al. [2023c] | fMRI | English | BERT and Probing Tasks | 18 | Narratives 21$^{st}$-Year | 9secs (6 TRs) |
| | Oota et al. [2023a] | fMRI | English | BERT, GPT-2, Wav2Vec2.0 | 6 | Moth-radio-hour | 12secs (6 TRs) |

Table 5.1: Summary of Brain Encoding Studies with constant HRF delays. Here, |S| denotes number of participants.

information are distributed across the brain language regions. However, all these studies perform their analysis at fixed delay. Hence, the study of joint syntactic processing between the brains and the language models, check for ways to improve across delays and further explain brain language processing.

Our work aims to examine how the intricate processing of diverse language regions at varying HRF delays in the human brain corresponds with Transformer based language models. For various HRF delays and context lengths, we analyze the impact on the alignment between brain recordings and language model representations (see Fig. 5.1 for a schematic). This analysis sheds light on how this alignment is influenced and how it relates to specific language regions. For this work, we focus on three popular language models–BERT [Devlin et al., 2019], GPT-2 [Radford et al., 2019] and Llama-2 [Touvron et al., 2023], and one speech-based language model Wav2vec2.0 [Baevski et al., 2020]–which have been studied extensively in natural language processing (NLP) and has been previously shown to significantly predict fMRI recordings of people processing language [Toneva and Wehbe, 2019, Schrimpf et al., 2021b, Antonello et al., 2021, Oota et al., 2022c, 2023c]. We use a popular dataset of fMRI recordings that are openly available such as Narratives [Nastase et al., 2020b] correspond to 22 subjects listening to a natural story.

Overall, our main contributions are as follows: (i) We perform an extensive study on evaluating linguistic brain encoding, examining various hemodynamic response function (HRF) delays. We specifically focus on both basic word-level syntactic, basic speech features, syntactic embeddings from constituent parsers and contextual representations derived from pretrained language models, considering different context lengths in the process. (ii) We show that syntactic information is early encoded in the brain, followed by semantic information, as the HRF delay increases. (iii) Detailed region and sub-region analysis reveal that longer context may impact the observed neural activity at different brain language regions, specifically at higher HRF delays. For instance, BERT with a context length of 20 has higher brain predictivity within language regions such as AG, IFG, ATL, and PTL, specifically for higher delays ranging from 9 to 12 seconds.

We will make all code available upon publication.

## 5.2 Dataset Curation

**Brain Imaging Dataset** We analyzed a publicly available Brain Imaging dataset, Narratives-Tunneling [Nastase et al., 2020b]. The fMRI data in these datasets were acquired from human participants actively engaged in *listening* to the Tunnel Under the World naturalistic story (in the Narratives-Tunneling dataset). In the Narratives tunneling, the dataset includes 22 subjects, and each functional scan was obtained at a time repetition of 1.5 secs (TR=1.5 sec), amounting to 1023 TRs. We use the multi-modal parcellation of the human cerebral cortex (Glasser Atlas: consists of 180 ROIs in each hemisphere) to display the brain maps [Glasser et al., 2016], since the Narratives dataset contains annotations tied to this atlas. The dataset is made available freely without restrictions by Nastase et al. [2020b]. The data covers six language brain regions of interest (ROIs) of the left hemisphere with the following subdivisions: (i) angular gyrus (AG: PFm, PGs, PGi, TPOJ2, and TPOJ3); (ii) anterior temporal lobe (ATL: STSda, STSva, STGa, TE1a, TE2a, TGv, and TGd); (iii) posterior temporal lobe (PTL: A4, A5, STSdp, STSvp, PSL, STV, TPOJ1); (iv) inferior frontal gyrus (IFG: 44, 45, IFJa, IFSp); (v) middle frontal gyrus (MFG: 55b); (vi) inferior frontal gyrus orbital (IFGOrb: a47r, p47r, a9-46v) [Baker et al., 2018, Milton et al., 2021, Desai et al., 2022].

**Estimating Participant Noise Ceiling** To account for the intrinsic noise in biological measurements and obtain a more accurate estimate of the model's performance, we estimate the noise ceiling approach proposed by Schrimpf et al. [2021b]. This is achieved by estimating the amount of brain response in one subject that can be predicted using only the data from a combination of other subjects, using an encoding model. We choose to use a kernel ridge regression model [1] as the encoding model. We first subsampled-the data with $n$ participants into all possible combinations of $s$ participants for all $s \in [2, n]$ (e.g. 2, 3, 4, ..., 22 for $n$=22). For each subsample, we then designated a random participant as the target that we attempt to predict from the remaining $s - 1$ participants (e.g., predict 1 subject from 1 (other) subject, 1 from 2 subjects, ..., 1 from 22, to obtain a mean score for each sensor in that subsample. As suggested in Schrimpf et al. [2021b], we extrapolate to infinitely many humans and thus to obtain the highest possible (most conservative) estimate.

Note that the estimated participant noise ceiling estimate is based on the assumption of a perfect model, which may not always be the case in real-world scenarios. Nonetheless, this approach can put the model's performance in a useful perspective. We report the estimated noise ceiling performance for each participant in the *supplementary* Figure 5.11.

### 5.2.1 Feature Representations

To simultaneously test both syntax and semantic representations and their alignment with brain recordings, we extract both word-level and auditory representations as follows: (i) Basic word-level syntax features (Part-of-Speech Tags (POS Tag) and Dependency Tags (DEP Tag)), (ii) Basic speech features (Phonological, MFCC, Mel and FBANK), (ii) Syntactic embeddings based on constituency parse trees, (iii) open-source language models such as

[1] https://scikit-learn.org/stable/modules/generated/sklearn.kernel_ridge.KernelRidge.html

BERT, GPT2 and Llama-2, and (iv) speech-based language model: Wav2vec2.0. We aim to understand the relative importance of these syntactic and language model representations while considering various HRF delays.

**Part-of-speech and dependency tags:** We use Spacy English dependency parser [Honnibal and Montani, 2017] to extract the POS and DEP tags. For each word, we generate a one-hot vector in which the corresponding POS tag location is 1 and the remaining tag values are 0. Similarly, we create a one-hot vector for dependency tags (DEP).

**Constituency Tree-based Embeddings:** Similar to [Reddy and Wehbe, 2021, Oota et al., 2023d], we build two types of constituency tree-based graph embeddings (ConTreGE): (i) ConTreGE Complete vectors (CC), and (ii) ConTreGE Incomplete vectors (CI) A CC vector is generated for every word using the largest subtree completed by that word. A subtree is considered complete when all of its leaves are terminals. The largest subtree a given word completes refers to the subtree with the most significant height. A CI vector is generated for every word using the incomplete subtree that contains all of the Phrase Structure Grammar productions needed to derive the words seen till then, starting from the root of the sentence's tree. Some examples for CC and CI are added in the Appendix (Figs. 5.20 and 5.21). Like Reddy and Wehbe [2021], we use Berkeley Neural Parser[2] for constituency parsing (i.e., for both CI and CC).

In the ConTreGE Complete tree (CC), the largest subtree completed by a given word refers to the subtree with the most significant height that also satisfies the following conditions - the given word must be one of its leaves, and all of its leaves must only contain words that have been seen till then.

In the ConTreGE Incomplete tree (CI), the embeddings are constructed using incomplete subtrees that are constructed by retaining all the phrase structure grammar productions required to derive the words seen till then, starting from the root of the sentence's tree. If incomplete subtrees are more representative of the brain's processes, it would mean that the brain correctly predicts specific phrase structures even before the entire phrase or sentence is read.

**BERT (encoder model): Devlin et al. [2019]** uses only encoder blocks of standard Transformer-based architecture with 12 layers and 768-dimensional representations. In order to extract BERT representations, we use BERT-base-uncased model from Huggingface [3]. We follow previous work to extract the hidden-state representations from each layer of these language models, given a fixed input length [Toneva and Wehbe, 2019].

**Varying the Context Length of BERT:** We constrained the model with maximum C words as context length to extract the stimulus features at different context lengths (C = 1, 5, and 20). Since the BERT model processes whole sentences, we input all the C context-length words into the BERT model and use the representation of the last word for the past context, similar to casual language model word representations GPT-2 [Radford et al., 2019]. For instance, given a story of M words and considering the context length of 5, while the third word's vector is computed by inputting the network with ($w_1$, $w_2$, $w_3$), the last word's vectors $w_M$ is computed by inputting the network with ($w_{M-5}$, $\cdots$, $w_M$). The pretrained

[2] https://spacy.io/universe/project/self-attentive-parser
[3] https://huggingface.co/bert-base-uncased

BERT model outputs word representations at different layers. We use the #words $\times$ 768 vector obtained from each hidden layer to obtain word-level representations.

**Speech-based language model:** Similar to text-based language models, we use popular pretrained Transformer speech-based model from Huggingface: Wav2Vec2.0 [Baevski et al., 2020]. Wav2Vec2.0 (encoder model) uses only encoder blocks of standard Transformer-based architecture with 12 layers and 768-dimensional representations. Here, the Transformer model was pretrained with contrastive loss as the objective function. To explore whether speech models incorporate linguistic information, we extract representations using context window of 16 secs with stride of 100 msecs and considered the last token as representation in each context window.

**Basic speech features:** We extract low-level speech features like filter banks (FBank), Mel Spectrogram, and MFCC from audio files using S3PRL toolkit [4], and phonological features using the DisVoice library [5]. (1) *FBank* divide the raw audio signal into multiple components (each one carrying a single frequency sub-band of the original signal) using a bandpass filter, results in a 26-dimensional vector. (2) *Mel Spectrogram* features are computed by applying a Fourier transform on the raw audio signal to analyze a signal's frequency content and converting it to the mel-scale, yielding an 80-dimensional vector. (3) *MFCC* features are Mel-frequency spectral coefficients obtained by taking the Discrete Cosine Transform (DCT) of the spectral envelope obtained from the logarithmic filter bank outputs. (4) *Phonological* features identify 108 phonological aspects (18 descriptors like vocalic, consonantal, back) across 6 statistical functions (mean, std, skewness, kurtosis, max, min). We employ the concatenation of all these basic speech features to model fMRI brain activity across hemodynamic response function (HRF) delays.

**Downsampling:** Since the rate of fMRI data acquisition (TR = 1.5sec for Narratives) was lower than the rate at which the text stimulus was presented to the subjects, several words fall under the same TR in a single acquisition. Hence, we match the stimulus acquisition rate to fMRI data recording by downsampling the stimulus features using a 3-lobed Lanczos filter. After downsampling, we obtain word embeddings corresponding to each TR. Similarly, for speech-based model, we perform this downsampling to obtain the chunk embedding corresponding to each TR.

**TR Alignment:** To account for the slowness of the hemodynamic response, we model the HRF using a finite response filter (FIR) per voxel and for each subject separately with various temporal delays. For instance, in Narratives listening, a temporal delay of 1 TR corresponds to 1.5 secs, and 5 TRs translates to a delay of 7.5 secs. Overall, the FIR filters were implemented by concatenating feature vectors that various delays had delayed.

## 5.3 Methodology

**Voxelwise Encoding Model:** We trained a bootstrapped ridge regression based encoding model [Tikhonov et al., 1977] to predict the fMRI brain activity associated with the stimulus representation obtained from syntax features and pretrained BERT. Before doing regression,

[4] https://github.com/s3prl/s3prl
[5] https://github.com/jcvasquezc/DisVoice

we first z-scored each feature channel separately for training and testing. This matched the features of the fMRI responses, which were also z-scored for training and testing. Each voxel value is predicted using a separate ridge regression model. Formally, at the time step (t), we encode the stimuli as $X_t \in \mathbb{R}^{N \times D}$ and brain voxels $Y_t \in \mathbb{R}^{N \times V}$, where $N$ denotes the number of training examples, $D$ denotes the dimension of concatenation of delayed TRs, and $V$ denotes the number of voxels. To find the optimal regularization parameter for each feature space, we use a range of regularization parameters that are explored using cross-validation. The main goal of each fMRI encoding model is to predict brain responses associated with each brain voxel given a stimulus.

**Cross-Validation:** We follow k-fold (k=4) cross-validation where k-1 folds were used for training and the remaining fold was held-out for testing.

**Evaluation Metrics:** We evaluate our models using the popular brain encoding evaluation metric described in the following. We compute the **Pearson correlation coefficient (PCC)** [Toneva and Wehbe, 2019, Caucheteux and King, 2020] between real and predicted fMRI brain activity to measure prediction performance for each voxel. PCC scores were then averaged over all voxels and across all folds. Finally, they are averaged across all subjects to obtain the final PCC score. PCC is computed as PCC=$\frac{1}{N}\sum_{i=1}^{n} corr[Y_i, \hat{Y}_i]$, where corr is the correlation function.

**Normalized Predictivity:** The neural model predictivity values were normalized by their respective subject ceiling values. The final measure of a model's performance ('normalized predictivity' or 'score') on a dataset is thus Pearson's correlation between model predictions and neural recordings divided by the estimated ceiling and averaged across voxels and participants.

**Hyper-parameter Settings**: We used banded ridge-regression with following parameters: MSE loss function, and L2-decay ($\lambda$) varied from $10^1$ to $10^3$. All experiments were conducted on a machine with 1 NVIDIA GEFORCE-GTX GPU with 16GB GPU RAM.

**Statistical Significance:** To estimate the statistical significance of the performance differences (across delays), we performed two-tailed paired-sample t-tests on the mean normalized predictivity scores for the subjects. Further, the Benjamni-Hochberg False Discovery Rate (FDR) correction Benjamini and Hochberg [1995] is used for all tests (appropriate because fMRI data is considered to have positive dependence [Genovese, 2000]). In all cases, we report significant p-values by representing the symbol as * (i.e., p$\leq$ 0.05).

## 5.4 Experimental Results

Here, we compare how the alignment with fMRI responses differs across a range of HRF delays for basic speech features, basic word-level syntactic features, syntactic embeddings, representations from pretrained language models considering different context lengths, and representations from pretrained speech-based language model, during listening across the brain. We then present the results of these analysis to shed light on the specific features that lead to this alignment at the whole brain, language regions and sub-regions. We calculate normalized brain predictivity independently for each types of model, averaging the results

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 15.58* | 28.23* | 40.06* | 45.46* | **47.86** | 47.57 | 46.36 | 45.58 |
| **BERT Context5** | 17.14* | 28.75* | 41.41* | 47.44 | 49.88 | 49.1 | **50.85** | 50.69 |
| **BERT Context20** | 22.83* | 34.05* | 44.67* | 46.0* | 53.62 | 53.0 | **53.81** | 53.42 |
| **Wav2vec2.0** | 25.2* | 34.22* | 41.45* | 44.57* | **47.12** | 45.94 | 46.24 | 46.14 |
| **POS Tag** | 5.82* | 11.2* | 16.35 | 18.55 | 17.77 | **19.14** | 18.5 | 16.44 |
| **DEP Tag** | 17.31* | 31.8* | 42.98* | **50.53** | 50.02 | 50.35 | 46.76 | 45.37* |
| **CC** | 15.24* | 30.45* | 43.66* | 47.91 | **48.08** | 47.55 | 45.88 | 44.18* |
| **CI** | 16.08* | 29.57* | 41.81* | 48.18 | **48.30** | 47.8 | 46.66 | 46.19 |
| **Basic Speech** | 17.86* | 21.49* | 24.98* | 27.6* | 29.53 | **29.62** | 28.93 | 29.34 |

Table 5.2: Whole Brain analysis across delays by fixing the delay5 as reference. Here, all the columns display average normalized brain predictivity across different feature representations. Underlined values show the best performance for given line. Starred $values*$ denotes a statistically significant in brain predictivity relative to delay 5. Highlighted in Red color denotes highest normalized predictivity across delays and models. Here, the delays are as follows: D1 (1.5 sec), D2 (3 sec), D3 (4.5 sec), D4 (6 sec), D5 (7.5 sec), D6 (9 sec), D7 (10.5 sec) and D8 (12 sec).

across participants within each model separately. We present the GPT-2 and Llama-2 results for Narratives tunneling dataset in the Appendix (see Figs. 5.16).

For assessing delay-wise performance of different stimuli representations, we utilize a standard HRF delay as a benchmark, reflecting the constant HRF delay referenced in prior studies. In all our findings, the set reference delay is 5 TRs (7.5 seconds) for the narratives tunneling dataset.

### 5.4.1 Whole Brain Analysis

We assess the degree to which each type of model aligns with different HRF delays across the whole brain. Here, we present the pretrained BERT model result for different context lengths during listening.

We show the normalized predictivity of each model obtained for a range of HRF delays in Fig. 5.2. Table 5.2 illustrates the variance analysis for Narratives listening.

**Narratives Listening** Using delay 5 (D5) as a reference, from Fig.5.2 (a) and Table 5.2, we make the following observations: (i) syntactic embeddings, including CC and CI, show higher normalized brain predictivity in the early delays, particularly D4, with a decrease in activity at later delays. (ii) The encoding of POS tag features within the brain starts from D3, which is not significant with higher delays (4-8). (iii) We also observe that DEP tags are early encoded in the brain at a delay of 6 secs (equivalent to 4 TRs). (ii) In contrast to syntactic embeddings, the normalized predictivity for BERT with different context lengths is significant for lower delays (1 to 4) and not significant for higher delays (6 to 8). This pattern suggests that the brain initially processes syntactic structures before moving on to sentence meaning. Semantic processing starting from the 5 delays (7.5 secs) and maintains this consistent processing up to the 8 delays (12 secs). For the BERT with different context lengths, we observe that the context 20 performs the best, implying that brain predictivity improves with increasing context length. Overall, the whole brain results indicate that syntactic information is processed during earlier delays, while semantic information is processed during

longer delays. When contrasting the speech-based language model Wav2vec2.0 with BERT, it was noted that the predictive capability of the speech-based model falls behind BERT after D3. However, it remains unclear what specific information speech models encode during early delays compared to text models.

When considering context lengths 5 and 20, there exists a slight enhancement in brain predictivity performance for delays 7 and 8. (ii) Likewise, for POS and DEP features, the performance of brain predictivity demonstrates improvement at higher delays. These findings collectively indicate that the brain engages in linguistic processing even during extended HRF delays, with context length notably influencing this connection with delayed responses.

Overall, the results show how the human brain may engage differently with language depending on whether it is being listened to or read. Listening might involve more dynamic and time-sensitive processing in the brain, especially in the early stages. In contrast, BERT processes written language with a more uniform level of predictivity over time, possibly because it does not replicate the time-sensitive aspects of human auditory processing.

### 5.4.2 Language ROIs Analysis

Fig. 5.3 displays the normalized predictivity scores that are examined within six language regions of interest (ROIs): AG, ATL, PTL, IFG, IFGOrb, and MFG. About D5 as the reference point, the subsequent insights are evident: (i) For the bilateral temporal lobes (ATL and PTL), syntactic embeddings, including CC and CI, as well as DEP tags, demonstrate increased normalized predictivity starting from the earlier delays (D4), implies that these regions encode word-level, parsing hierarchical structure and complex syntactic information. Conversely, BERT representations with contexts 5 and 20 exhibit higher predictivity in the later delays, specifically for the PTL, whereas the ATL maintains a consistent level of predictivity across all stimuli representations. These observations very loosely support the theory by Matchin and Hickok [Matchin and Hickok, 2020], which stipulates that parts of the PTL are The ATL is a knowledge store of entities in the later delays and is involved in hierarchical lexical-syntactic structure building (D1-D2: lexical, D3-D5 syntactic, and D6-D8 semantic). (ii) For the AG region, which processes high-level semantic information, BERT with context-20 showcases higher predictivity across delays and is significant with lower delays (1-4). This finding strongly suggests that the AG region is actively processing high-level semantic information. (iii) For the IFG region, the normalized predictivity for both BERT representations and DEP features is initially similar for delays up to 4. However, as the delays progress, BERT representations with longer contextual information outperform short contexts and DEP features regarding predictivity. This suggests that incorporating broader context information through BERT representations becomes more advantageous for predicting brain activity in the IFG region as the delay increases.

### 5.4.3 Sub-ROI-Level Analysis

Each language brain region is not necessarily homogeneous in function across all voxels it contains. Therefore, an aggregate analysis across an entire language region may mask some

nuanced effects. Thus, we further analyze several important language sub-regions that exemplify the variety of functionality across some of the broader language regions. Fig. 5.10 (see in the Appendix) illustrates the normalized predictivity scores for several important language sub-ROIs: IFG sub-regions (44, 45, IFJa), ATL (STGa, STSda), PTL (A5, PSL, STV). Using the delay-5 as a reference, we make the following observations: (i) Although syntactic embedding CI is not significant at early delay D4 in the IFG region, it is significant for 45 sub-ROI, hinting that these areas encode complex syntactic information [Reddy and Wehbe, 2021]. Further, DEP tags are good at encoding local-word information and are significant at D4. This implies that 45 sub-ROI encode both local-word complex syntactic information. On the other hand, for the sub-ROI 44, BERT with context 20 is only significant in the later delays D6-D8, which implies that this region encodes word-level and hierarchical syntax structure in early delays and processes semantic information in later delays. Another interesting observation is that syntactic embeddings, including CC and CI, are significant at an early delay D4 for the sub-ROIs PSL and STV in the PTL region, as these sub-ROIs are involved in many cognitive processes, including grammatical and syntactic processing. Furthermore, we note that BERT with context 20 exhibits superior normalized predictivity in the IFSp region compared to other stimuli representations. It is plausible that the IFSp region plays a role in retrieving auditory memories and creating short memories from verbal instructions. In the case of sub-ROIs PGp and PGs within the AG region, the normalized predictivity remains higher for BERT representations across delays. This observation suggests that these sub-ROIs primarily process high-level semantic information. Tables. 5.4 and 5.5 report the variance analysis across delays for each language sub-ROI (please see in the Appendix).

**Qualitative Analysis** To present the normalized predictivity above at an even finer grain, we show them now at the voxel-wise level across different feature representations, including, basic speech, hierarchical syntax features (CC), complex syntax features (CI), BERT with context length 20, Wav2vec2.0 in Figures 5.4, 5.5, 5.7, 5.8, and 5.9, respectively. (1) Basic speech features exhibit higher normalized brain predictivity in the early auditory area, starting from D1, and demonstrate increased predictivity in later delays. This suggests that low-level auditory information, such as phonological features and other basic speech features, is encoded at the initial stage of information processing. We also observe similar findings from the speech-based model Wav2vec2.0, where early delays show greater predictivity in the early auditory cortex. (2) In contrast to basic speech features and Wav2vec2.0, syntactic information such as basic word-level and syntactic embeddings are processed early in the bilateral temporal lobes (ATL & PTL), starting from delay 2. An intriguing discovery is that even syntactic embeddings such as CC and CI exhibit some predictivity in the auditory area where phonological features are processed. We empirically verified that syntactic embeddings share some phonological information, which accounts for their predictivity in that region. (3) The effect of high-level semantic regions (i.e. both Frontal and Parietal regions) are highly predictive from delay4. (4) The linguistic information is highly effective in the posterior cingulate cortex (PCC) from delay5. Surprisingly, neither basic speech nor Wav2vec2.0 embeddings demonstrate any predictivity in the PCC or dmPFC regions. These regions are recognized for processing semantic properties such as tense, subject number, and object number [Oota et al., 2023c].

### 5.4.4 Ablation Studies

**BERT vs. GPT2 vs. Llama-2** We extended our analysis to GPT2 and Llama-2 models, extracting representations at different context lengths (1, 5, and 20). We observe that both GPT2 and Llama-2 demonstrate similar pattern of normalized predictivity across context lengths as BERT model. Comprehensive results for the GPT2 and Llama-2 models can be found in the *Appendix*. To determine whether representations of these language models share information across layers and different context lengths, we first encoded model representations using other model presentations and then computed the $R^2$-score, as shown in Figure 5.12.

**Do syntactic embeddings share basic phonological information?** Figures 5.4 and 5.5 display brain maps for basic speech features and hierarchical syntactic embeddings (CC) respectively. Observations from the Figure 5.5 dedicated to hierarchical syntactic embeddings reveal that even syntactic embeddings possess predictive power in early auditory areas. To determine if these syntactic embeddings share information with basic speech features, we first encoded model representations using other feature presentations and then computed the $R^2$-score, as shown in Figure 5.14.

## 5.5 Discussion and Conclusion

We examine how the intricate processing of diverse language regions at varying HRF delays in the human brain corresponds with word-level syntactic features, syntactic embeddings obtained from constituent parsers, and Transformer-based language model representations. To do this, we build encoding models for various HRF delays. These models enable us to analyze the impact of the alignment with fMRI brain recordings acquired while participants listened to or read naturalistic stories. We show that word-level syntactic information, particularly DEP Tags, is significantly encoded at early delays (D4) in specific language regions IFG and IFGOrb; these regions are known to process syntactic information [Friederici et al., 2003, Friederici, 2012]. Using constituent syntactic embeddings, we find that hierarchical syntax information is significantly encoded in the MFG region at early delays D4. This region is implicated in higher-level cognitive functions, including processing complex and hierarchical structures, which aligns with the nature of hierarchical syntax parsing [Scholz et al., 2022]. Additionally, we find that complex syntax information, represented by (CI), is encoded in the IFGOrb region. We also find that these embeddings have improved normalized brain activity at early delays across whole brain and language ROIs [Friederici et al., 2003, Friederici, 2012]. Using pretrained language model representations, the detailed region and sub-region analysis reveals that longer context may impact the observed neural activity at different brain language regions, specifically at higher HRF delays. For instance, BERT with a context length of 20 has higher brain predictivity within language regions such as AG, IFG, ATL, and PTL, specifically for higher delays ranging from 9 to 12 seconds.

**Implications of our findings** The insights gained from our work could have implications for AI engineering, neuroscience, and the interpretability of models. **Neuro-AI engineering:** Our work immediately fits in with the neuro-AI research direction that specifically investigates the relationship between representations in the brain and representations learned

by powerful LMs. This direction has gained recent traction, especially in the domain of language, thanks to advancements in language models [Toneva and Wehbe, 2019, Schrimpf et al., 2021b, Goldstein et al., 2022, Aw and Toneva, 2023, Oota et al., 2023c]. **Model Explainability:** The recent studies explored syntactic and semantic differences in brain language processing by generating pure syntactic embeddings from parse trees (constituent and dependency parsers) [Reddy and Wehbe, 2021, Oota et al., 2023d, Zhang et al., 2022a], or disentangle the language model representations by controlling the syntactic information [Caucheteux et al., 2021a]. However, all these studies perform their analysis at a fixed delay. In the longer term, our method of varying across delays aims to enhance more detailed language processing by disentangling the LM representations into distinct components: syntactic (constituents & dependencies) and semantic aspects (discourse, emotion, etc,.). These variations across LM layers can further increase the model interpretability and brain insights this line of work enables.

## 5.6 Limitations

One limitation of our approach is the interpretation of the syntactic vs. semantic differences by disentangling the pretrained language models representations like BERT, GPT-2, and Llama-2 to observe the brain alignment across delays. However, we tested word-level syntactic and constituent syntactic embeddings and compared them with BERT embeddings. However, BERT encodes a hierarchy of linguistic properties ranging from surface-syntactic to semantic across layers. Hence, it is worth disentangling these representations to understand detailed language processing across delays further. Another limitation is that some of the differences in brain alignment we observe are due to confounding differences between model types (BERT vs. GPT-2), and there is value in investigating these questions in the future with models that are controlled for architecture, objective, and training data amounts.

# *Appendix for: Interpretaion of HRF delays*

---

## 5.7 Narratives Tunneling

### 5.7.1 Language ROIs results

Using delay 5 (D5) as a reference, Table. 5.3 reports the variance analysis across delays for each language ROI. We make the following observations: (i) BERT with a context length of 20 consistently displays higher normalized predictivity at delay5 across ROIs and is always significant only for higher delays. For shorter context lengths of 1 and 5, several language ROIs like AG, ATL, and PTL exhibit the highest normalized brain predictivity at delay5. On the other hand, the other language ROIs reach their highest normalized predictivity at delay 4. This implies that the variance tends to be more pronounced for shorter contexts as

the delay increases. The predictive power of BERT representations with longer contexts remains relatively stable after five delays. We observe similar findings for pretrained language models like GPT-2 and Llama-2.

### 5.7.2 Language sub-ROIs results

Each language brain region is not necessarily homogeneous in function across all voxels it contains. Therefore, an aggregate analysis across an entire language region may mask some nuanced effects. Thus, we further analyze several important language sub-regions that are thought to exemplify the variety of functionality across some of the broader language regions. Fig. 5.10 illustrates the normalized predictivity scores for several important language sub-ROIs: IFG sub-regions (44, 45, IFJa), ATL (STGa, STSda), PTL (A5, PSL, STV). Using the delay-5 as a reference, we make the following observations: (i) Although syntactic embedding CI is not significant at early delay D4 in the IFG region, it is significant for 45 sub-ROI, hinting that these areas encode complex syntactic information [Reddy and Wehbe, 2021]. Further, DEP tags are good at encoding local-word information and are significant at D4. This implies that 45 sub-ROI encode both local-word complex syntactic information. On the other hand, for the sub-ROI 44, BERT with context 20 is only significant in the later delays D6-D8, which implies that this region encodes word-level and hierarchical syntax structure in early delays and processes semantic information in later delays. Another potentially interesting observation is that syntactic embeddings, including CC and CI, are significant at an early delay D4 for the sub-ROIs PSL and STV in the PTL region, as these sub-ROIs are involved in many cognitive processes, including grammatical and syntactic processing. Furthermore, we note that BERT with context 20 exhibits superior normalized predictivity in the IFSp region compared to other stimuli representations. It is plausible that the IFSp region plays a role in retrieving auditory memories and creating short memories from verbal instructions. In the case of sub-ROIs PGp and PGs within the AG region, the normalized predictivity remains higher for BERT representations across delays. This observation suggests that these sub-ROIs primarily process high-level semantic information.

Tables. 5.4 and 5.5 report the percentage change across delays for each language sub-ROI. We make the following observations: (i) BERT with a context length of 20 consistently displays higher normalized predictivity at delay5 across sub-ROIs. (ii)For language sub-ROIs, we observe that the percentage change tends to be more pronounced for shorter contexts and basic syntactic features, similar to language ROIs.

## 5.8 GPT2: Whole Brain Analysis

Fig. 5.15 displays the whole brain analysis across various delays across different context lengths. By considering delay 5 as a reference, we observe that the normalized predictivity for GPT2 with different context lengths is significant for lower delays (1 to 4) and not significant for higher delays (6 to 8). This clearly shows that the brain processes linguistic information starting from the 5 delays (7.5 secs) and maintains this consistent processing up to the 8 delays (12 secs). In contrast to BERT, for the GPT2 with different context

lengths, we observe that (i) Context 1 and 5 perform the best, implying that brain predictivity improves with shorter context length.

## 5.9 GPT2: Language sub-ROI Analysis

Fig 5.16 illustrates the normalized predictivity scores for several important language sub-ROIs: IFG sub-regions (44, 45, IFJa, IFSp), ATL (STGa, STSda, STSva), PTL (A5, STSdp, STSvp), AG (PGi, PGs, PGp, PFm). Considering the delay-5 as a reference, We make the following observations: (i) GPT-2 with context 1 shows higher predictivity in all sub ROIs (44, 45, IFJa, IFSp) of IFG region, it is noteworthy that these language sub ROIs are mainly involved in processing lexical information, syntax processing . (ii) Furthermore, we note that AG sub ROIs exhibit superior normalized predictivity for GPT2 with context 5 compared to other context lengths. It's plausible that the AG region plays a role in processing semantic comprehension and transitions to handling semantic roles between tokens.

## 5.10 Narratives Tunneling: Llama Results

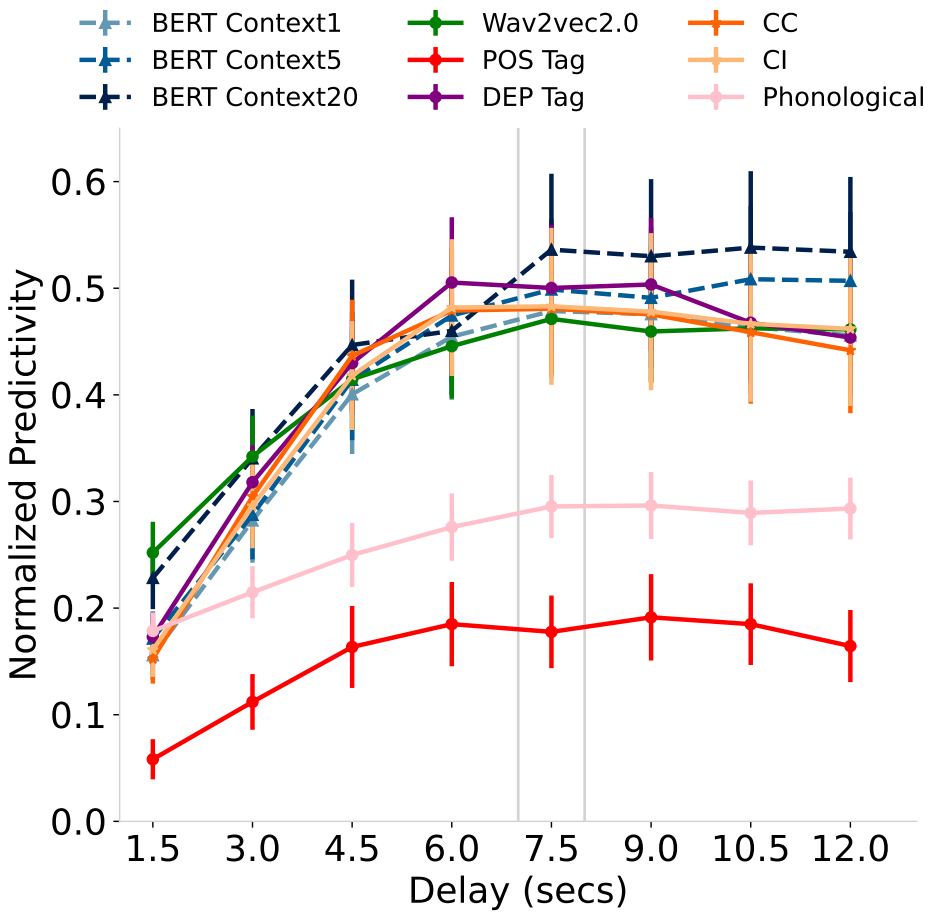



Figure 5.2: Whole Brain Normalized Predictivity: This plot provides a comparison of delay-wise performance for various stimuli representations, averaged across participants and layers. The vertical grey line serves as a reference point, representing the constant HRF delay used in previous studies. In this context, the reference delay is 5 TRs (7.5 seconds). Error bars denote standard error across participants.

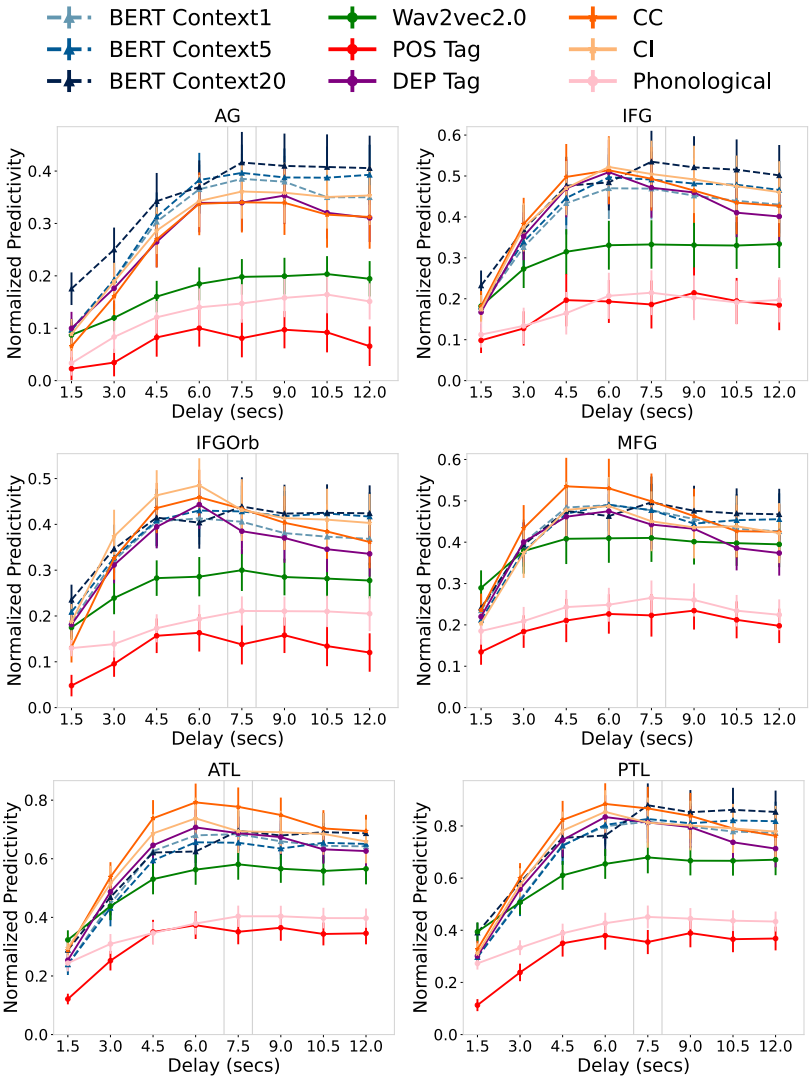



Figure 5.3: Language ROIs-based normalized brain predictivity was computed by averaging across participants, layers, and voxels. Dotted lines patterned for BERT representations and solid lines report basic syntax and constituent syntax embeddings.

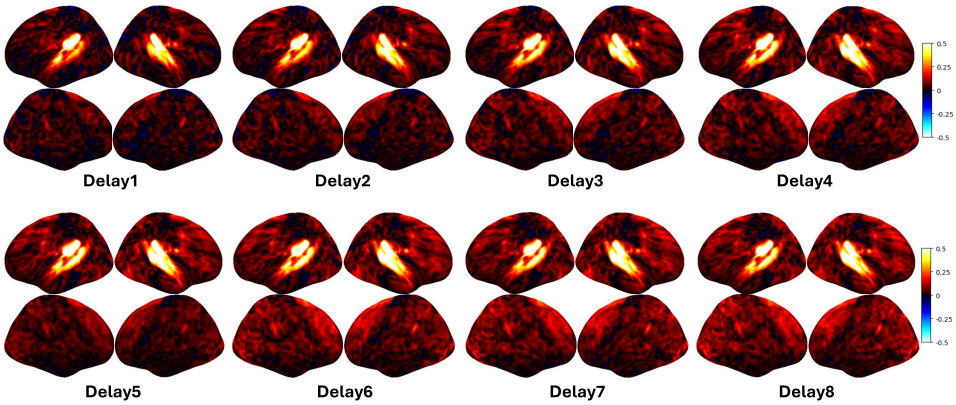



Figure 5.4: Basic Speech features: Brain Maps: Voxelwise normalized brain predictivity average across layers and subjects for different HRF delays.

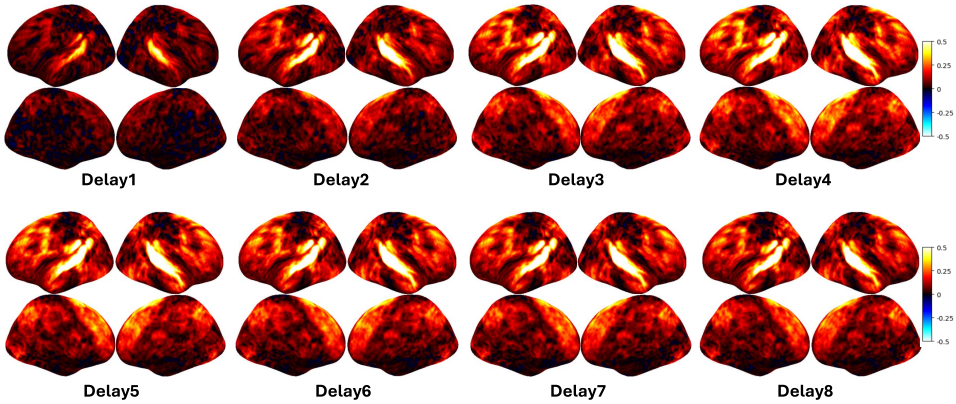



Figure 5.5: Hierarchical syntax features (CC): Brain Maps: Voxelwise normalized brain predictivity average across layers and subjects for different HRF delays.

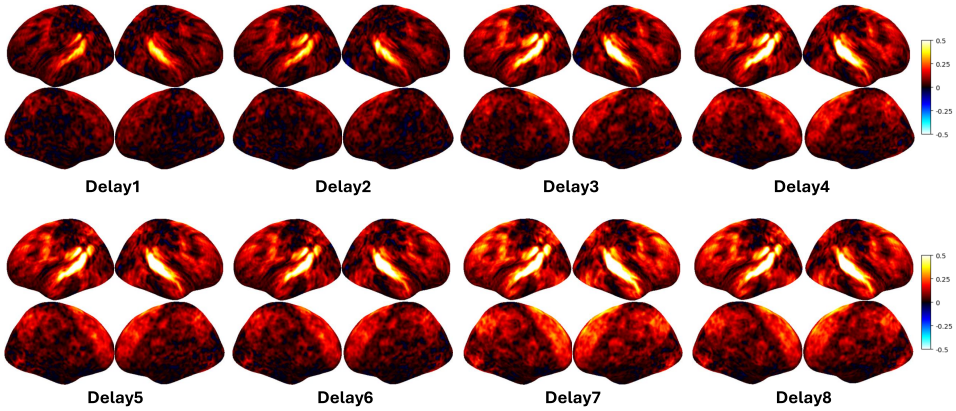



Figure 5.6: Residual brainmaps after removal of phonological features from CC: Voxelwise normalized brain predictivity average across layers and subjects for different HRF delays.

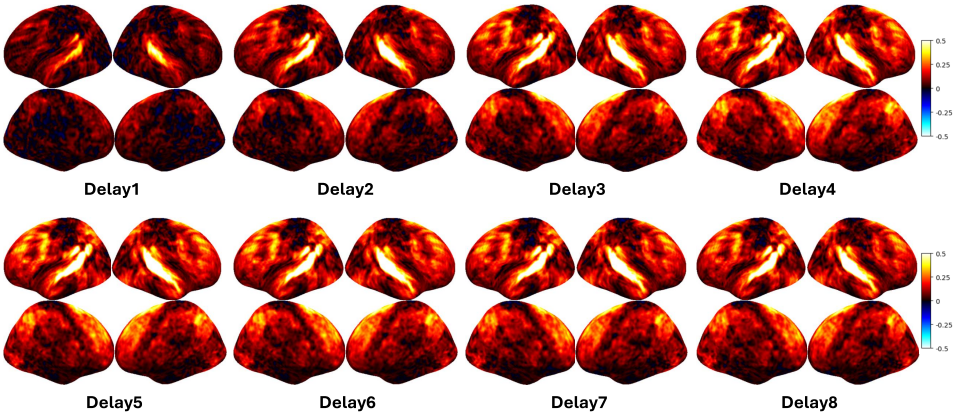



Figure 5.7: Complex syntax features (CI): Brain Maps: Voxelwise normalized brain predictivity average across layers and subjects for different HRF delays.

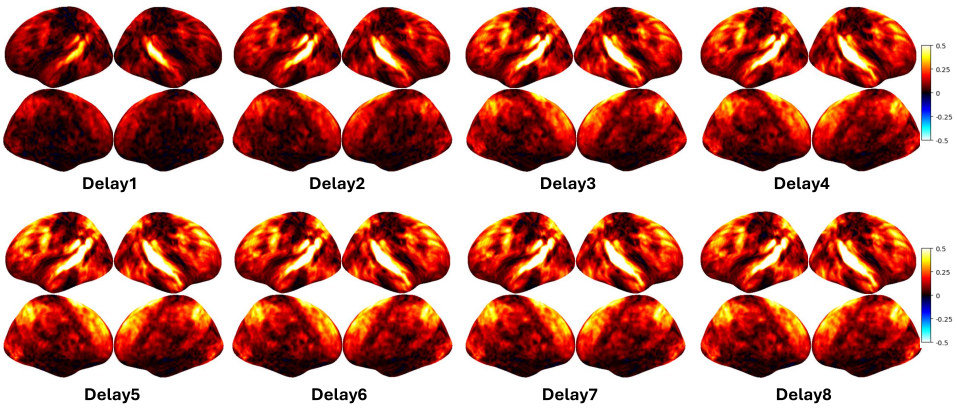



Figure 5.8: BERT Context20 Brain Maps: Voxelwise normalized brain predictivity average across layers and subjects for different HRF delays.

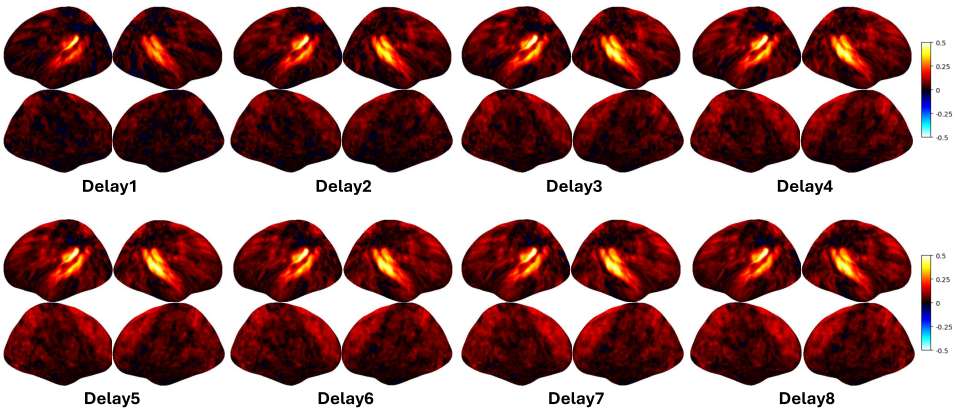



Figure 5.9: Wav2vec2.0 Brain Maps: Voxelwise normalized brain predictivity average across layers and subjects for different HRF delays.

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 9.79* | 19.17* | 30.33* | 36.52 | **38.51** | 37.97 | 35.0 | 35.0 |
| **BERT Context5** | 9.63* | 19.29* | 31.29* | 38.29 | **39.65** | 38.78 | 38.73 | 39.29 |
| **BERT Context20** | 17.54* | 25.01* | 34.25* | 36.97* | **41.63** | 40.99 | 40.78 | 40.56 |
| **Wav2vec2.0** | 8.72* | 11.99* | 16.04* | 18.46 | 19.8 | 19.95 | **20.34** | 19.44 |
| **POS Tag** | 2.27 | 3.45* | 8.24 | **10.00** | 8.1 | 9.71 | 9.22 | 6.57 |
| **DEP Tag** | 9.99* | 17.62* | 26.48* | 33.92 | 33.98 | **35.32** | 32.04 | 31.07 |
| **CC** | 6.55* | 16.09* | 26.95* | 33.75 | **34.03** | 33.95 | 31.68 | 31.24 |
| **CI** | 8.88* | 18.9* | 28.64* | 34.29 | **36.1** | 35.86 | 35.05 | 35.34 |
| **Basic Speech** | 3.33* | 8.33* | 12.08 | 14.0 | 14.72 | 15.8 | **16.42** | 15.12 |

(a) AG

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 23.96* | 44.66* | 62.61* | 67.98 | **68.40** | 65.95 | 64.42* | 64.24* |
| **BERT Context5** | 23.94* | 43.22* | 59.42* | **65.61** | 65.46 | 63.48 | 65.45 | 65.06 |
| **BERT Context20** | 28.89* | 46.88* | 62.12* | 62.48* | **69.53** | 67.94 | 69.08 | 68.68 |
| **Wav2vec2.0** | 32.31* | 43.84* | 53.04* | 56.28* | **58.10** | 56.6 | 55.83 | 56.55 |
| **POS Tag** | 12.13* | 25.22* | 35.01 | **37.33** | 35.08 | 36.45 | 34.33 | 34.55 |
| **DEP Tag** | 25.37* | 48.71* | 64.61 | **70.69** | 68.73 | 67.39 | 63.2 | 62.63* |
| **CC** | 28.81* | 53.79* | 73.82* | **79.24** | 77.73 | 74.92 | 70.34* | 69.46* |
| **CI** | 29.57* | 51.55* | 68.64 | **73.78** | 69.39 | 69.04 | 68.48 | 65.91 |
| **Basic Speech** | 24.39* | 30.93* | 34.64* | 37.9* | **40.37** | 40.36 | 39.75 | 39.73 |

(b) ATL

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 29.84* | 51.79* | 72.67* | 79.9 | **81.61** | 79.74 | 77.87* | 77.33* |
| **BERT Context5** | 29.67* | 51.24* | 72.42* | 80.42 | **82.66** | 81.02 | 82.13 | 81.81 |
| **BERT Context20** | 39.08* | 58.38* | 75.44* | 76.32* | **87.93** | 85.27 | 86.18 | 85.38 |
| **Wav2vec2.0** | 39.43* | 50.73* | 61.05* | 65.44* | **67.93** | 66.67 | 66.64 | 67.05 |
| **POS Tag** | 11.29* | 23.84* | 35.0 | 37.92 | 35.49 | **38.92** | 36.58 | 36.84 |
| **DEP Tag** | 30.08* | 55.76* | 74.6 | **83.42** | 81.37 | 79.52 | 73.74* | 71.31* |
| **CC** | 32.69* | 60.0* | 82.3 | **88.45** | 86.72 | 83.86 | 79.03* | 76.19* |
| **CI** | 31.07* | 57.37* | 78.22 | **85.41** | 81.27 | 80.16 | 79.05 | 77.82 |
| **Basic Speech** | 27.29* | 33.32* | 38.84* | 42.69 | **45.11** | 44.46 | 43.67 | 43.35 |

(c) PTL

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 18.01* | 32.63* | 43.34 | **47.00** | 46.92 | 45.21 | 43.94* | 42.99* |
| **BERT Context5** | 18.11* | 33.94* | 44.63* | **49.69** | 49.12 | 48.15 | 47.86 | 46.6 |
| **BERT Context20** | 23.23* | 36.91* | 47.61* | 48.44* | **53.49** | 52.11 | 51.55 | 50.16* |
| **Wav2vec2.0** | 18.2* | 27.31* | 31.48 | 33.08 | 33.27 | 33.12 | 33.01 | **33.38** |
| **POS Tag** | 9.83 | 12.75 | 19.66 | 19.34 | 18.6 | **21.44** | 19.45 | 18.47 |
| **DEP Tag** | 16.72* | 35.27* | 46.99 | **50.94*** | 47.14 | 45.98 | 41.03 | 40.11* |
| **CC** | 18.17* | 38.26* | 49.77 | **51.40** | 49.32 | 46.44 | 43.47* | 42.69* |
| **CI** | 17.39* | 37.44* | 47.03 | **52.21** | 50.48 | 49.17 | 47.51 | 46.06* |
| **Basic Speech** | 11.25* | 13.38* | 16.47* | 20.72 | **21.48** | 20.23 | 19.13 | 19.72 |

(d) IFG

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 18.66* | 32.11* | 40.63 | **41.31** | 40.57 | 38.09 | 37.31* | 36.84* |
| **BERT Context5** | 20.88* | 32.71* | 40.86 | **43.02** | 42.82 | 41.78 | 42.37 | 41.71 |
| **BERT Context20** | 23.53* | 34.65* | 41.47 | 40.38 | **43.89** | 42.35 | 42.53 | 42.42 |
| **Wav2vec2.0** | 17.49* | 23.92* | 28.26 | 28.58 | **30.00** | 28.5 | 28.16 | 27.74 |
| **POS Tag** | 4.8* | 9.55 | 15.68 | **16.32** | 13.79 | 15.8 | 13.43 | 12.02 |
| **DEP Tag** | 18.12* | 31.13* | 39.45 | **44.32*** | 38.51 | 37.04 | 34.57 | 33.57* |
| **CC** | 13.2* | 32.6* | 43.56 | **45.91** | 43.39 | 40.35 | 38.44 | 36.14* |
| **CI** | 18.74* | 37.39 | 46.33 | **48.52*** | 43.13 | 41.37 | 41.03 | 40.31 |
| **Basic Speech** | 12.99* | 13.86* | 17.31 | 19.35 | **21.10** | 21.05 | 21.0 | 20.51 |

(e) IFGOrb

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 23.5* | 39.9* | 48.38 | **48.95*** | 47.72 | 45.6 | 43.23* | 42.63* |
| **BERT Context5** | 21.19* | 37.92* | 46.87 | **49.07** | 47.73 | 44.49* | 45.32 | 45.56 |
| **BERT Context20** | 24.15* | 39.37* | 47.57 | 46.32 | **49.63** | 47.59 | 46.95 | 46.75 |
| **Wav2vec2.0** | 28.95* | 37.86 | 40.82 | 40.94 | **41.03** | 40.13 | 39.76 | 39.47 |
| **POS Tag** | 13.46* | 18.39 | 21.06 | 22.64 | 22.3 | **23.45** | 21.22 | 19.77 |
| **DEP Tag** | 22.0* | 40.03 | 46.16 | **47.49** | 44.26 | 43.24 | 38.57* | 37.37* |
| **CC** | 23.61* | 43.37* | **53.51*** | 53.03* | 49.88 | 46.31 | 42.66* | 42.51* |
| **CI** | 20.35* | 37.38* | 47.81 | **48.74** | 45.05 | 43.57 | 43.84 | 42.19 |
| **Basic Speech** | 18.51* | 20.91* | 24.3 | 24.87 | **26.55** | 26.03 | 23.43 | 22.44 |

(f) MFG

Table 5.3: Language ROIs analysis of BERT and syntactic features: variance analysis across delays by fixing the delay5 as constant.

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 18.1* | 32.22* | 41.46 | **44.42** | 43.96 | 41.75 | 40.44 | 38.72* |
| **BERT Context5** | 20.02* | 33.8* | 41.33* | **45.74** | 45.58 | 43.47 | 43.45 | 41.99 |
| **BERT Context20** | 23.23* | 34.77* | 43.29* | 43.75* | **47.59** | 45.53 | 45.46 | 44.01 |
| **Wav2vec2.0** | 19.9* | 26.85* | 29.84 | 31.32 | **31.41** | 29.79 | 29.41 | 29.78 |
| **POS Tag** | 7.7 | 11.08 | 14.54 | **15.97** | 12.42 | 15.04 | 12.95 | 12.67 |
| **DEP Tag** | 13.75* | 29.28* | 41.16 | **46.52** | 42.71 | 42.1 | 38.24 | 37.17* |
| **CC** | 18.5* | 39.17* | 48.62 | **51.10** | 49.04 | 45.65 | 42.47* | 40.83* |
| **CI** | 21.22* | 36.65* | 44.66 | **50.82** | 46.90 | 45.25 | 43.76 | 41.36* |
| **Basic Speech** | 10.5* | 10.66* | 12.97* | 15.64 | **18.13** | 17.35 | 16.5 | 17.34 |

(a) 44

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 21.28* | 37.45* | 47.8 | **48.57** | 46.77 | 44.13 | 42.82* | 42.33* |
| **BERT Context5** | 22.88* | 36.72* | 46.88 | **49.97** | 49.47 | 48.03 | 47.71 | 46.52 |
| **BERT Context20** | 25.2* | 38.92* | 46.95 | 45.99 | **49.79** | 47.76 | 47.72 | 47.31 |
| **Wav2vec2.0** | 22.11* | 30.51* | 35.29 | 36.37 | **37.65** | 35.34* | 34.91 | 34.19 |
| **POS Tag** | 3.96* | 10.88 | 18.67 | **19.35** | 16.13 | 18.46 | 16.06 | 13.78 |
| **DEP Tag** | 19.97* | 35.32* | 44.45 | **49.49*** | 43.93 | 41.59 | 39.54 | 37.77* |
| **CC** | 16.35* | 38.56* | 52.04 | **54.30** | 51.50 | 47.93 | 45.88* | 42.58* |
| **CI** | 18.88* | 39.12* | 51.42 | **53.51*** | 47.02 | 45.17 | 44.2 | 44.15 |
| **Basic Speech** | 17.33* | 18.69* | 21.78 | 23.78 | **25.76** | 24.64 | 23.99 | 22.97 |

(b) 45

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 14.24* | 29.6* | 41.52 | 43.31 | **43.84** | 41.72 | 40.11 | 39.86 |
| **BERT Context5** | 14.25* | 32.86* | 43.36 | **47.52** | 47.21 | 46.92 | 46.53 | 45.82 |
| **BERT Context20** | 19.76* | 36.48* | 46.69* | 46.48 | **52.67** | 51.3 | 50.44 | 48.86* |
| **Wav2vec2.0** | 18.27* | 29.87 | 33.86 | 34.54 | 34.59 | 36.22 | **36.24** | 35.84 |
| **POS Tag** | 8.68* | 11.3* | 20.72 | 19.97 | 20.88 | **26.27** | 25.6 | 23.33 |
| **DEP Tag** | 18.85* | 38.47* | 50.09 | **52.21** | 50.68 | 48.35 | 42.4 | 41.88 |
| **CC** | 18.83* | 36.44* | **48.10** | 47.61 | 46.26 | 43.78 | 40.55 | 39.17* |
| **CI** | 19.28* | 41.95* | 50.45 | 51.29 | **51.50** | 49.43 | 46.54 | 44.95* |
| **Basic Speech** | 12.9* | 16.3 | 18.73 | **23.48** | 22.31 | 21.09 | 20.08 | 20.69 |

(c) IFJa

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 21.93* | 36.46* | 47.79* | **54.39** | 54.14 | 53.54 | 52.68 | 52.02 |
| **BERT Context5** | 19.73* | 35.28* | 50.35* | **57.27** | 55.86 | 55.67 | 55.15 | 53.55 |
| **BERT Context20** | 26.98* | 40.2* | 54.31* | 56.75 | **62.19** | 61.69 | 60.8 | 59.68 |
| **Wav2vec2.0** | 15.87* | 25.15 | 31.11 | 33.85 | 34.32 | 34.18 | 34.29 | **35.49** |
| **POS Tag** | 13.88 | 16.51 | **25.29** | 23.11 | 24.34 | 24.71 | 21.44 | 20.92 |
| **DEP Tag** | 18.34* | 39.76 | 51.38 | **55.44*** | 49.19 | 48.56 | 43.25 | 42.09 |
| **CC** | 17.02* | 39.01* | 53.1 | **55.89** | 52.99 | 50.35 | 47.95 | 48.93 |
| **CI** | 10.27* | 33.64* | 46.48* | **55.04** | 54.11 | 54.05 | 53.51 | 53.47 |
| **Basic Speech** | 10.45 | 13.84* | 18.67* | 24.46 | **25.01** | 23.13 | 21.57 | 21.8 |

(d) IFSp

Table 5.4: Language sub-ROIs analysis of BERT and syntactic features: variance analysis across delays by fixing the delay5 as constant.

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 14.2* | 30.15* | 43.52 | **47.84** | 47.7 | 45.52 | 45.61 | 45.38 |
| **BERT Context5** | 18.16* | 32.06* | 42.91 | **47.00** | 46.34 | 44.28 | 46.17 | 44.08 |
| **BERT Context20** | 22.76* | 36.12* | 45.35* | 46.53 | **50.11** | 48.13 | 49.7 | 48.61 |
| **Wav2vec2.0** | 12.78* | 20.95* | 27.27 | 28.74 | 29.53 | 29.1 | 28.26 | 28.33 |
| **POS Tag** | 7.28* | 13.49* | 22.09 | **24.08** | 22.16 | 23.66 | 20.04 | 19.46 |
| **DEP Tag** | 16.31* | 34.62* | 45.7 | **51.46** | 49.3 | 45.75 | 42.43* | 42.97* |
| **CC** | 18.43* | 34.79* | 49.52 | **55.43** | 54.97 | 51.71 | 49.79 | 49.47 |
| **CI** | 18.15* | 32.89* | 43.24 | **47.43** | 45.14 | 47.1 | 46.29 | 46.11 |
| **Basic Speech** | 13.39* | 19.57 | 19.49 | 20.52 | 22.59 | 22.17 | 21.87 | **22.77** |

(e) STGa

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 15.34* | 28.65* | 38.26 | 39.79 | **39.89** | 38.4 | 37.59 | 38.19 |
| **BERT Context5** | 14.69* | 25.63* | 35.0 | **36.3** | 35.08 | 32.97 | 33.82 | 33.52 |
| **BERT Context20** | 17.79* | 27.57* | 35.91 | 35.53 | **38.48** | 37.03 | 38.16 | 37.81 |
| **Wav2vec2.0** | 26.95* | 34.57* | 41.19* | 43.09 | **43.97** | 42.52 | 41.5 | 41.73 |
| **POS Tag** | 4.73* | 12.23 | 15.17 | 16.28 | **18.34** | 17.16 | 17.11 | 16.8 |
| **DEP Tag** | 21.99* | 32.02* | 38.97 | **39.81** | 38.77 | 38.39 | 33.86* | 34.17* |
| **CC** | 20.27* | 36.54* | 45.75 | **48.24** | 47.32 | 44.82 | 43.05 | 40.83* |
| **CI** | 20.28* | 33.71* | **41.56** | 41.26 | 39.9 | 40.19 | 39.07 | 38.64 |
| **Basic Speech** | 31.72 | 32.6 | 31.59 | 32.38 | 33.36 | 32.87 | **33.53** | 32.35 |

(f) TA2

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 26.24* | 46.97* | 63.33 | **66.13** | 64.35 | 63.82 | 62.18 | 62.47 |
| **BERT Context5** | 27.09* | 46.48* | 61.34* | 66.29 | 66.54 | 65.48 | **67.32** | 66.52 |
| **BERT Context20** | 33.15* | 51.68* | 63.61* | 63.62* | **70.97** | 68.78 | 69.51 | 68.86 |
| **Wav2vec2.0** | 39.74* | 54.86* | 64.03 | 66.59 | **67.66** | 65.25 | 65.16 | 66.53 |
| **POS Tag** | 10.68* | 19.61 | 23.08 | **27.58** | 24.73 | 27.15 | 27.33 | 26.45 |
| **DEP Tag** | 30.78* | 53.71* | 65.38 | **69.4** | 65.36 | 64.09 | 58.05* | 55.55* |
| **CC** | 28.42* | 55.03* | 72.11 | **73.62*** | 69.22 | 66.29 | 61.13* | 62.52* |
| **CI** | 28.18* | 54.47* | 69.07 | **71.17*** | 64.0 | 61.06* | 59.85 | 59.97 |
| **Basic Speech** | 42.2* | 45.39* | 49.98 | 50.27 | **51.83** | 50.04 | 49.42 | 47.68 |

(g) PSL

| ↓Models / Delays→ | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **BERT Context1** | 46.66* | 78.41* | 109.12* | 117.13 | **118.19** | 113.68* | 110.35* | 109.24* |
| **BERT Context5** | 48.93* | 80.99* | 111.55* | 121.2 | **122.06** | 119.55 | 119.35 | 119.67 |
| **BERT Context20** | 61.55* | 89.85* | 117.73* | 115.24* | **129.71** | 124.47 | 125.61 | 124.45 |
| **Wav2vec2.0** | 62.97* | 85.42* | 104.4* | 109.98* | **113.48** | 111.84 | 109.95* | 110.96 |
| **POS Tag** | 14.45* | 36.57* | 51.55 | 51.38 | 52.8 | 51.32 | 52.29 | **56.04** |
| **DEP Tag** | 55.95* | 96.44* | 127.4 | **136.18** | 130.21 | 124.67* | 114.72* | 110.14* |
| **CC** | 53.21* | 96.47* | 128.39 | **135.94*** | 130.77 | 123.86 | 114.93* | 112.22* |
| **CI** | 51.52* | 95.91* | 126.72 | **131.55*** | 123.51 | 122.23 | 120.12 | 116.87 |
| **Basic Speech** | 53.61* | 62.07* | 73.73* | 81.14 | 83.07 | 80.48* | 77.31* | 75.17* |

(h) STV

Table 5.5: Language sub-ROIs analysis of BERT and syntactic features: variance analysis across delays by fixing the delay5 as constant.

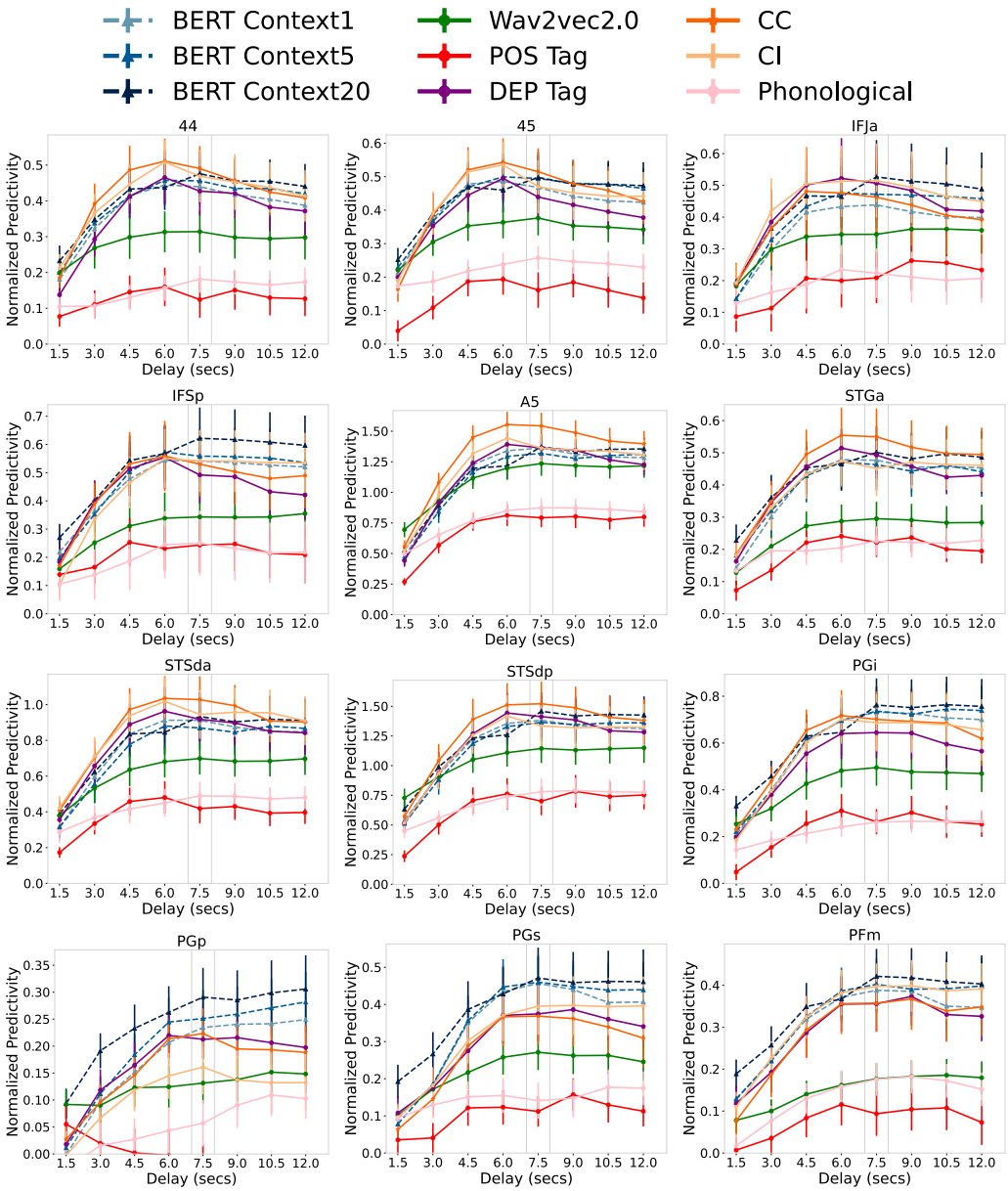



Figure 5.10: Narratives Tunneling: Language sub-ROIs-based normalized brain predictivity was computed by averaging across participants, layers, and voxels. Dottled lines patterned for BERT representations and solid lines report basic syntax and constituent syntax embeddings.

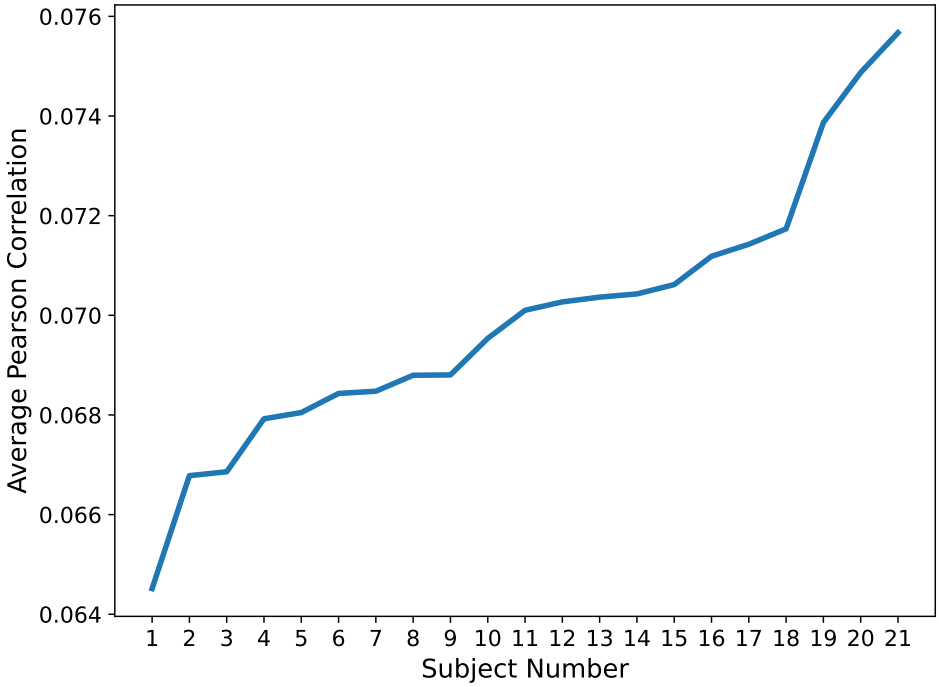



Figure 5.11: Estimated Noise Ceiling: Average Pearson Correlation across voxels for each subject.

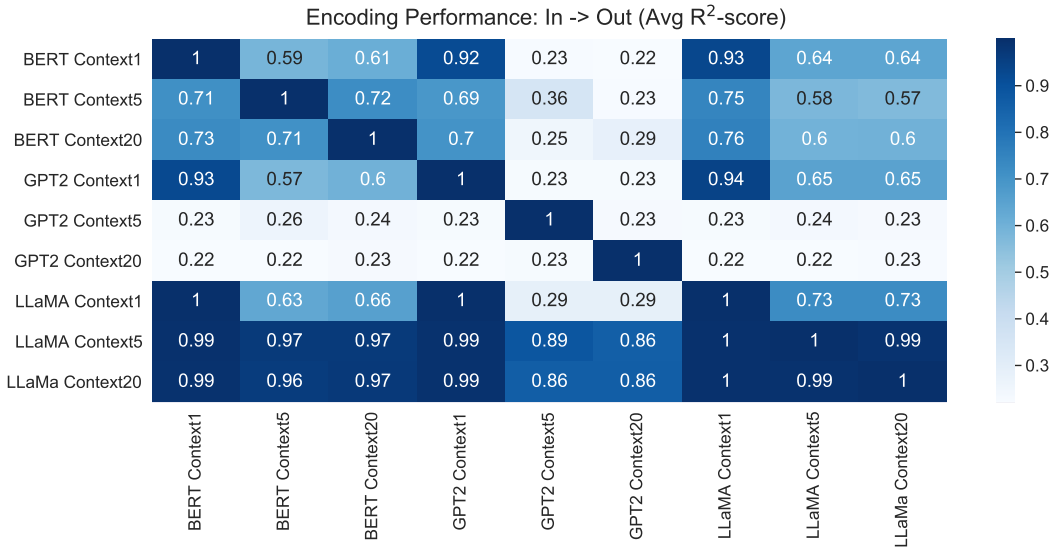



Figure 5.12: BERT vs. GPT2 vs. Llama2 - Average $R^2$-score was calculated by encoding the representations of one model with those of another model.

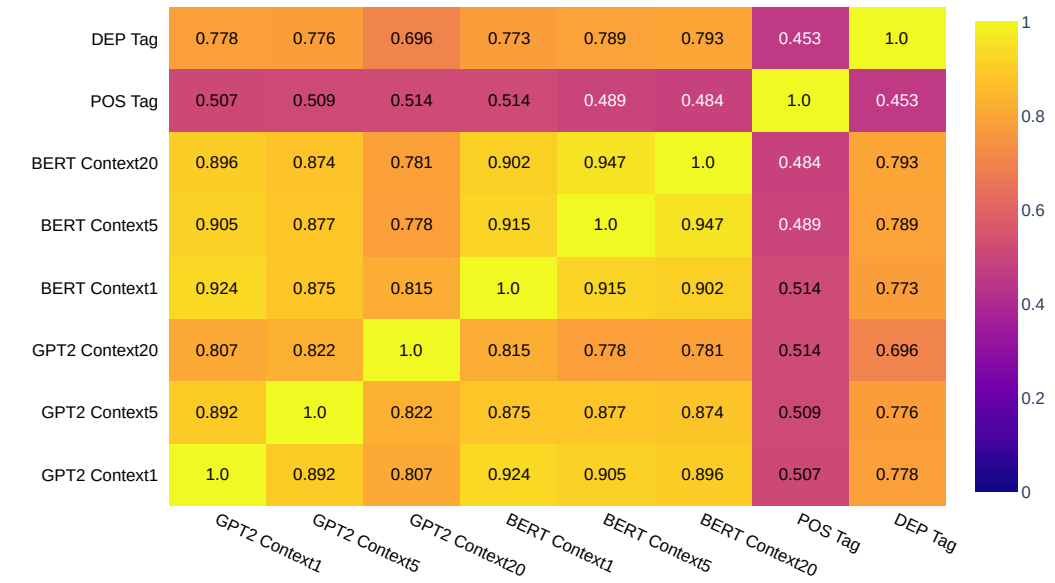



Figure 5.13: BERT vs. GPT2 - Task Similarity (Pearson Correlation Coefficient) constructed from the model-wise brain predictions averaged across various delays. We observe a high correlation only between BERT Context 5 vs. BERT Context 20, GPT2 context 1 vs. BERT Context 5.

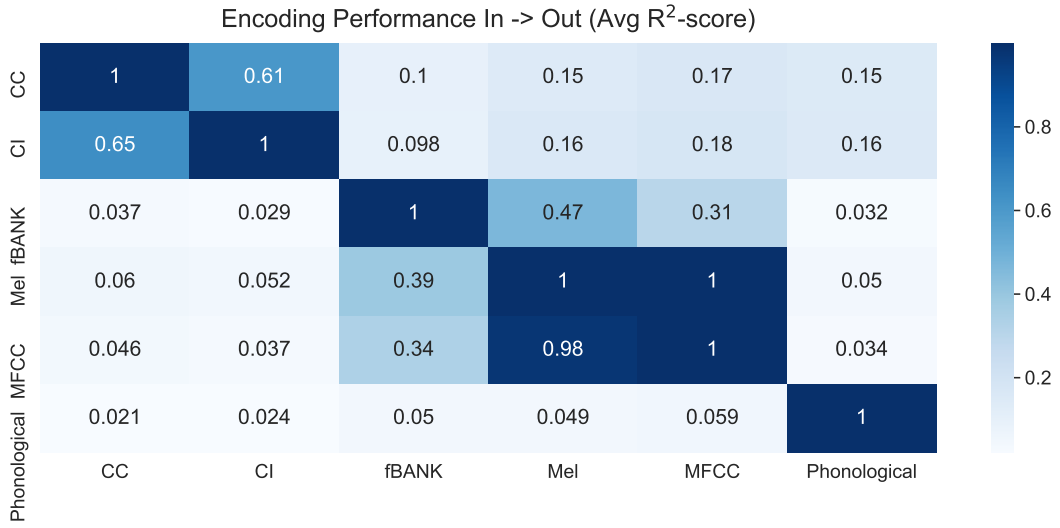



Figure 5.14: Basic speech features vs. syntactic embeddings - Average $R^2$-score was calculated by encoding the representations of one model with those of another model.

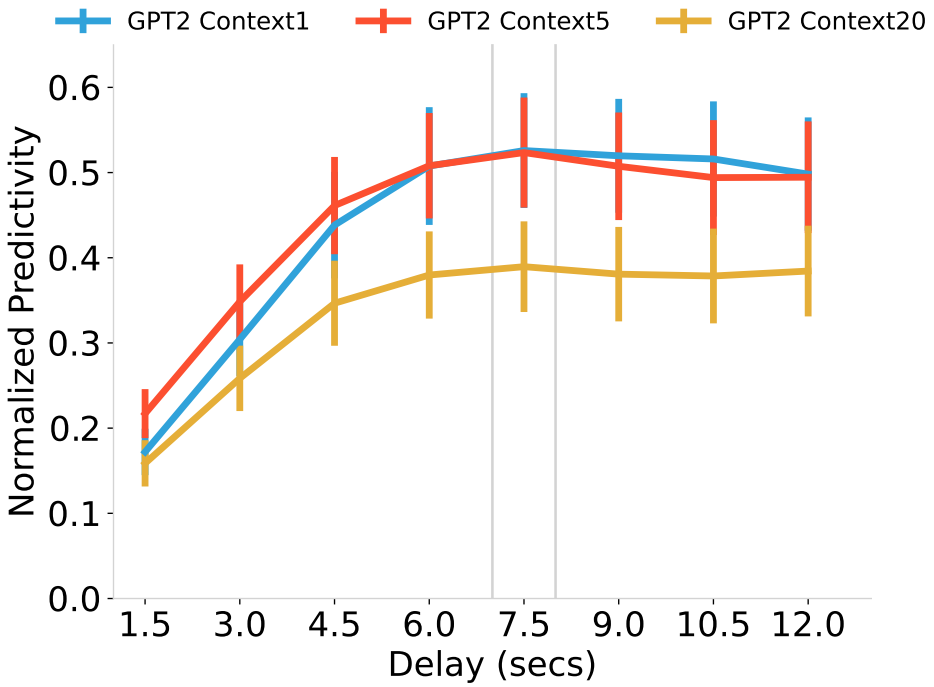



Figure 5.15: GPT2 Whole Brain Normalized Predictivity: The plot provides a comparison of delay-wise performance for various stimuli representations, averaged across subjects and layers. The vertical grey line serves as a reference point, indicating delay5 (7.5 seconds).

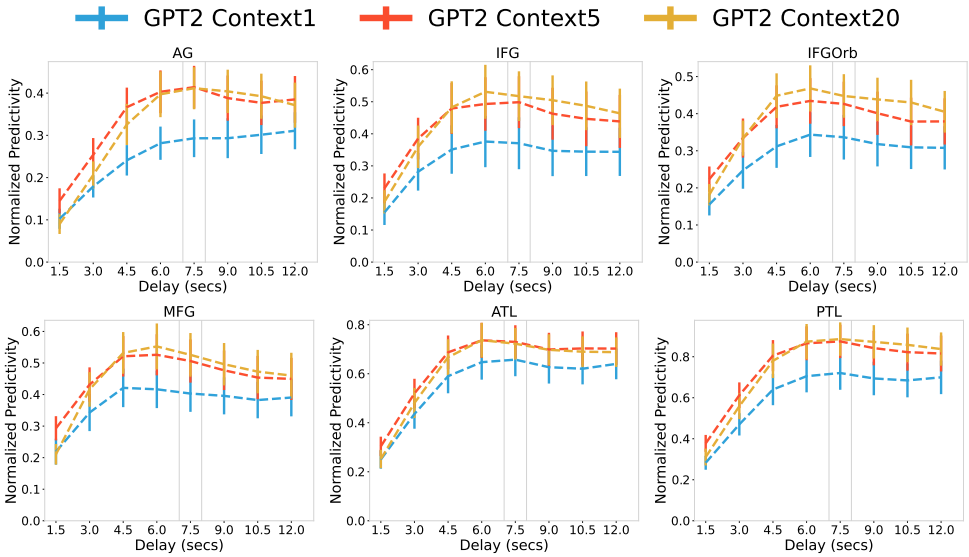



Figure 5.16: Language ROIs Normalized Predictivity.

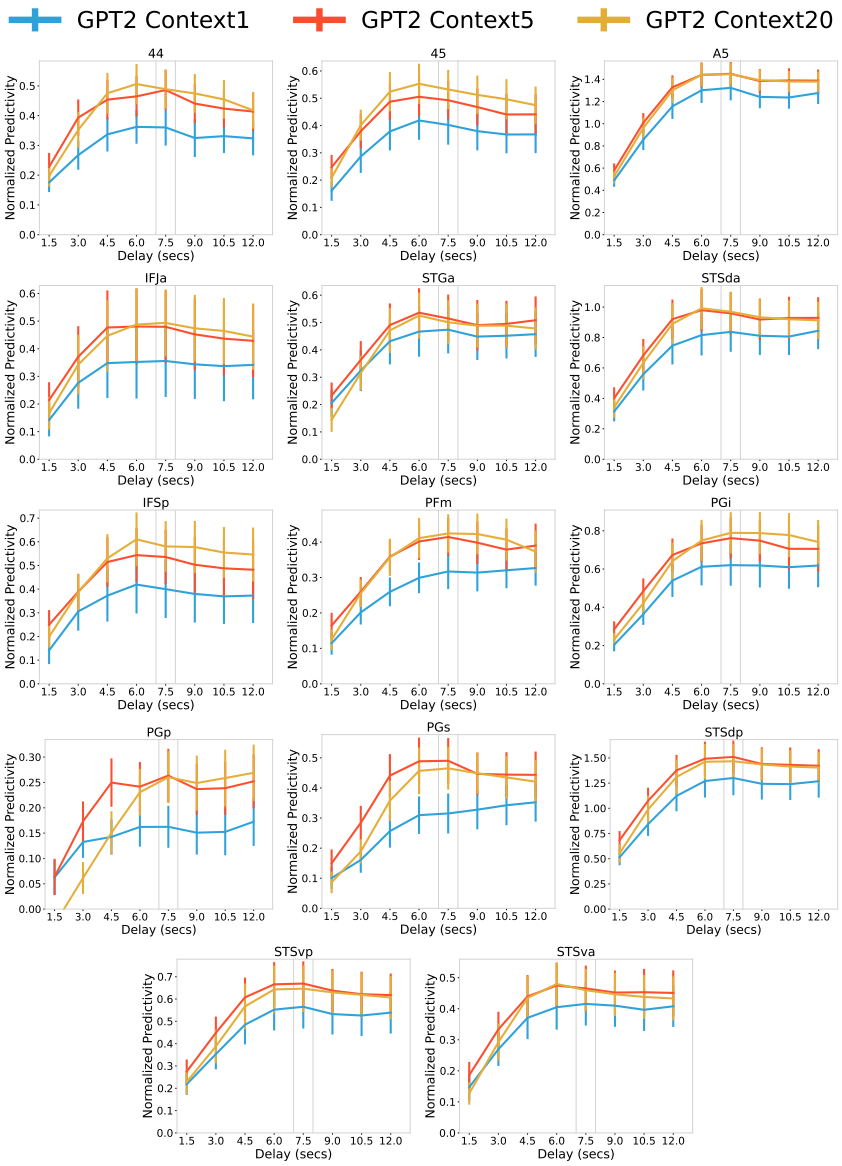



Figure 5.17: Language sub ROIs Normalized Predictivity.

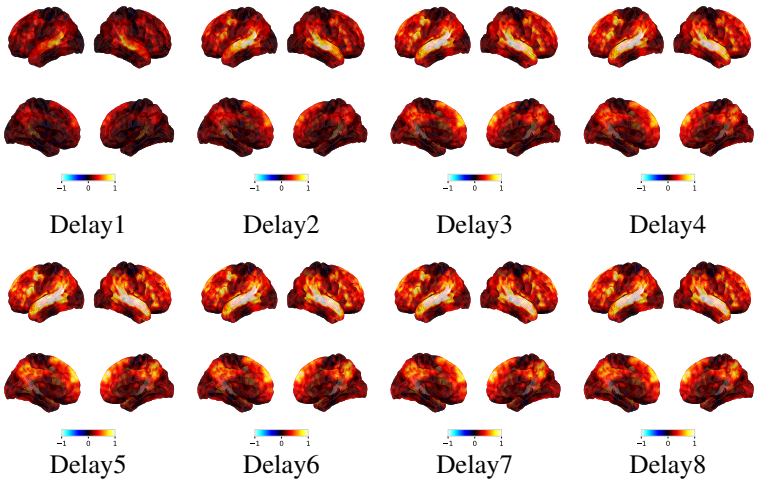



Figure 5.18: BERT Context20: Brain Maps: Voxelwise normalized brain predictivity average across layers and subjects for different HRF delays.

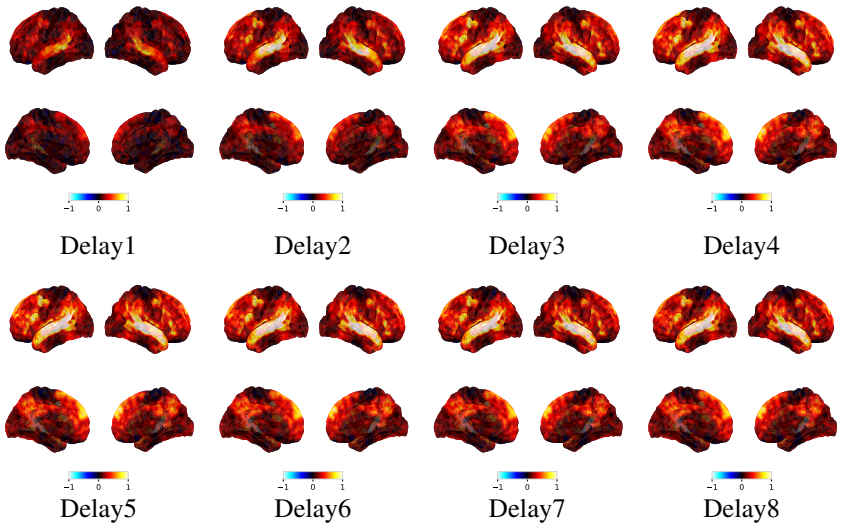



Figure 5.19: Brain Maps for GPT2 Context5: Voxelwise normalized brain predictivity average across layers and subjects for different HRF delays.

| Normalized Predictivity | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **Llama Context1** | -20.2* | -9.93* | -0.46 | **+2.14** | 38.32 | -0.91 | -0.9 | -2.59 |
| **Llama Context5** | -22.15* | -11.87* | -1.14 | **+2.13** | 42.01 | -2.55* | -1.88 | -2.12 |
| **Llama Context20** | -20.04* | -10.98* | -0.54 | **+3.24** | 40.6 | -2.23 | -1.51 | -1.89 |

(a) IFGOrb

| Normalized Predictivity | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **Llama Context1** | -32.38* | -16.58* | -6.24* | -0.13 | 47.93 | **+0.94** | +0.73 | -0.5 |
| **Llama Context5** | -35.8* | -18.78* | -4.81 | **+1.33** | 57.39 | -2.56 | -2.42 | -2.48 |
| **Llama Context20** | -34.24* | -18.25* | -6.55* | **+0.73** | 58.11 | -3.0 | -3.01 | -2.9 |

(b) IFG

| Normalized Predictivity | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **Llama Context1** | -30.93* | -13.22* | -2.33 | **+1.94** | 48.99 | -1.21 | -0.68 | -2.26 |
| **Llama Context5** | -29.75* | -13.47* | +0.52 | **+3.72** | 51.59 | -3.9* | -2.45 | -3.15 |
| **Llama Context20** | -29.87* | -14.69* | -0.91 | **+3.61*** | 52.07 | -4.92* | -2.98 | -4.35* |

(c) MFG

| Normalized Predictivity | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **Llama Context1** | -45.33* | -24.32* | -6.39* | **+1.18** | 65.51 | -2.07 | -1.17 | -0.63 |
| **Llama Context5** | -45.21* | -25.54* | -6.74* | **+1.23** | 67.3 | -1.75 | +0.22 | 1.17 |
| **Llama Context20** | -43.71* | -26.26* | -9.7* | -0.09 | **68.41** | -1.58 | -0.67 | 0.17 |

(d) ATL

| Normalized Predictivity | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **Llama Context1** | -53.97* | -33.22* | -11.72* | 0.0 | **81.85** | -1.37 | -1.56 | -2.53 |
| **Llama Context5** | -54.39* | -33.09* | -11.85* | **+0.13** | 87.74 | -2.43 | -0.51 | -0.99 |
| **Llama Context20** | -50.52* | -32.2* | -13.94* | -1.19 | **88.53** | -2.6 | -1.3 | -1.15 |

(e) PTL

| Normalized Predictivity | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 |
|---|---|---|---|---|---|---|---|---|
| **Llama Context1** | -25.59* | -14.56* | -5.76* | **+0.69** | 34.4 | -0.65 | -1.05 | -2.17 |
| **Llama Context5** | -25.83* | -14.7* | -4.93 | **+1.5** | 39.08 | -1.28 | -0.71 | -3.34 |
| **Llama Context20** | -24.61* | -15.03* | -7.28* | **+0.52** | 42.27 | -3.48 | -1.98 | -2.42 |

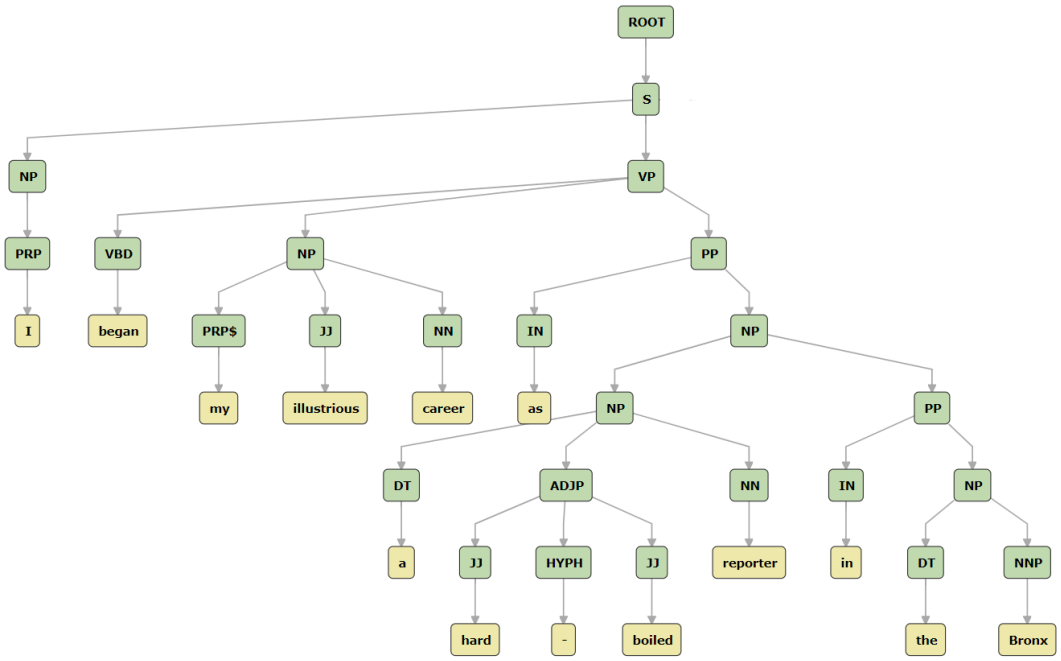
(f) AG

Table 5.6: Llama results: Variance analysis across delays by fixing the delay5 as constant for different language ROIs.


(a) I




(b) began




(c) Bronx

Figure 5.20: Complete trees for the words: I, began, and Bronx, for the sentence "I began my illustrious career as a hard-boiled reporter in the Bronx where I toiled for the Ram, uh, Fordham University's student newspaper."

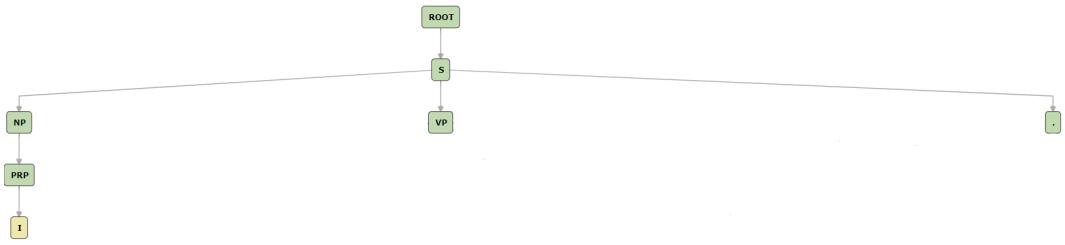



Figure 5.21: Incomplete trees for the word: I, for the sentence "I began my illustrious career as a hard-boiled reporter in the Bronx where I toiled for the Ram, uh, Fordham University's student newspaper."

# 6 MEG Encoding using Word Context Semantics in Listening Stories

Brain encoding is the process of mapping stimuli to brain activity. There is a vast literature on linguistic brain encoding for functional MRI (fMRI) related to syntactic and semantic representations. Magnetoencephalography (MEG), with higher temporal resolution than fMRI, enables us to look more precisely at the timing of linguistic feature processing. Unlike MEG decoding, few studies on MEG encoding using natural stimuli exist. Existing ones on story listening focus on phoneme and simple word-based features, ignoring more abstract features such as context, syntactic, and semantic aspects. Inspired by previous fMRI studies, we study MEG brain encoding using basic syntactic and semantic features, with various context lengths and directions (past vs. future), for a dataset of 8 subjects listening to stories. We find that BERT representations predict MEG significantly but not other syntactic features or word embeddings (e.g., GloVe), allowing us to encode MEG in a distributed way across auditory and language regions in time. In particular, past context is crucial in obtaining significant results.

This chapter has been finalized based on our previously published paper at $24^{th}$ INTERSPEECH conference (August 2023, Dublin, Ireland) [Oota et al., 2023e].

## 6.1 Introduction

Over the past decade, Brain-Computer Interface (BCI) helped to make significant progress in understanding language processing in the brain using a popular computational paradigm: Brain encoding, the process aiming to map stimuli features to brain activity. The central aim of brain encoding for language processing analysis is to unravel how the brain represents linguistic knowledge (i.e., semantic and syntactic properties) and carries out sentence-processing information [Wehbe et al., 2014, Huth et al., 2016, Caucheteux et al., 2021b, Reddy and Wehbe, 2021, Zhang et al., 2022a] by modeling the effect of such information on brain recordings. For instance, using functional Magnetic Resonance Imaging (fMRI) brain recordings, several previous studies have investigated the alignment between text stimuli representations extracted from language models (e.g., Bi-directional Encoder Representation Transformer (BERT) [Devlin et al., 2019]) and brain recordings of people comprehending language [Toneva and Wehbe, 2019, Schrimpf et al., 2021b, Oota et al., 2022c,b, 2023c].

While a large part of brain encoding literature uses fMRI brain recordings to study linguistic contrasts involved in language processing, the low temporal resolution of fMRI makes it challenging to link brain activation to specific linguistic processes. Conversely, MEG recordings have a better temporal resolution (generally understood as the smallest period of brain activation that can be distinguished) and allow us to understand better the neural dynamics of the underlying language processing network. However, few studies use MEG to study how word embeddings such as BERT can be related to the brain activity of subjects reading one word at a time from a story [Toneva and Wehbe, 2019]. We propose to uncover insights into human sentence processing during a naturalistic story-listening task.

Studies using word embedding representations and fMRI have revealed that syntactic features are distributively represented across brain language networks and overlapped mainly with semantic networks [Reddy and Wehbe, 2021, Zhang et al., 2022a]. Despite the great strides in learning sentence comprehension at a functional level, many problems could benefit from further improvements in understanding sentence structure and meaning at the temporal level. Therefore, investigating how the brain encodes semantic and fine-grained syntactic features of words using MEG recordings seems crucial to understanding the timing of language comprehension mechanisms. Some critical questions remain to be explored: (1) How much context is maintained through time to process words? (2) Is the direction of context important (past context vs. future context)? The main objective of this work is to address these questions using MEG activity, in time at different sensor locations, for both syntactic and semantic representations during naturalistic story listening.

**Brain Regions of Interest (ROIs) for sentence processing:** Several MEG studies report evidence from well-formed natural language expressions for the role of the left posterior temporal lobe (PTL) in incremental syntactic processing. Similarly, post-nominal adjectives were relayed to the inferior frontal gyrus (IFG) and influence of semantic type in the left anterior temporal lobe (ATL) [Flick and Pylkkänen, 2020, Kochari et al., 2021, Law and Pylkkänen, 2021]. Further, Toneva et al. [2020] conclude that the involvement of a language network with task-specific settings (e.g., question-answering task) is localized to the frontal and the left temporal lobes. These findings correspond to many fMRI studies [Cara-

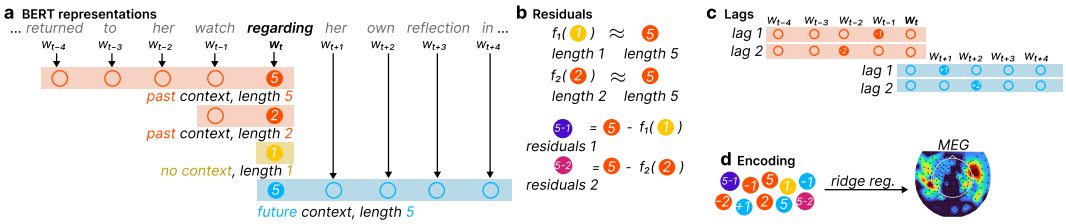



Figure 6.1: *Global schema of the study.* Each circle is a vector embedding for a particular word. Here, "*regarding*" is the current word $w_t$ encoded. ***(a)*** BERT representations are computed with varying context lengths and directions (past vs. future). "*Past (future) context of length* $n$" means that word $w_t$ is encoded with its $n$ preceding (following) words $w_{t-1}, \dots w_{t-n-1}$ ($w_{t+1}, \dots w_{t+n-1}$). (Red) Past contexts (lengths 5 and 2). (Blue) Future context 5. (Yellow) Absence of context (context 1: static representation of the word). ***(b)*** For a given past (future) context around a word $w_t$, we name "*residuals* $n$" the results of filtering information of the $n$ nearest words from the representation of current word $w_t$ (e.g. past residuals 2 are the result of removing information of past context 2 $w_t^2 = [w_t, w_{t-1}]$ from past context 5 $w_t^5$). Filtering is performed by first fitting a length $n$ sub-context $w_t^n$ to the 5 context $w_t^5$, and then computing residuals between estimated and real 5 context. ***(c)*** For a given past (future) context around a word $w_t$, we name "*lag* $l$" the representation of the word $w_{t-l}$ ($w_{t+l}$). ***(d)*** For each word, all of these representations are used to predict the subject's MEG activity at word onset in the story using ridge regression.

mazza and Zurif, 1976, Friederici et al., 2006, Friederici, 2011, Zaccarella and Friederici, 2015, Humphries et al., 2006, Rogalsky and Hickok, 2009, Bemis and Pylkkänen, 2011]. However, the time at which different brain regions are sensitive to distinct syntactic and semantic properties remains unclear.

**Word stimulus representations for brain encoding**: Several studies have used basic syntactic features such as part-of-speech (POS), dependency relations (DEP), complexity metrics [Caucheteux et al., 2021a, Reddy and Wehbe, 2021, Oota et al., 2023d], and semantic word embeddings [Oota et al., 2018, Jain and Huth, 2018, Hollenstein et al., 2019, Toneva and Wehbe, 2019, Vaidya et al., 2022, Oota et al., 2022b] to represent words for fMRI brain encoding with text stimulus. However, modeling these basic syntactic and semantic features for MEG recordings still needs to be explored. In this paper, to understand when the brain processes linguistic structure in sentences, we leverage text representations using basic syntactic features and semantic features with various context lengths, directions (past vs. future), and within-context relative importance.

Overall, our main contributions are as follows. (1) We explore: (a) basic syntactic features, (b) GloVe embeddings, and (c) semantic BERT embeddings for MEG brain encoding. We found that only BERT embeddings were predictive of MEG activity. (2) We find that prediction of the MEG activity using BERT is in regions such as the bilateral temporal lobes, frontal lobe and parietal lobe between 250ms to 750 ms (word onset is at 200ms). (3) We report that past context has greater predictive power than future context. When dealing with past context, $R^2 scores$ are proportional to context length.

## 6.2 Feature Representations

We used different features computed per word to simultaneously test different syntactic and semantic representations.
**(1) Basic Syntactic Features:** Similar to prior studies in Wang et al. [2020a], Reddy and Wehbe [2021], Zhang et al. [2022a], we use various multi-dimensional syntactic features such as Complexity Metrics (Node Count (NC), Word Length (WL), Word Frequency (WF)), Part-of-speech (POS) and Dependency tags (DEP), described briefly below. **Node Count (NC)** The node count for each word is the number of subtrees completed by incorporating each word into its sentence. **Word Length (WL)** Word length is the number of characters present in the word. **Word Frequency (WF)** Word frequency reports log base-10 of the number of occurrences per billion of a given word in a large text corpus. **Syntactic Surprisal (SS)** Syntactic surprisal is computed using incremental top-down parser [Roark, 2001]. It measures how unexpected it is to read a given word in the current syntactic context. Both of these metrics aim to measure the amount of effort that is required to integrate a word into the syntactic structure of its sentence. **Part-of-speech (POS)** We use the Spacy English dependency parser [Honnibal and Montani, 2017] to extract the Part-of-speech (POS). We generate a one-hot vector for each word in which the corresponding POS tag location is 1 and the remaining tag values are 0. **Dependency tags (DEP)** We use the Spacy English dependency parser [Honnibal and Montani, 2017] to extract the dependency tags. In DEP, we generate a one-hot vector for each word and dependency tag in which the corresponding dep tag location is 1, and the remaining tag values are 0. **(2) Semantic Features** We use two semantic representations: (1) GloVe (distributed word representations) [Pennington et al., 2014] and (2) BERT (contextualized representations) [Devlin et al., 2019], described briefly below. **GloVe:** word vectors (each word is a 300-dimension vector) [Pennington et al., 2014], and the model always represents unique embedding irrespective of the word appearing in different contexts.
**BERT:** Given an input sentence, the pretrained BERT [Devlin et al., 2019] outputs word representations at each layer. In this paper, we have used the pretrained BERT-base model. We have not performed any fine-tuning here. Since BERT embeds a rich hierarchy of linguistic signals: surface information at the bottom, syntactic information in the middle, semantic information at the middle to higher layers [Jawahar et al., 2019]; hence, we use the #words $\times$ 768D vector from the intermediate layer (layer-7) to obtain the embeddings.
**(3) Varying the Context Length of BERT** To extract the stimulus features at different context lengths (C = 1, 2, 3, 4, 5, 20), we constrained the model with maximum C words as context length (Fig. 6.1 (a)). Since the BERT model processes whole sentences, we input all the C context-length words to the BERT model and use the representation of the last word for the past context and the first word for the future context. For instance, given a story of M words and considering the context length of 5, while the third word's vector is computed by inputting the network with ($w_1$, $w_2$, $w_3$), the last word's vectors $w_M$ is computed by inputting the network with ($w_{M-5}$, $\cdots$, $w_M$). Here, we extracted representations for both past and future contexts.
**(4) Residuals** To compute residuals from pretrained BERT representations at different context lengths, we use a ridge regression method in which the context $w_M$ (M=1,2,3) as input

and the context $w_5$ is the target vector (Fig. 6.1(b)). We compute the residuals by subtracting the predicted context from the actual context, resulting in the (linear) removal of a particular context from context $w_5$ (see Fig. 6.1 for a schematic). Because the MEG brain prediction method is also a linear function (see Section 6.4), this linear removal limits the contribution of the word importance to the eventual brain prediction performance.
**(5) Lags** To extract lag $l$ representations, we take as an embedding vector, for a given context length $t$, the vector of the word $w_{t-l}$ for past context (or $w_{t+l}$ for future context) (Fig. 6.1(c)). Contrary to residuals, these lag representations still contain information from the current word $w_t$. Encoding MEG using lag representations assesses how lag word information is correlated to current word MEG activity.

## 6.3 Dataset and Experiments

We used data from 8 subjects of the MEG-MASC dataset [Gwilliams et al., 2023a]. The activity from 208 MEG sensors was recorded. At the same time, each subject listened to naturalistic spoken stories selected from the Open American National Corpus (*"Cable spool boy"*, *"LW1"*, *"Black willow"* and *"Easy money"*).
**MEG preprocessing** We performed the minimal processing steps described in Gwilliams et al. [2023a]. On raw MEG data and for each subject separately, using *MNE-Python* defaults parameters, we (i) bandpass filtered the MEG data between 0.5 and 30.0 Hz, (ii) temporally-decimated the data 10x, (iii) segmented these continuous signals between -200 ms and 600 ms after word onset, (iv) applied a baseline correction between -200 ms and 0 ms, and (v) clipped the MEG data between the fifth and ninety-fifth percentile of the data across channels.
**Word Processing** Since MEG data is sampled at a higher rate ($1000Hz$) than word presentation, epoching and downsampling yields, for each word, 81 times points recorded at 208 sensors. There are total of 8567 words across four stories. In our experiments, for each word, the model makes a prediction of MEG activity for all of these 16848 = 208 $\times$ 81 values. Here, each word is transformed into one of the feature representations described in section 6.2.

## 6.4 Models and evaluations

### 6.4.1 Encoding Model

We extracted different features describing each stimulus word to explore how and when syntactic and semantic specific features are represented in the brain when listening to stories. We used them in an encoding model to predict brain responses (Fig. 6.1(d)). MEG encoder models attempt to predict brain responses associated with each MEG sensor and each time point when given audio stimuli (spoken words in our case). We trained a model per subject separately. Following the literature on brain encoding [Wehbe et al., 2014, Toneva et al., 2020, Caucheteux et al., 2021b, Reddy and Wehbe, 2021], we used a ridge regression as an encoding model. The ridge regression objective function for the stimulus features is

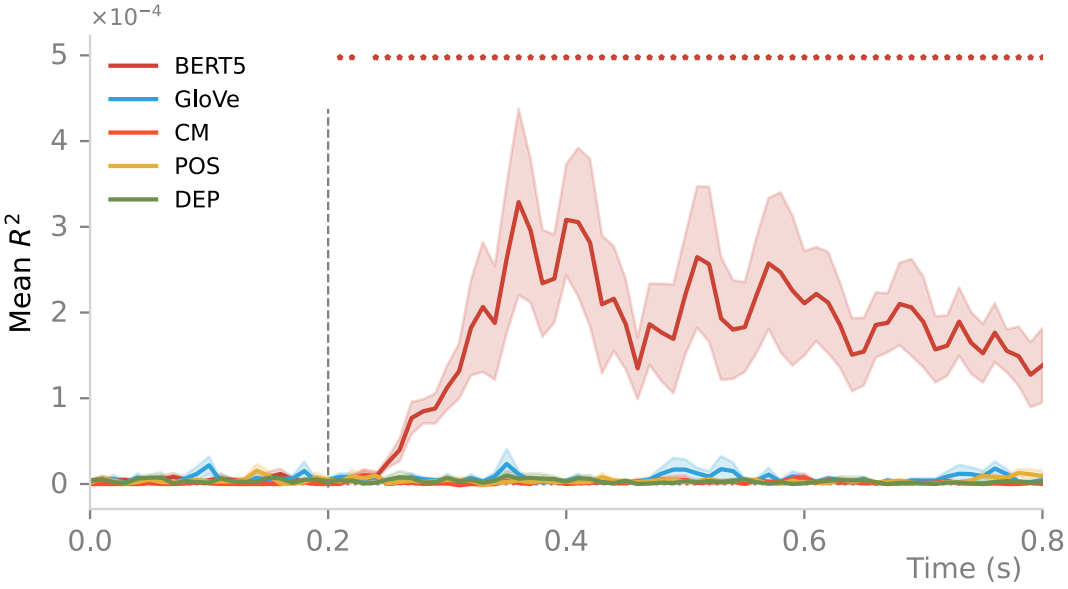



Figure 6.2: *Only BERT representations significantly encode MEG activity.* Plain lines represent mean significant $R^2$ score (permutation test, $p < 0.05$, FDR correction) between predicted and real MEG activity, across sensors and subjects. Areas around lines represent standard error across subjects. Dots above the figure represent significant difference with 0, for all timestep (one-sample t-test, $p < 0.05$, FDR correction) (color is matching the legend). Word onset is at 200ms.

$f(X_s) = \min_{W_s} \|Y_b - X_s W_s\|_F^2 + \lambda \|W_s\|_F^2$. Here, $X_s$ denotes the input stimuli representation, $W_s \in \mathbb{R}^{F_s \times LT}$ are the learnable weight parameters, $F_s$ denotes the number of features in stimuli representation (768), $L$ corresponds to number of MEG sensors (208), $T$ represents the time dimension of the brain activity (81), $s$ denotes the sample stimulus $s \in \mathbb{R}^{F_s}$, $\|.\|_F$ denotes the Frobenius norm, and $\lambda > 0$ is a tunable hyper-parameter representing the regularization weight. $\lambda$ was tuned on a small disjoint validation set obtained from the training set.

### 6.4.2 Cross-Validation

We follow 4-fold (K=4) cross-validation. All the data samples from K-1 folds (3 stories data) were used for training, and the model was tested on samples of the left-out fold (1 story data).

### 6.4.3 Evaluation Metrics

We compute the coefficient of determination $R^2$ score [Pedregosa et al., 2011] between real and predicted MEG activity to measure prediction performance for each sensor location and each timepoint within epochs. $R^2$ scores were then averaged over all epochs and across all folds. Along with $R^2$ score, we also use Root-Mean-Square (RMS) to measure

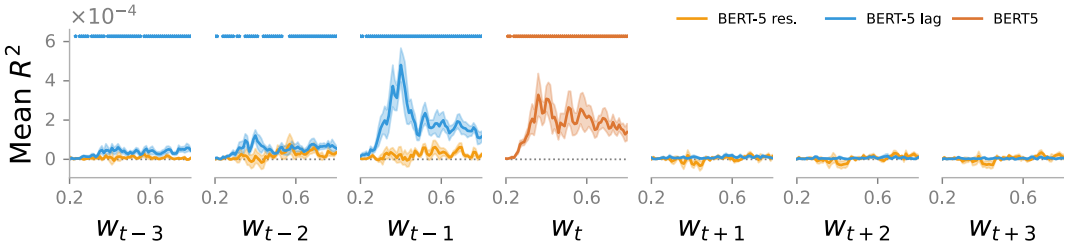



Figure 6.3: $R^2$-score performance of encoding for different lag and residuals of BERT representations. Plots $w_{t-n}$ report, for $n \in [3, 1]$, the performance of lag $n$ and residuals $n$ when encoding MEG activity at word $w_t$ using its past context. Similarly, plots $w_{t+n}$ reports performance of lag and residuals $n$ using future context. Plot $w_t$ displays performance of context 5. Lines, areas and dots figures same metrics as Fig. 6.2. Word onset is at 200ms.

the predicted evoked response, averaged across all MEG sensors, tasks and subjects. RMS scores are reported in supplementary materials. **Statistical Significance** We check $R^2$ scores statistical significance using a permutation test. We permute blocks of MEG predictions and compute $R^2$ scores between permuted predictions and real data 5000 times to estimate an empirical distribution of chance performance and corresponding p-values. Finally, the Benjamini-Hochberg False Discovery Rate (FDR) correction [Benjamini and Hochberg, 1995] is applied on all tests to control the type I error rate. **Implementation Details for Reproducibility** All experiments were conducted on a machine with 1 NVIDIA GEFORCE-GTX GPU with 4GB GPU RAM. We used ridge-regression with the following parameters: MSE loss function, and L2-decay ($\lambda$) varied from $10^1$ to $10^3$.

## 6.5 Results

In order to assess the performance of MEG encoder models learned using syntactic and semantic representations, we computed the $R^2$-score between predicted MEG and ground-truth recordings of the evoked response at word onset, across all sensors, folds and subjects. Each figure reports the average $R^2$-scores of the different features, where all values are first filtered by significance for each time point (i.e. we set to 0 the score values for sensors where $p < 0.05$ after the permutation test and FDR correction procedure described in section 6.4.3).

### 6.5.1 Encoding Performance of Syntactic and Semantic Methods

From Fig. 6.2, we make the following observations: (i) Only BERT-based feature representations significantly correlate to MEG activity, starting around 0.25s (0.05s after word onset). (ii) Basic syntactic (CM, POS and DEP), and non-contextual semantic features (GloVe) are, on average, not correlated with the considered window of MEG activity. These features poor performance may be explained by their overly simple nature or their limited contextual information. To better visualize the predicted MEG performance using these

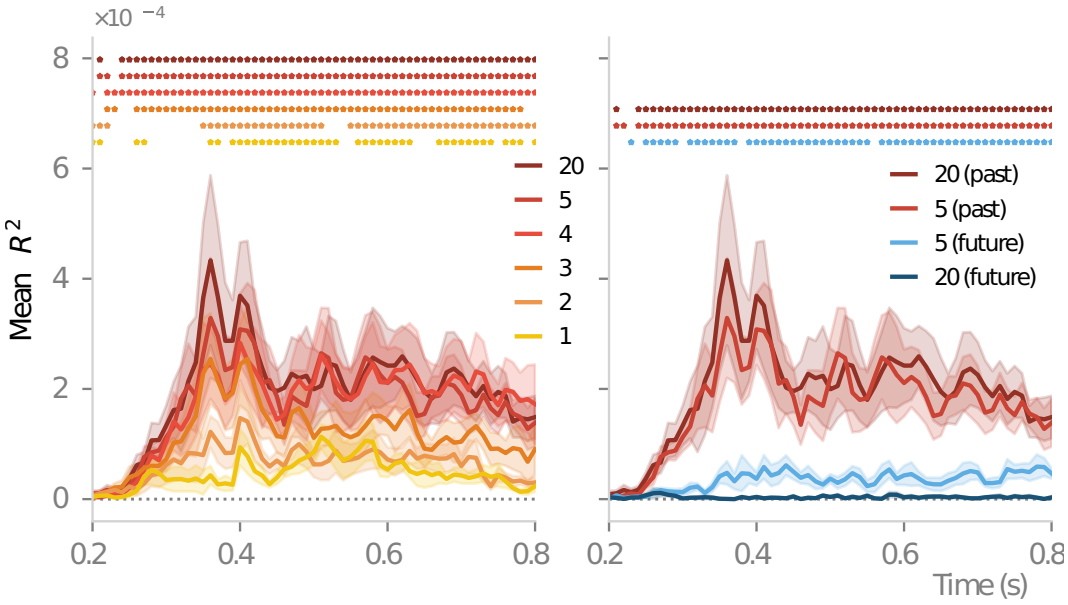



Figure 6.4: *Long past contexts enable better encoding than future or short-scale present contexts.* (left) $R^2$ given BERT representation context length, from 1 (no context) to 20. (right) $R^2$ given BERT representation context direction (past vs. future) and length (5 vs. 20). Lines, areas and dots figures same metrics as Fig. 2. Word onset is at 200ms.

simple features, we report the RMS plot in supplementary. It is observed that the RMS plot for these methods is not closer to the original MEG in comparison to BERT.

### 6.5.2 Contextual BERT Embeddings: Effect of Length

To assess whether the direction and length of context are important for predicting MEG activity during story listening, we report the $R^2$-score performance from both past and future BERT contextual representations in Fig. 6.4. From Fig. 6.4 (left), we observe that context length plays a crucial role in predicting MEG activity. The performance of this prediction is proportional to the length of the context. However, above a context length of 5, no significant improvement in MEG predictivity is noticeable. Moreover, the context performance difference is mainly observed between 300ms and 425ms (100–325ms from word onset). This suggests that MEG activity results from integrating past auditory information on a short time horizon.

### 6.5.3 Contextual BERT Embeddings: Effect of Direction

From Fig. 6.4 (right), we observe a significant effect on the direction of context. All features created from future context display a low correlation with features created using past context. Interestingly, this effect is inversely proportional to context length for future context, where BERT features extracted from a future context of length 5 achieve better $R^2$ scores than the same features created from a future context of length 20. This suggests that

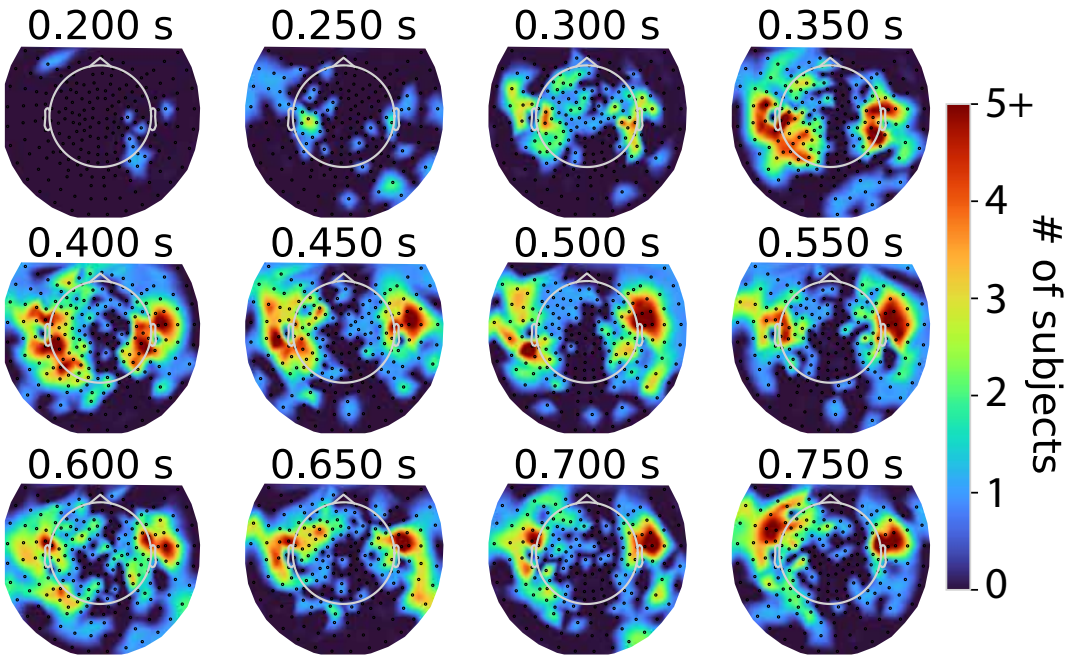



Figure 6.5: Significantly predicted MEG activity for each timepoint and each sensor position (permutation test, $p < 0.05$, FDR correction) using BERT past context 5 word embeddings. Color denotes, for each sensor and timepoint, the number of subjects whose MEG activity was significantly predicted. Word onset is at 200ms.

the MEG brain activity mostly correlates to past and current contexts. Inference of future context could not be detected, and if present, only on a short time horizon. Since length five context contains the current word, the relative importance of the current word could account for its relatively correct performance in its 5-word context, which is diluted in a 20-word context.

### 6.5.4 Contextual BERT Embeddings (Residuals vs. Lag)

To investigate whether the removal of word-level information (current word, two nearest words, three nearest) from context has any effect in predicting MEG activity, we report the $R^2$-score performance of residuals in both past and future contexts, as shown in Fig. 6.3. We also report the lag representations performance, which represents the performance of the previous word representations in predicting the current word MEG. From Fig. 6.3, we make the following observations: (i) Complete removal of current word information from past context through residual representations (i.e., $w_{t-n}$) has a significant drop in $R^2$-score. (ii) Similarly, in the future context, the $R^2$-score performance of residuals is always zero or significantly below chance. (iii) Unlike residuals, lag representations display significant performance for lag 1, 2 and 3 in the past context, with lag 1 demonstrating the most notable performance equal to current word prediction performance of MEG activity. (iv) Similar to future context residuals, future context lag representations yield below-chance performance. From these results, we hypothesize that the current word MEG activity is the

product of short-term past context and current word information. Both pieces of information are required to render MEG activity at a given word onset accurately. Future context information is not detectable in MEG activity.

### 6.5.5 Cognitive Insights

Fig. 6.5 reports MEG sensor locations which are significantly predicted across subjects by 5-context BERT representations (permutation tests, $p < 0.05$, FDR correction). Best brain MEG alignments are in the bilateral temporal and frontal regions between 250ms to 750ms (word onset is at 200ms).

## 6.6 Discussion & Conclusion

In this paper, we evaluated the alignment of basic syntactic, distributed word embeddings, and contextualized word representations (varying different context lengths, past vs. future context, residuals, and lags) with MEG brain responses in time. We showed that BERT representations, contrary to other features or GloVe, lead to a significant prediction in brain alignment across auditory and language regions between 50-550ms (250ms to 750ms with word onset at 200ms). Noteworthy, this prediction performance is a function of the amount of available past context, and only past context future or current word.

It is surprising that BERT current word representation alone $w_t^1$ (BERT-1) allows so weak predictions compared to $w_t^{past\geq 3}$ (BERT with contexts higher than 3) (Fig. 6.4). Moreover, lag results of Fig. 6.3 shows that the previous BERT-5 word $w_{t-1}^{past3\&future1}$ allows higher $R^2$ score than current word with low context $w_t^{past\leq 5}$. Additionally, it is surprising that near future context $w_t^{future5}$ which includes the current word is not relevant for MEG prediction, as if the brain was making no or very few predictions of future incoming words.

This suggests that the "word encoding center of mass" is few words behind the current word, as if the brain would wait for more future context before encoding "fully" the word, or similarly that the current representation of the incoming word is encoded in a transient representation that is changing until the next words come in. This is coherent with previous studies Gwilliams et al. [2022] from that showed that the several past phonemes information (with position and order in sequence) are kept in memory, and that current incoming word lexical information is retrieved in a context-sensitive manner (rather than using the most probable lexical category of the word) [Gwilliams et al., 2023b].

We hypothesise that such "encoding center of mass" lying in the past is also what is happening in the speaker's brain. Songbirds such as canaries need to keep track of long-time dependencies in the sequences of phrases performed in order to produce the next syllables at syntax branching points correctly [Cohen et al., 2020]: the brain area managing these dependencies preferentially encodes past actions rather than future actions. Specific neuron populations preferentially encoding past actions were actually more active during the rare phrases that involve history-dependent transitions in song [Cohen et al., 2020]. This is also coherent with the results of Gwilliams et al. [2022] where phoneme representations are sustained longer when linguistic identity is uncertain. Overall, it seems that the representations

of past events or actions are kept in memory until they have been used to disambiguate future events/actions.

# *Appendix for: MEG ENCODING USING WORD CONTEXT SEMANTICS IN LISTENING STORIES*

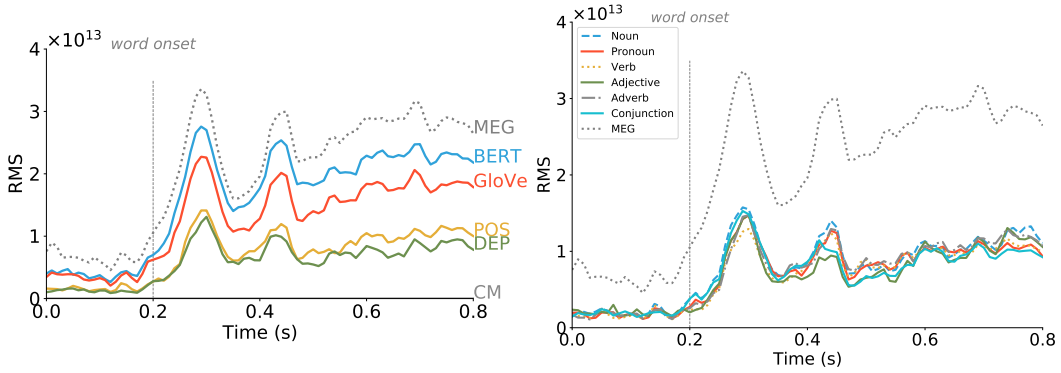
---

**RMS Plots** In order to assess the performance of the MEG encoder models learned using the representations from a syntactic and semantic methods, we computed the Root-mean-square (RMS) of the evoked response at word onset, time averaged across all sensors, folds and subjects, for both ground-truth and predicted MEG recordings using different feature sets, as shown in Fig. 6.6.

**FDR Correction: P-values** We report the number of subjects for which predictions resulted in higher than chance average $R^2$ score ($p < 0.05$, FDR correction), for every MEG sensor and times around word onset, in the Fig. 6.7. Prediction performance was measured for each subject and feature representation, using $R^2$ score, averaged over all epochs and all folds. $R^2$ scores were controlled for significance using the permutation test procedure described in main paper Section 4. From this figure, we observe that BERT contextual representations are significantly correlated with the MEG signal for most of the subjects, between word onset (200ms) and end of epoch (around 700ms and onward). Other representations predict the MEG signal with very local significance in space and time, and for subsets of subjects only.

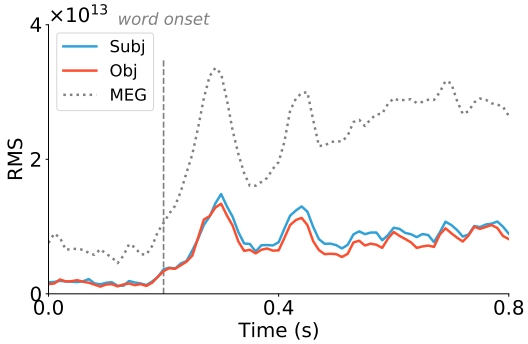
**Topomaps for All Feature Representations** To further investigate which brain activities are predicted by each feature representation, we reported in Figs. 6.8 and 6.10, the count of subject whose MEG signal was significantly predicted on topomaps, using the *mne* package [Gramfort et al., 2013]. Fig. 6.10 presents the same results as Fig. 6.7, for BERT predictions only, where subject count with significantly correlated predictions is displayed for each sensor location on the scalp at a 50ms time intervals during epoch (word onset at 0.2s). We observe the following insights: (1) the effect of contextual word representations begins at 250ms to 300ms (word onset at 200ms), distributed in the left and right temporal lobes. (2) Further, the effect extends until the end of the considered time, with major contribution in the left-right temporal lobes, left-right frontal lobes and left parietal lobe between 350 - 750ms. BERT representations learn linguistic structure from full sentence [Jawahar et al., 2019]; hence these representations involve different linguistic properties. To differentiate the syntax and semantic effects, we further report the topomaps for CM, POS tags, DEP tags and GloVe in Fig. 6.8.

**Syntactic Effects** From Fig. 6.8, it can be seen that the syntax information provided by the CM, POS tags and DEP tags have no clear effect at all, but without correction there is a low effect (that might not be significant).




(a) Different Feature Representations

(b) POS Tags




(c) DEP Tags

Figure 6.6: Predicted MEG Root-Mean-Square (RMS) activity following word onset, (a) encoded from different feature representations, (b) encoded from different POS tags, (c) encoded from different DEP tags. Activity is averaged over subjects and stories. Dashed gray (MEG): Ground truth RMS. The legend displays colors corresponding to each feature representation in fig (a), POS tags (Noun, Verb, Adjectives) in the fig (b) and DEP tags (Subject and Object) in the fig (c).

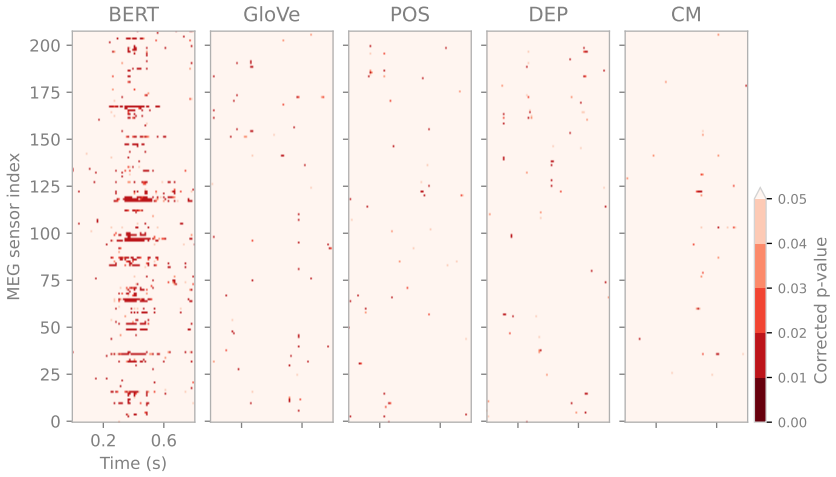



Figure 6.7: Number of subjects where encoding models display significant average $R^2$ on all folds ($p < 0.05$ with 5000 permutations and FDR correction), for each sensor and each time point.

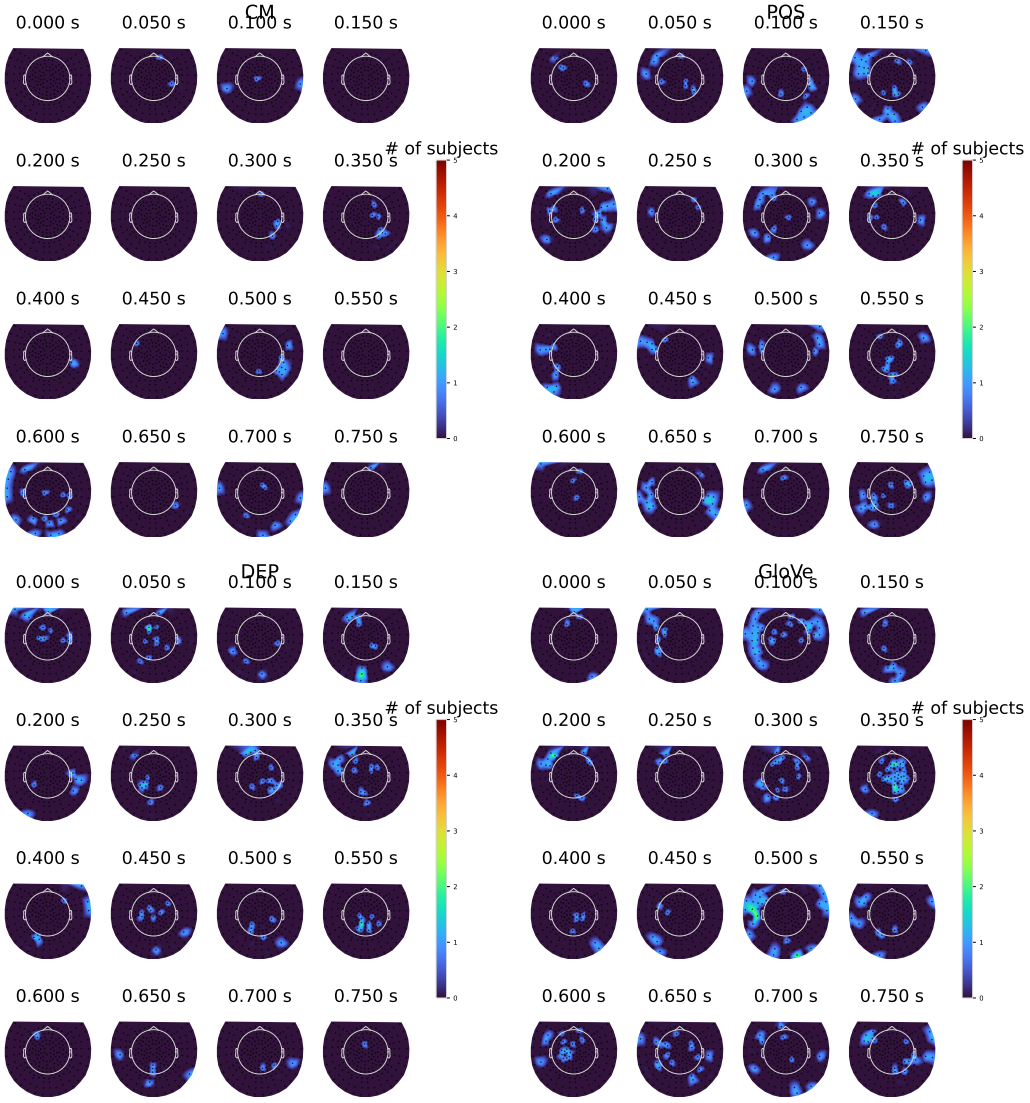



Figure 6.8: The significantly predicted MEG recordings (p-values corrected using FDR correction) for BERT word embeddings: 5 words context left and right.

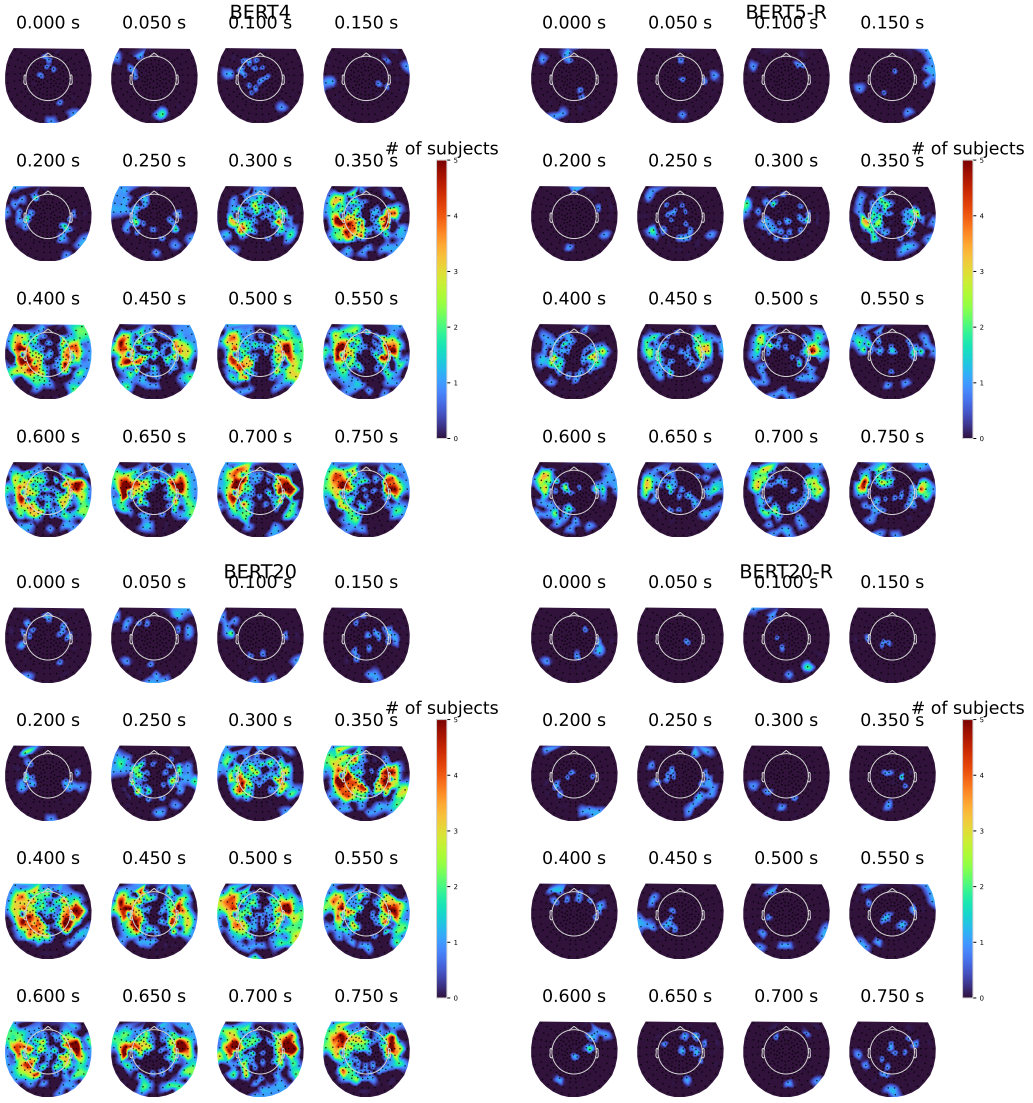



Figure 6.9: The significantly predicted MEG recordings (p-values corrected using FDR correction) for BERT word embeddings: context lengths (4-past, 5-future, 20-past, 20-future).

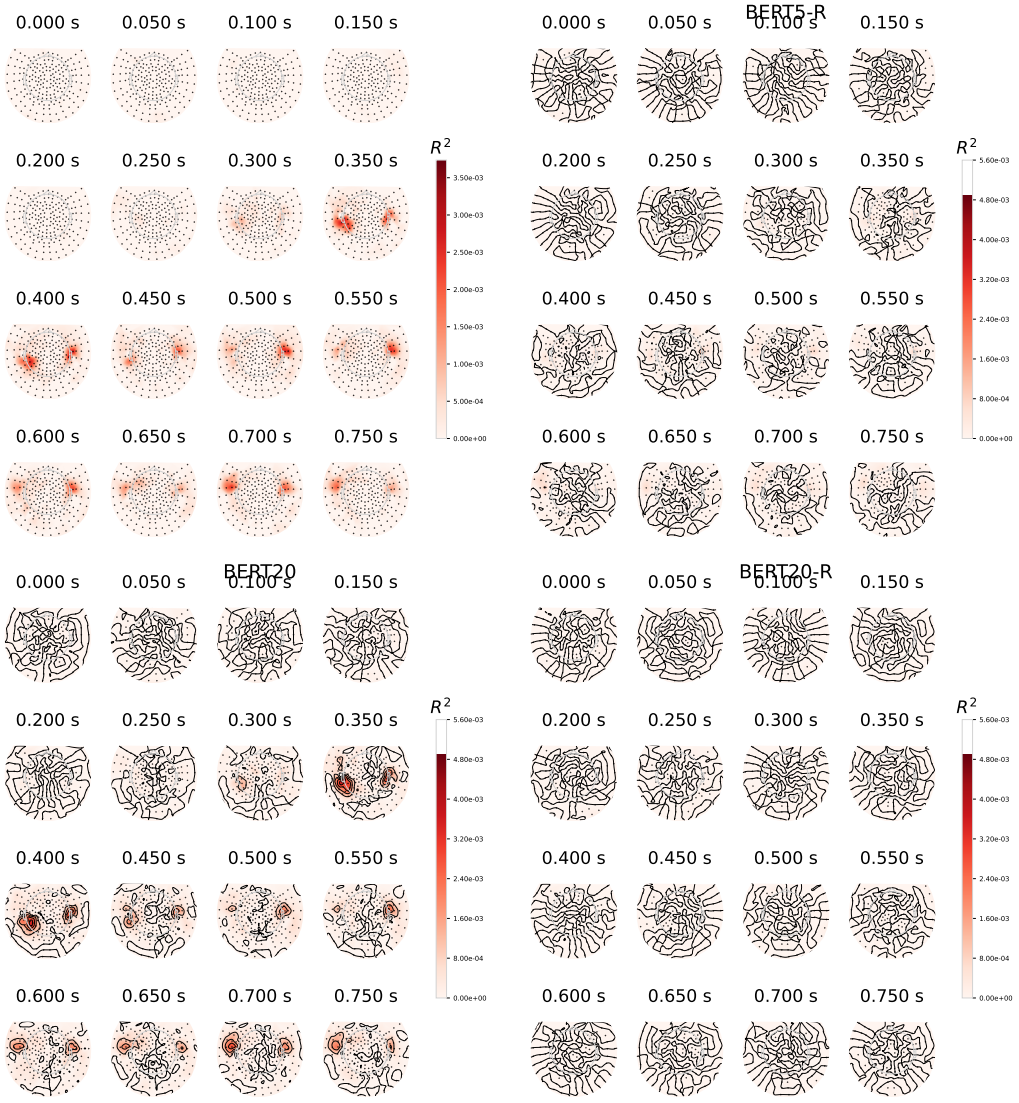



Figure 6.10: Average $R^2$-score across all the subjects for significant sensors for BERT word embeddings: context lengths (5-past, 5-future, 20-past, 20-future).

# 7 Cross-Situational Learning Towards Language Grounding

How do children acquire language through unsupervised or noisy supervision? How does their brain process language? We take this perspective to machine learning and robotics, where part of the problem is understanding how language models can perform grounded language acquisition through noisy supervision. This would also help us understand how language models account for brain learning dynamics. Most prior works primarily focused on tracking the co-occurrence between individual words and referents to model how infants learn word-referent mappings. This paper studies cross-situational learning (CSL) with complete sentences, aiming to understand the brain mechanisms enabling children to learn mappings between words and their meanings from complete sentences in early language learning. We investigate CSL on a few training examples with two sequence-based models: reservoir computing (RC) and long-short term memory networks (LSTMs). Importantly, we study how robust these models are when dealing with several word embeddings, including One-Hot, GloVe, pre-trained BERT, and fine-tuned BERT representations. We apply our approach to three datasets with varying complexities. We observe that (1) One-Hot, GloVe, and pre-trained BERT representations are less efficient when compared to representations obtained from fine-tuned BERT. (2) ESN online with final learning (FL) yields superior performance over ESN online continual learning (CL), offline learning, and LSTMs, indicating the more biological plausibility of ESNs and the cognitive process of sentence reading. (3) LSTMs with fewer hidden units exhibit higher performance for small datasets, while LSTMs with more hidden units are needed to perform reasonably well on larger corpora. (4) ESNs demonstrate better generalization than LSTM models, especially with increasingly large vocabularies. These models can learn from scratch to link complex relations between words and their corresponding meaning concepts, handling polysemous and synonymous words. Moreover, we argue that such models can extend to help current human-robot interaction studies, particularly in language grounding, and better understand children's developmental language acquisition. We make the code publicly available https://github.com/subbareddy248/cross_situational_learning.

This chapter has been finalized based on an initial report presented at the Splu-RoboNLP workshop at ACL in July 2021, which was later extended into a journal article. It is currently undergoing minor revisions for publication in the journal Nature Scientific Reports [Oota et al., 2022a].

## INTRODUCTION

Experimental and modeling studies of language acquisition [Dominey and Boucher, 2005, Chen and Mooney, 2008, Tomasello, 2009, Thomason et al., 2018, Vanzo et al., 2020, Roembke et al., 2023] try to understand how infants can learn language by observing their environments and interacting with others. Before one year of age, children can segment words from speech based on statistical learning mechanisms [Saffran et al., 1996, Yu and Smith, 2007]. Moreover, what children observe when hearing "the blue glass is on the left" and how they map these sounds with visual concepts is crucial, particularly when distinguishing between a blue glass and a green one. To navigate this, children have to learn from several presentations of the same word in various contexts, a phenomenon often referred to as cross-situational learning (CSL) [Taniguchi et al., 2017, Juven and Hinaut, 2020, Warren et al., 2020, Dinh and Hinaut, 2020, Variengien and Hinaut, 2020, Roembke et al., 2023]. For instance, in Fig. 7.1(a), the CSL paradigm is illustrated. In this scenario, the early language learners (infants) are presented with multiple referents and multiple words in one naming moment, they are unable to decide which word maps onto which object.
Traditional approaches to language grounding mainly focus on mapping natural language commands and task representations, essentially sequences of primitive robot actions [Chen and Mooney, 2011, Matuszek et al., 2013, Tellex et al., 2011]. In recent years, a large amount of research has been focused on grounded language learning, exploring how robots can learn to correlate natural language with the physical world, recognizing objects by their names and attributes, using either multimodal or natural language data [Thomason et al., 2016, Beinborn et al., 2018]. Further, several studies have performed computational experiments on CSL by tracking the co-occurrence between word forms and referents (objects) to model how infants could do it [Romberg and Yu, 2013]. In this paradigm, initially, the word-referent mappings appear entirely random. With repeated trials, the robot must learn to identify the appropriate object properties and representations of the meaning of individual words. However, existing robotic frameworks [Taniguchi et al., 2017, Roesler et al., 2018] do not adequately model how children learn to understand directly from full sentences through cross-situational learning without providing specific cues including: (i) social cues from speakers, such as gaze direction, head orientation, body movements, gestures, speech intonation, and facial expressions [Yu and Ballard, 2007, MacDonald et al., 2017]; (ii) visual cues from objects held by the teacher, including gestures like pointing at, hovering over, or displaying an object, as well as the movement of the object [Roy, 2002, Krenn et al., 2017]; and (iii) auditory cues, such as requests for action and the naming of objects [Räsänen and Rasilo, 2015, Krenn et al., 2017, Escudero et al., 2023], etc,.

Recently, the Transformer-based pretrained language model, specifically bidirectional encoder representation transformer (BERT) [Devlin et al., 2019], has brought large improvements in the field of natural language processing (NLP) on a wide variety of tasks, including machine translation [Devlin et al., 2019], sentence representation [Devlin et al., 2019], and semantic role labeling [Shi and Lin, 2019, Zhang et al., 2020b]. These models are potentially appealing as cognitive models because they can learn from raw linguistic stimuli, something previous cognitive models have not addressed. Furthermore, it is unclear how these Transformer-based representations would compare as accounts of models able to

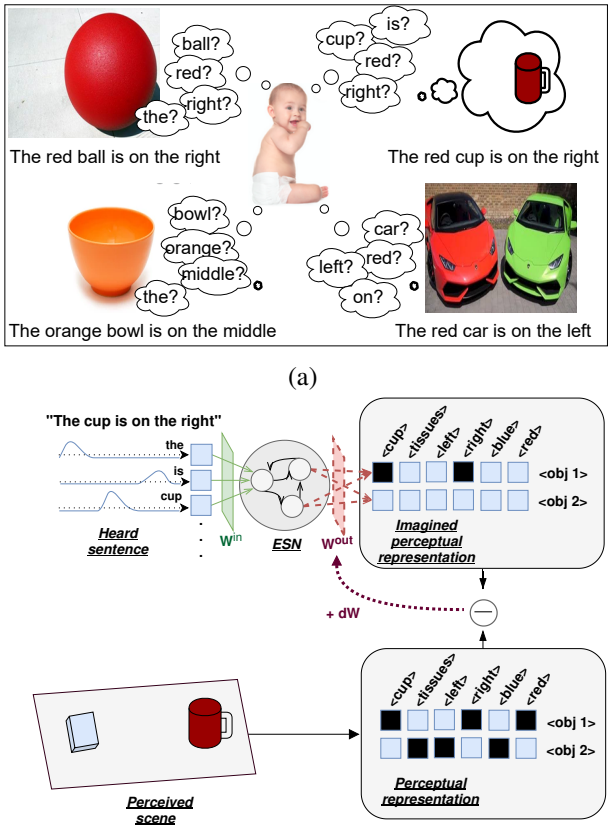



Figure 7.1: (a) At start, children don't know how to map words from an utterance to features of a scene they observe. This concerns content words (e.g. "car", "red") that should be mapped to observed features, but also functions words (e.g. "on", "the") that should not be mapped. They have to learn this mapping while perceiving sequence of words; words are rarely spoken in isolation. (b) The Cross-Situational Learning (CSL) task (on simple Juven's dataset) tested with Recurrent Neural Networks (ESNs and LSTMs). The model has to reconstruct an imagined scene from the sentence given word by word. This image is sourced from Juven and Hinaut [2020].

capture a wide range of key phenomena in cross-situational learning. Given that pretrained language models are trained in the self-supervised setting, and these language models exhibit slower learning of words in longer utterances in a similar way as children acquire language [Chang and Bergen, 2022], it poses a challenge for researchers to investigate the use of transformer models in robotics. Inspired by their success of pretrained Transformer models (BERT [Devlin et al., 2019], T5 [Raffel et al., 2020], and GPT-2 [Radford et al., 2019]) when applied to robotics and reinforcement learning tasks [Hill et al., 2020, Marzoev et al., 2020], we leverage BERT model to encode the sentences.

Usually, robotic implementations or models emphasizing grounding tend to focus on single words "Run!, Stop!" or simple sentences like "Put yellow cube right", which are

more like a predefined sequence of words than natural sentence processing [Taniguchi et al., 2017]. Some recent works utilizing Transformers are able to parse more complex sentences. However, such sophisticated deep learning architectures need a huge dataset (or pretrained models) to learn such relations. Moreover, they lack insights into how children's brains perform such tasks because they are not biologically plausible in the training schema nor in how biological neural networks process utterances, usually taking input sentences as a whole. At the same time, humans have to process word by word the incoming flow of speech. Furthermore, in the study of Pedrelli and Hinaut [2021], a model using a hierarchy of reservoirs could convert raw speech into semantic role labels by recognizing intermediate representations like phonemes and words. This raises the question of how reservoirs handle input sentences (word by word) in a manner analogous to human sentence processing.

The motivation to perform the CSL task is interesting, as it involves employing simple neural architectures to generalize efficiently with few noisy trials, mimicking the learning conditions experienced by children. In this task, some words may appear only a few times in the training set. Similarly to children who do not have an oracle that gives the correct labels for each word, the models do not have access to true teacher output but to a noisy version of it, based on the concepts a child could extract from visual information. This implies that visual scenes often contain more objects and features than what a given sentence will explicitly describe. The core principle of CSL—deriving associations between symbols and their referents by observing their co-occurrences—acts as a stand-in for the intricate, real-world process of language learning in infants. Nonetheless, our methodology incorporates three language grounding datasets, specially curated for this study and varying in complexity to emulate learning environments ranging from simpler to more intricate, similar to those encountered by robots. Each dataset contains 1000 training examples, meticulously chosen to test and assess the capability of Echo State Networks (ESNs) [Jaeger, 2001, 2002, 2007] and Long-Short Term Memory networks (LSTMs) [Hochreiter and Schmidhuber, 1997] to learn from limited and subtle data. While our study does not employ datasets directly sourced from infant learning experiments, we have selected a psycholinguistic task known as CSL due to its conceptual resemblance to the manner in which infants acquire language. Originating from developmental psychology, this task seeks to emulate how infants learn to link words with their meanings within contexts that are naturally ambiguous. Overall, our study aims to establish foundational insights through initial experiments in the area of language grounding, prior to advancing towards practical applications involving robots, which necessitate a considerably longer experimental duration.

Importantly, we do not want to focus on engineered neural architectures for biologically plausible purposes. Instead, we are keen on exploring how relatively simple recurrent neural networks could generalize in such conditions while using incremental learning. In particular, one of the models we use, Echo State Networks (ESN) and, more generally, the Reservoir Computing paradigm, have already been used in several neuroscience models [Buonomano and Merzenich, 1995, Maass et al., 2002, Hinaut and Dominey, 2013] and are often referred to as a plausible computational principle for electrophysiological results [Machens et al., 2010, Rigotti et al., 2013, Enel et al., 2016]. Moreover, ESNs are more biologically plausible than LSTMs because they do not need to rely on back-propagation through time, which involves virtualizing time for several time steps, which is not biolog-

ically relevant. Also, ESNs can learn incrementally by seeing each utterance only once (which is thus closer to what children experience), contrary to LSTMs, which needs to process the data for several epochs. The CSL procedure for the ESN architecture is shown in Fig. 7.1(b).

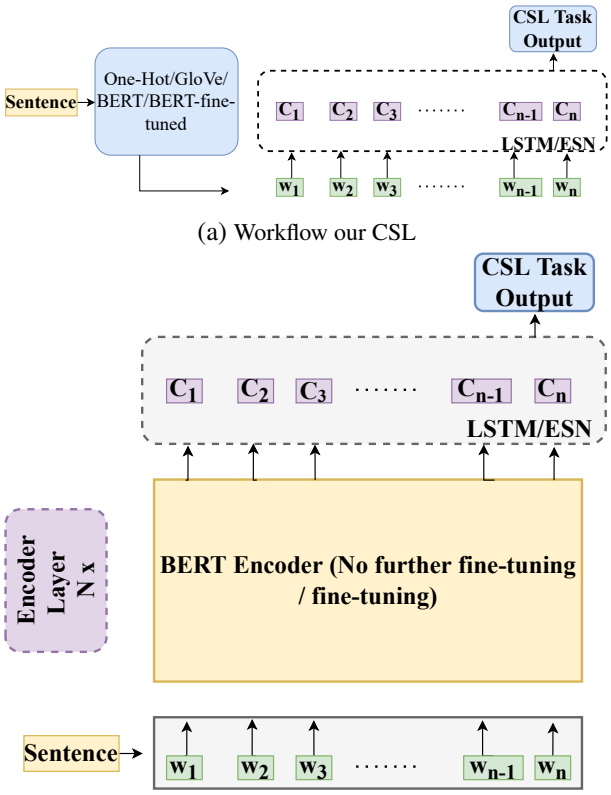
This study also investigates how well the one-hot, GloVe [Pennington et al., 2014], and transformer-based representations perform in the cross-situational learning task, using the biologically plausible ESN model [Jaeger, 2001] and non-plausible[1] LSTMs [Gers et al., 1999]. Figs. 7.2 (a) and 7.2 (b) depict the workflow of our CSL task. The proposed framework, incorporating input featurization, dynamic memory, and learning modules, offers a flexible and biologically plausible architecture for investigating CSL tasks on diverse datasets. Our experimental results demonstrate that fine-tuned BERT representations are more efficient and better at capturing the complex relations between words than other word representations. For sentences with a smaller number of objects, LSTMs outperform the ESNs across two datasets. On the other hand, for the increasingly larger number of objects, ESNs display better performance than both LSTM and RandLSTM. These results demonstrate the biological plausibility of ESNs, as the internal weights of the underlying reservoir network remain unchanged during learning. In contrast, LSTMs require training of both internal weights and reservoir-to-output connections." Finally, we try to interpret the inner working details of two models and plot the evolution of the output activation during the processing of a sentence.

## 7.1 Related Work

### Cross situational learning models

Human infants learn word meanings from several presentations of the same word in different contexts, including uncertainty and ambiguity in the language environment; this is often referred to as cross-situational learning [Yu and Smith, 2007, Taniguchi et al., 2017, Juven and Hinaut, 2020, Warren et al., 2020]. Traditional approaches to cross-situational learning use three types of models: computational [Kachergis et al., 2012, McMurray et al., 2012], statistical [Trueswell et al., 2013, Stevens et al., 2017], and Bayesian models [Frank et al., 2009, Yurovsky and Frank, 2015], aiming to examine the plausibility of language models for language learning. In particular, a body of research has demonstrated that both adults and infants can effectively exploit cross-situational learning information when learning a small number of words, using both naturalistic and more controlled stimuli [Akhtar and Montague, 1999, Yu and Smith, 2007, Medina et al., 2011, Trueswell et al., 2013]. Additionally, several studies have performed computational experiments on cross-situational learning by tracking the co-occurrence between word forms and referents (objects) to model how infants could do it [Smith and Yu, 2008]. However, existing robotic frameworks only model how children learn to understand directly from full sentences through cross-situational learning,

[1] LSTMs are not biologically plausible because they use an engineered mechanism to perform back-propagation on time-unfolded representations.




(a) Workflow our CSL

(b) Workflow of our CSL with BERT model

Figure 7.2: (a) The sentences are passed as input to One-Hot/GloVe/BERT/fine-tuned BERT for extracting the word embeddings. (b) The sentence are passed as input to BERT encoder for extracting the token embeddings from the last output layer in two setups: (i) no further fine-tuning of encoder weights, (ii) fine-tuning of encoder weights. These token embeddings are passed as input to the LSTM/ESN models for the final prediction of CSL task outputs. It's important to note that the language grounding datasets employed in our study are composed of textual representations designed to simulate the process of language acquisition and grounding in a controlled and measurable way.

providing specific cues such as visual cues [Roy, 2002], social cues [MacDonald et al., 2017], and auditory cues [Räsänen and Rasilo, 2015, Escudero et al., 2023], etc,.

### Bio-inspired Language Models

Deep neural architectures such as ESN and LSTM are shown to be successful in handling sequential tasks. Recently, ESNs have been successfully applied to understand how infants learn the meaning of words in a fuzzy context. ESNs need to make associations between symbols and referents [Juven and Hinaut, 2020], building upon previous studies using supervision to model human sentence parsing [Dominey et al., 2006, Hinaut and Dominey, 2013] and multilingual processing [Hinaut et al., 2015, Hinaut and Twiefel, 2019], adapt-

ing it for human-robot interactions [Hinaut et al., 2014, Twiefel et al., 2016, Hinaut, 2018]. Similarly, several authors have explored language acquisition tasks with LSTMs [Zhong et al., 2017] and GRUs [Ororbia et al., 2020], engaging in learning robotic multi-modal tasks when provided with sentences. To this extent, ESNs and LSTMs, coupled with the cross-situational task, offer a more plausible learning perspective from a human brain than a purely supervised task. Moreover, this type of task is also interesting for practical applications where exact target outputs are not always available.

#### Grounded language models and human-robot interaction

Human-robot interaction using natural language often requires language to be grounded to the agent's perceptions of the physical world and its interactions with others. Specifically, cross-situational learning in human-robot interaction involves statistical mechanisms whereby robots learn to associate words with sensory experiences, drawing heavily on repeated exposure to varied contexts [Yu and Smith, 2007]. Prior studies of language acquisition involves the teaching strategies employed by humans, such as providing feedback and guidance, play a crucial role in optimizing the robot's learning trajectory [Thomaz et al., 2006]. However, the challenge of disambiguating words that appear in multiple contexts remains a significant hurdle, requiring sophisticated algorithms that allow robots to interpret commands correctly based on contextual information [Kollar et al., 2014]. More recently, joint attention mechanisms have been introduced in Matuszek et al. [2013], which allow robots to focus on the same objects or actions as humans during teaching phases, significantly improving the efficiency of language acquisition. Overall, this body of research not only advances our understanding of how robots can effectively learn language but also underscores the complex interplay of human input and algorithmic processing in creating more adaptable and intuitive robotic systems.

## 7.2 Methodology

In this section, we propose to employ a CSL task using two sequence-based models, including ESN (i.e. Reservoir Computing) and LSTM, to build the grounded language acquisition models. Here, we recall the definitions of reservoir computing and random features in ESN, and LSTM, and introduce the details of the model architecture.

#### Echo State Networks (ESN)

Reservoir Computing [Lukoševičius and Jaeger, 2009] is an effective paradigm as Recurrent Neural Network (RNNs) receives the sequential input $x_t \in \mathbb{R}^d$ and producing the output $y_t$, where internal weights are fixed randomly and only the output layer (called the "read-out") is trained [Jaeger, 2001].Let N be the number of neurons in the reservoir, the reservoir state $r_t$ is updated by using the following recurrent equation:

$$r_t \leftarrow (1-\alpha)r_{t-1} + \alpha \tanh(\mathbf{W}_{rec} r_{t-1} + \mathbf{W}_{in} x_t) \tag{7.1}$$

where $\mathbf{W}_{rec} \in \mathbb{R}^{NXN}$ and $\mathbf{W}_{in} \in \mathbb{R}^{NXd}$ are respectively the reservoir and input weight matrices, and the parameter $\alpha$ denotes the leak rate.

**Offline Learning vs. Online Learning** Offline learning refers to a training process in which the model is trained on a complete dataset at once. The entire training dataset is available from the start, and the training involves adjusting the network's output weights to minimize error across all the training data. The internal state and recurrent weights of the ESN typically remain unchanged; only the readout weights are optimized during the training phase. This approach is ideal when you have a complete dataset and the task doesn't require the model to adapt to new data over time. Online Learning with ESNs involves updating the model's weights incrementally as new data becomes available, without the need to retrain the model from scratch with the entire dataset [Hinaut and Dominey, 2012]. This method is particularly useful for tasks where data arrives in a stream or when the system needs to adapt to changes in the data distribution over time.

During the testing phase, particularly with Echo State Networks (ESN) and the distinctions between offline and online learning, the parameters of the model—primarily those in the readout layer (given that the reservoir of an ESN typically remains unchanged)—are "frozen." This means that the parameters are not updated any further once the training phase is completed. Freezing the parameters during testing facilitates a fair evaluation of the model's performance.

### ESN Offline Learning

To refine the control of the reservoir dynamics, we add a constant bias to the reservoir state $s_t \in \mathbb{R}^N$ and then multiply this reservoir state $s_t$ by the output matrix $\mathbf{W}_{out}$ to get the output vector $y_t$ as described in the Equation 7.2. The output predicted by the network $y_t$ closer to the teacher vector is obtained by optimizing the output weight matrix $\mathbf{W}_{out}$ after a final layer.

$$s_t = \begin{pmatrix} 1 \\ r_t \end{pmatrix} \qquad y_t = \mathbf{W}_{out} s_t \tag{7.2}$$

Since only the output weights $\mathbf{W}_{out}$ are trained, the optimization problem boils down to simple linear regression, called an offline learning method.

### ESN with FORCE/Online Learning

In the context of online learning models such as ESN with Final Learning (FL) or Continual Learning (CL), the model's parameters, particularly those of the readout layer, are updated after each training example. Final Learning (FL) involves applying the FORCE algorithm to the reservoir's state after processing the entire sentence [Sussillo and Abbott, 2009]. This means that the model waits until the last word of the sentence before updating its parameters. Continual Learning (CL), on the other hand, employs the FORCE algorithm after each word within a sentence. This method requires the ESN to update its parameters and potentially alter its output predictions as each word is processed. This distinction highlights the trade-offs between the two approaches: FL focuses on accuracy by utilizing complete sentence contexts for updates, while CL emphasizes adaptability and interpretability, accepting potentially lower performance for insights into the model's processing of sequential data.

To update the output weights $\mathbf{W}_{out}$ for each learning example, the online FORCE learning algorithm [Sussillo and Abbott, 2009] that is a more biologically plausible model to train the network than usual ESN offline learning. This method does not unfold time while training the network like back-propagation through time. The matrix P "acts as a set of learning rates for the RLS (Recursive Least Squares) algorithm" [Sussillo and Abbott, 2009]. Let $e_t$ be the error between the prediction of the network and the ground truth at time t, and the output weights are updated as follows:

$$\mathbf{W}_{out}(t) = \mathbf{W}_{out}(t-1) - e_t \mathbf{P}(t) r_t \tag{7.3}$$

$$\mathbf{P}(0) = \frac{I}{\in} \tag{7.4}$$

$$\mathbf{P}(t) = \mathbf{P}(t-1) - \frac{\mathbf{P}(t-1) r_t r_t^T \mathbf{P}(t-1)}{1 + r_t^T \mathbf{P}(t-1) r_t} \tag{7.5}$$

where I is an identity matrix and $\epsilon$ is a regularisation term.

**ESN with Final Learning (ESN FL)** For the final learning method, the FORCE algorithm is applied to the reservoir state after the last word of the sentence.

**ESN with Continual Learning (ESN CL)** Unlike ESN FL, the reservoir states are updated after each word of a sentence using the FORCE learning method; an equivalent method with offline learning was used in Hinaut and Dominey [2013].

### LSTM

An LSTM [Gers et al., 1999] network with sequential time steps that computes an output $y_t$ as a function of the input vector $x_t$, and weights of hidden state obtained using three gates (forget gate, input gate, output gate). The weights of LSTMs are learned using the error back-propagation through time, BPTT, an algorithm to maximize the log-likelihood of the training data given the parameters. In order to compare the performance of ESNs with LSTMs, we employ unidirectional LSTMs in our CSL tasks.

### RandLSTM

In RandLSTM model [Bai et al., 2018], the LSTM weight matrices and their corresponding biases are initialized uniformly at random and kept frozen (i.e both Input and LSTM connection weights are random) [Wieting and Kiela, 2018]. Hence, the output layer parameters are only trainable and the remaining parameters are frozen in RandLSTM model [Bai et al., 2018].

## 7.2.1 Availability of Data and Materials

Here, we describe the three diverse datasets: Juven's (simple sentences) [Juven and Hinaut, 2020], GoLD (consists of simple to very complex sentences) [Jenkins et al., 2020], and *Knowledge Technology Train Robots (KTTR)* (sentences that describe complex robot actions) [Twiefel, 2020] – which we will call *Robot Data* for simplicity in the paper. We reused these three publicly available datasets for this work, we did not collect any new dataset.

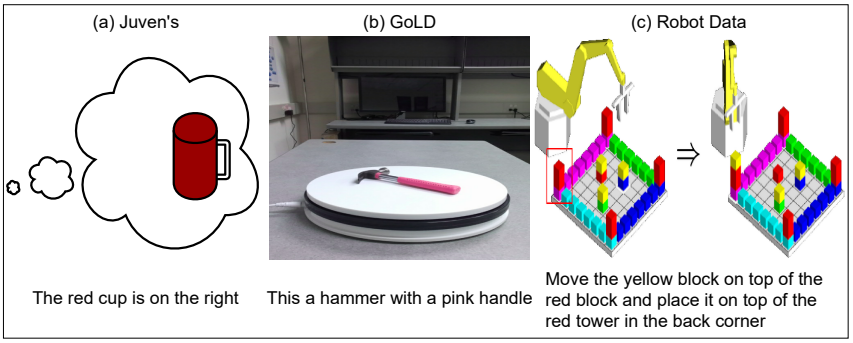



Figure 7.3: Example sentences with concepts from three datasets: Juven's, GoLD, and Robot data. Here, the image from the GoLD dataset is sourced from Jenkins et al. [2020], and the image from the Robot dataset is sourced from Twiefel [2020].

- Juven's dataset can be downloaded from this link [2]. Please read their terms of use for more details.
- GoLD dataset can be downloaded from this link [3]. Please read their terms of use for more details.
- Robot dataset can be downloaded from this link [4]. Please read their terms of use for more details.

Fig. 7.3 showcases the example of sentences corresponding to each dataset, and we present the detailed statistics of each dataset in Table 7.3 (refer in supplementary).
**Juven's CSL:** Juven's CSL dataset [Juven and Hinaut, 2020] comprises approximately 70,000 sentences, of which we randomly sampled 1000 training sentences and 1000 testing sentences, where each sentence describes one or two objects. The sampled dataset has 700 sentences with two objects and 300 with one object in training and testing. We validated our models on Juven's dataset by varying the number of object classes from 4 to 50, three actions, and four colors. These objects were chosen to reflect and provide data for three different domains: home, kitchen, and tools, in which the model learns to ground a complex sentence, describing a scene involving different objects into a perceptual representation space.
**GoLD:** Grounded language dataset (GoLD) [Jenkins et al., 2020] is a collection of visual, speech, and language data in five different domains: food, home, medical, office, and tools. There are 8250 textual descriptions consisting of 47 object classes spread across five different groups, seven actions, and eight colors.
**Robot Grounding Dataset:** Robot dataset [Twiefel, 2020] is a collection of visual, speech, and language data that focuses on contextual semantic parsing of robotic spatial commands. There are 2500 textual descriptions in the training set, of which we randomly sampled 1000

[2] https://github.com/aJuvenn/JuvenHinaut2020_IJCNN
[3] https://github.com/iral-lab/gold
[4] https://alt.qcri.org/semeval2014/task6/index.php?id=data-and-tools

sentences for training. The test consists of 909 textual descriptions (different from 2500 sentences from the training set). Overall, the dataset consists of 11 object classes, three actions, eight colors, nine directions, and nine positions. Unlike Juven's data, both GoLD and Robot datasets consist of simple to very complex sentences.

### CSL Task: Input and Output

In our experiments, as illustrated in Figure 7.2, we exclusively utilized text-based data for our analysis. The language grounding datasets employed in our study are composed of textual representations designed to simulate the process of language acquisition and grounding in a controlled and measurable way. This approach allowed us to focus on the computational models' ability to learn associations between words and their meanings within various contexts, a core aspect of the Cross-Situational Learning (CSL) task that our study aims to explore.

For each sequence-based model, we give the input and output as follows: (i) The input is a sequence of words (i.e. a sentence) with one-hot, GloVe, or word representation from pretrained/fine-tuned BERT. (ii) The target output is a constant vector corresponding to concepts units (i.e. objects, colors, positions). (iii) Since the CSL task is defined in noisy supervision, the target may include additional concepts outputs that are not in the input sentence. Fig. 7.9 displays the target vectors corresponding to input sentences for Juven's and GoLD datasets. The semantic output structure and the binary encoding of semantic structure for the Robot dataset are shown in Figs. 7.8 (a) and 7.8 (b), respectively.

| Dataset | Sentences with Single/Two Objects | Concepts |
|---|---|---|
| Juven's | A bowl is on the middle and a glass on the right is green | Obj1: bowl (middle)<br>Obj2: glass (right, green) |
| GoLD | This is a hammer with a pink handle | Obj1: hammer (pink)<br>Obj2: NA |
| Robot | Move the yellow block on top of the red block and place it on top of the red tower in the back corner | move (cube1, above, cube2),<br>drop (cube1, above, group3),<br>is (group3, above, corner 4)<br>cube1 (yellow)<br>cube2 (red)<br>group3 (red)<br>corner4 (right) |

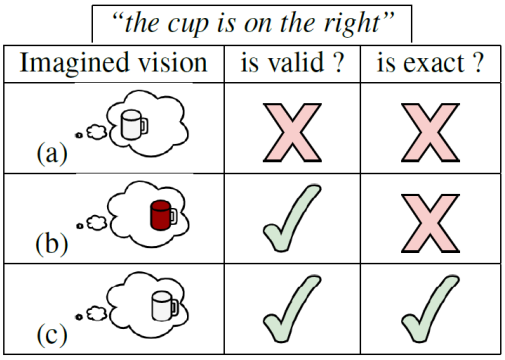



Figure 7.4: (left) Example sentences with concepts from three datasets: Juven's, GoLD, and Robot data. (right) Evaluations of different imagined scenes. (a) It is not valid or exact because the cup is not on the right. (b) It is not exact because the sentence does not mention the cup color. (c) It is both valid and exact because the imagined scene is same as a textual description. This Figure (right) is source from Juven and Hinaut [2020].

### Evaluation Methodology

During model training, we use cross-entropy as the loss between prediction and ground truth. To evaluate the performance of two models on prediction of test sentences, we use

the two error metrics: Valid and Exact [Juven and Hinaut, 2020] for Juven's and GoLD, as shown in Fig. 7.4 (right). In the Robot dataset, we use only the exact error because the model predictions for a sentence are classified correctly (actions, relations, objects, and attributes), and the output is considered correct. Partly incorrect outputs are considered incorrect, making it a strict metric [Twiefel, 2020]. Since the visual representation can contain more information about the scene than what is described in the sentence in a cross-situational learning task, we cannot simply quantify the performances of a model with the distance to the desired teacher vector. The percentage of sentences from the testing set considered invalid or not exact is then used as a quantitative error measurement. The error metrics are defined as follows:

$$\text{Valid Error} = 1 - \frac{\#\text{Valid Representations}}{\#\text{Instances}} \tag{7.6}$$

$$\text{Exact Error} = 1 - \frac{\#\text{Exact Representations}}{\#\text{Instances}} \tag{7.7}$$

where #Instances denote the number of test instances, Valid Representation=1 if every concept mentioned in the sentence is present, else 0. Similarly, Exact Representation=1 if the representation contains all the sentence information and nothing more, else 0.

To enable a fair comparison between the two models (ESN and LSTM), we set the threshold is fixed to $1.3/K_c$ throughout the paper. In the CSL task settings, $K_c$ represents the number of possible values for each concept c. Each concept c (e.g., color, position, object category) in the model can take on a specific number of values. For instance, if the concept is "position," the possible values might be "right," "middle," and "left," making $K_c$=3 for this concept. Similarly, for the concept "color", the possible values might be "red", "blue", "green" and "orange", making $K_c$=4 for this concept. Essentially, $K_c$ quantifies the diversity or the range of values that a particular concept c can assume in the model. Table 7.3 reports the number of concepts and possible values associated with each concept for 3 datasets.

We selected the threshold factor (1.3) based on the study by Variengien and Hinaut [2020], which utilized the Juven's dataset. The choice of threshold factor (1.3) impacts the performance of the models. As the threshold factor increases, the exact error decreases while the valid error increases. Essentially, the exact error functions similarly to a false positive rate, and the valid error corresponds to a false negative rate. With a threshold of 1.3, we achieve the lowest error rates for both LSTM and ESN models without favoring one over the other. Therefore, we applied the threshold factor of 1.3 across all datasets. In summary, $1.3/K_c$ serves as a customized threshold to assess the significance of a model's prediction for each concept c, where $K_c$ is the diversity of the concept, and 1.3 amplifies the minimum probability required to accept a prediction as significant.

### Cross-Validation

In the cross-validation setup, the training set was indeed fixed across all models for the purpose of ensuring a fair comparison. By randomly sampling the training instances in each run and then running each model (ESN-Offline, ESN-Online FL, ESN-Online CL, RandLSTM, and LSTM) through this process, repeated across five iterations. This we ensured that

each model was trained and evaluated on an identical dataset in each run. Averaging the results across these five runs for each model helps to mitigate the effects of simple to complex sentences on model performance, providing a more reliable measure of each model's capability to learn from the dataset. This approach aligns with best practices in machine learning research, where controlling for variables such as the training data is crucial for accurately assessing and comparing the performance of different models. In our model training, we use a small data set (1000 sentences) to see what the model can learn with few-shot learning; some words may appear only a few times in the training set.

## 7.3 Experimental Setup

We evaluated our cross-situational learning task on three datasets across twenty different settings: 5 architectures x 4 feature representations. The five architectures are: (i) ESN-Offline, (ii) ESN-Online FL, (iii) ESN-Online CL, (iv) RandLSTM, and (v) LSTM.

### Feature Representations

We use the feature representations such as one-hot encoding, GloVe, pretrained BERT, and fine-tuned BERT as input for the models ESN and LSTM.
**One-Hot Encoding:** In one-hot encoding, each word is represented as a binary vector that is all zero values except the index of the word from the unique vocabulary, which is marked with a 1. Thus, the dimension of the input vector will be equal to the vocabulary size (see Table 7.10 in sup. mat.).
**GloVe:** We use the existing pretrained word embeddings, GloVe based word vectors (each word is a 300-dimension vector) [Pennington et al., 2014] to perform the CSL task.
**BERT:** Pretrained BERT model [Devlin et al., 2019] provides word contextual information by looking at previous and next words, which is one of the main limitations in earlier language models. For every sentence, BERT yields $1 \times$ #tokens $\times$ 768 dimensions, where #tokens denote the number of tokens (i.e. each token will be represented as 768 vectors).
**Fine-tuned BERT:** Here, we use the BERT-base-cased model and fine-tuned on the last layer of BERT model for each dataset. Like BERT, we obtained $1 \times$ #tokens $\times$ 768 dimensions for every sentence from fine-tuned BERT.

### Model Training

**ESN Training:** We use the ReservoirPy library [Hinaut and Trouvain, 2021][5] to build the ESN model, where the model is trained on 1000 sentences and tested on 1000 sentences. When tuning hyperparameters for an ESN using the ReservoirPy library, the approach involves a systematic exploration of the hyperparameter space to identify the combination that yields the best performance. This process is crucial because the choice of hyperparameters can significantly affect the model's ability to learn and generalize from the data. We chose four hyper-parameters to explore: spectral radius (SR), leak rate (LR), sparsity,

[5] https://github.com/reservoirpy/reservoirpy

and ridge regularization parameter. We also chose to fix at least one of the more important hyper-parameter, to reduce the complexity of the search: input scaling (IS) will be kept constant and equal to 1 during this first step. Unlike grid search, which exhaustively tests all possible combinations of hyperparameters, random search samples a subset of combinations from the defined hyperparameter space. In our random search, we performed 100 evaluations is sufficient enough to to explore the space and identify a well-performing set of hyperparameters. Figs. 7.15, 7.16 and 7.17 (please refer the supplementary) display the cross-entropy loss along with both exact and valid error performance with exploration of four hyper-parameters. Similarly, we report the hyperopt plots for GoLD and Robotic datasets in the supplementary (please refer the Figs. 7.18, 7.19, 7.20, 7.21 and 7.22). We obtain the following parameters by performing the random hyper-parameter using hyperopt[6] for each dataset as follows: For **Juven's dataset:** {Spectral Radius = 0.025, Leak Rate = 0.0097, Sparsity (on Reservoir Weight Matrix - $\mathbf{W}_{rec}$) = 0.5, Regularization coefficient = 1.3e$^{-10}$, Input Scaling = 1.0}. For **Robot dataset:** {Spectral Radius = 0.839, Leak Rate = 0.0735, Sparsity (on Reservoir Weight Matrix - $\mathbf{W}_{rec}$) = 0.5, Regularization coefficient = 3.91e$^{-5}$, Input Scaling = 1.0}. For **GoLD dataset:** {Spectral Radius = 2.29, Leak Rate = 0.003, Sparsity (on Reservoir Weight Matrix - $\mathbf{W}_{rec}$) = 0.2, Regularization coefficient = 0.01, Input Scaling = 1.0}.

**LSTM Training:** Our parameter selection for the LSTM model is guided by the use of Keras Tuner for hyperparameter tuning [O'Malley et al., 2019]. Through Keras Tuner, we configured the Dropout_rate to vary between a minimum of 0 and a maximum of 0.5, with increments of 0.1. We also explored learning_rate values within the range of [1e-2, 1e-3, 1e-4], opting for the mean squared error as the loss function and Adam as the optimizer. Employing RandomSearch from Keras Tuner enabled us to methodically test a broad spectrum of hyperparameters, including the number of LSTM units from 10 to 160, learning rate, and batch sizes of 8, 16, and 32. By defining this search space, RandomSearch could randomly evaluate numerous configurations, optimizing for the best performance based on predefined criteria such as minimizing loss. In summary, for the ESN model, hyperparameter tuning is conducted solely on the reduced corpora, and the derived parameters were subsequently applied to the complex corpora. To ensure a fair comparison of model performance, we adopted a similar approach for the LSTM model by performing hyperparameter tuning on the reduced corpora and applying the identified parameters to the complex corpora.

Following the determination of the hyperparameters, we experimented with one layer of LSTM to capture the meaning of the concepts. The model is implemented in Keras with TensorFlow backend [Abadi et al., 2016] with meas squared error as loss, Adam optimizer [Kingma, 2014], the number epochs set to 70, the batch size is of 8, and tried LSTM with different hidden units (20, 40, 80). Since the number of trainable parameters in 20-unit LSTM is equivalent to the ESN model with 1000 reservoir units, we use these two settings for baseline comparison. We used the early-stopping method to stop model training when the loss started to plateau with patience of 5.

Our decision to set the early stopping patience at 5 is driven by a balance between model performance and computational efficiency. LSTM models, being a form of recurrent neural

[6] http://hyperopt.github.io/hyperopt/

network (RNN), are particularly sensitive to the risk of overfitting due to their capacity to model complex, long-term dependencies in sequential data. A lower patience value, like 5, helps in avoiding overfitting by stopping the training when the model's performance on the validation set does not improve for a consecutive number of epochs. This approach also conserves computational resources by preventing unnecessary training iterations. Furthermore, given our dataset contains only 1,000 sentences, opting for an early stopping patience of 5 is a wise choice. It allows us to maintain an effective learning process while preventing overfitting, a crucial consideration given the limitations in both computational resources and dataset size.

Choosing a higher patience value, such as 10 or 20, could potentially allow for more subtle improvements in model performance over a longer period. However, it also increases the risk of overfitting and requires more computational time and resources. By evaluating the model's learning dynamics and considering the trade-off between performance gains and computational cost, we determined that a patience of 5 yield an optimal balance for our specific scenario.

## 7.4 Results

### 7.4.1 CSL task performance of Sequence-based models

In this section, we report our two sequence-based model results on the CSL task using three datasets viz. Juven's, GoLD, and Robot. We used the four different word representations such as one-hot encoding, GloVe, BERT (bert-base-case)[7], and fine-tuned BERT to extract the features for every sentence, and the error metrics are computed from the two sequence-based models. To compare the effectiveness of the models with an approximately equal number of parameters (ESN with 1000 units, RandLSTM with 1000 units and a 20-unit LSTM) using different token representations as input feature vectors, we report the Valid and Exact errors for the Juven's and GoLD, and Exact error for Robot datasets, respectively, described in Tables 7.1 (a) and 7.1 (b). To verify statistical difference between pairs of these sequence-based models, we have employed the two-sample paired t-test to rigorously assess the differences in performance metrics (*'Valid'* and *'Exact'* errors) observed between our ESN models and LSTM models across multiple runs within our cross-validation framework. This statistical test is particularly suited for our analysis as it compares the means of two independent samples, aligning perfectly with our objective to evaluate performance distinctions across distinct model architectures.

**Reduced-size Corpora Results:** In Table 7.1(a), we evaluate the performances on a smaller number of objects datasets, we chose 4-objects for Juven's and 10 objects for GoLD data. From Table 7.1 (a), we found that both models are able to learn the CSL task with low error successfully and outperform the one-hot, GloVe, and pretrained BERT results; we make the following observations. (i) We observe that the LSTM outperforms the ESN on both *Valid* and *Exact* errors on Juven's and GoLD datasets. (ii) On the other hand, ESN displays better performance than RandLSTM while considering the same number of neurons in both

[7] https://huggingface.co/bert-base-cased

| | Juven's CSL Data | | GoLD Data | |
|---|---|---|---|---|
| **Model** | **Valid** | **Exact** | **Valid** | **Exact** |
| ESN-offline + One-Hot | 33.10±0.16 | 43.90±0.21 | 17.46±0.24 | 25.39±0.58 |
| ESN-offline + GloVe | 16.70±0.11 | 25.00±0.13 | 25.88±2.70 | 30.19±4.40 |
| ESN-offline + fine-tuned BERT | 2.20±0.02 | 10.90±0.05 | 21.43±0.14 | 25.51±0.47 |
| ESN-offline + BERT | 2.30±0.02 | 12.20±0.07 | 24.47±0.21 | 43.45±0.40 |
| ESN-online FL + One-Hot | 0.28±0.01 | 05.64±0.03 | 12.29±0.24 | 20.89±0.28 |
| ESN-online FL + GloVe | 0.10±0.01 | 12.20±0.09 | 12.69±0.46 | 25.15±0.60 |
| ESN-online FL + fine-tuned BERT | **0.00±0.00** | 06.28±0.01 | 12.11±0.13 | 22.22±0.21 |
| ESN-online FL + BERT | 0.20±0.00 | 07.72±0.07 | 20.62±0.16 | 42.22±0.21 |
| ESN-online CL + One-Hot | 2.32±0.01 | 12.10±0.09 | 14.82±1.96 | 26.58±2.37 |
| ESN-online CL + GloVe | 7.80±0.08 | 25.50±0.14 | 15.38±2.40 | 28.93±3.40 |
| ESN-online CL + fine-tuned BERT | 2.41±0.01 | 13.70±0.10 | 13.83±0.76 | 27.79±0.88 |
| ESN-online CL + BERT | 2.78±0.01 | 14.60±0.11 | 20.13±0.78 | 38.08±0.87 |
| RandLSTM + One-Hot | 7.30±0.10 | 10.00±0.14 | 21.91±0.43 | 24.64±0.34 |
| RandLSTM + GloVe | 23.80±0.78 | 48.70±1.02 | 49.93±1.01 | 51.66±1.40 |
| RandLSTM + fine-tuned BERT | 4.30±0.04 | 7.40±0.09 | 17.19±0.23 | 18.33±0.21 |
| RandLSTM + BERT | 8.00±0.06 | 35.00±0.89 | 20.23±0.24 | 23.37±0.28 |
| LSTM + One-Hot | 0.10±0.00 | 03.50±0.01 | 16.40±0.13 | 22.85±0.18 |
| LSTM + GloVe | 2.90±0.01 | 17.50±0.10 | 41.11±0.18 | 48.88±0.09 |
| LSTM + fine-tuned BERT | 0.20±0.01 | **01.30±0.02** | **10.35±0.09** | **14.66±0.01** |
| LSTM + BERT | **0.00±0.0** | 04.56±0.02 | 12.33±0.12 | 21.72±0.14 |

(a) Reduced-size corpora

| | Juven's CSL Data | | GoLD Data | | Robot Data |
|---|---|---|---|---|---|
| **Model** | **Valid** | **Exact** | **Valid** | **Exact** | **Exact** |
| ESN-offline + One-Hot | 46.60±0.27 | 63.30±0.35 | 29.49±0.25 | 30.38±0.45 | 42.30±0.14 |
| ESN-offline + GloVe | 44.40±0.31 | 61.00±0.37 | 48.93±0.28 | 53.90±0.24 | 57.42±0.23 |
| ESN-offline + fine-tuned BERT | 20.70±0.16 | 40.20±0.18 | 44.57±0.26 | 47.48±0.41 | 43.00±0.11 |
| ESN-offline + BERT | 24.50±0.20 | 43.60±0.24 | 52.20±0.24 | 54.78±0.35 | 45.50±0.14 |
| ESN-online FL + One-Hot | 02.90±0.01 | 29.40±0.24 | 19.23±0.22 | **26.92±0.29** | 37.12±0.06 |
| ESN-online FL + GloVe | 06.00±0.07 | 40.20±0.31 | 20.27±0.26 | 32.56±0.24 | 38.09±0.14 |
| ESN-online FL + fine-tuned BERT | **02.52±0.01** | **26.00±0.18** | **17.45±0.11** | 28.89±0.19 | **34.20±0.06** |
| ESN-online FL + BERT | 02.72±0.01 | 28.50±0.20 | 27.24±0.12 | 54.40±0.21 | 35.34±0.10 |
| ESN-online CL + One-Hot | 18.64±0.13 | 39.52±0.31 | 21.69±0.46 | 32.48±0.48 | 57.10±0.55 |
| ESN-online CL + GloVe | 42.60±0.56 | 72.90±1.01 | 22.14±0.64 | 36.42±0.76 | 59.96±0.64 |
| ESN-online CL + fine-tuned BERT | 27.28±0.19 | 54.00±0.34 | 18.37±0.40 | 34.04±0.28 | 58.86±0.20 |
| ESN-online CL + BERT | 32.86±0.20 | 60.88±0.41 | 22.30±0.46 | 52.49±0.44 | 60.17±0.33 |
| RandLSTM + One-Hot | 100.0±0.0 | 100.0±0.0 | 71.11±1.61 | 75.34±1.82 | 79.53±1.51 |
| RandLSTM + GloVe | 100.0±0.0 | 100.0±0.0 | 84.48±2.32 | 84.83±2.10 | 88.88±1.04 |
| RandLSTM + fine-tuned BERT | 100.0±0.0 | 100.0±0.0 | 72.02±1.64 | 72.02±2.03 | 87.34±0.89 |
| RandLSTM + BERT | 100.0±0.0 | 100.0±0.0 | 76.31±1.45 | 80.17±1.67 | 87.91±1.21 |
| LSTM + One-Hot | 99.64±0.01 | 99.82±0.01 | 42.89±0.56 | 48.14±0.65 | 75.67±0.54 |
| LSTM + GloVe | 99.20±0.01 | 99.99±0.00 | 65.18±0.84 | 70.89±0.91 | 86.57±0.87 |
| LSTM + fine-tuned BERT | 97.84±0.01 | 98.90±0.01 | 44.18±0.46 | 47.26±0.68 | 72.47±0.41 |
| LSTM + BERT | 98.10±0.01 | 99.99±0.01 | 48.28±0.44 | 52.40±0.66 | 78.60±0.45 |

(b) Complex corpora

Table 7.1: Results for *reduced-size corpora datasets* (left) and *complex corpora datasets* (right) for 4 input representations (BERT, One-Hot, GloVe, fine-tuned BERT) using the 5 model settings (ESN-offline, ESN-online FL, ESN-online CL, 1000-unit RandLSTM, 20-unit LSTM). Object vocabulary sizes: *reduced-size corpora datasets* (4 for Juven's; 10 for GoLD), *complex corpora datasets* (50 for Juven's; 47 for GoLD, 11 for Robot Data). (bold) Best result for each column, and (underlined) 2nd and 3rd results for each column. ESNs outperform LSTMs for all complex datasets.

models; these results demonstrate the biological plausibility of ESNs than RandLSTM. (iii) ESN-online FL performed significantly better than ESN-online CL and offline methods.

**Complex Corpora Results:** To evaluate the performances of two models on complex datasets, we chose the larger number of objects from three datasets: 50 for Juven's, 47 for GoLD, and 11 for Robot, as shown in Table 7.1. Considering complementary results to Table 7.1 : (i) For three datasets with more objects, ESN showcases a better *Valid* and *Exact* error performance than LSTM. Thus, in the general case, the ESN outperforms the LSTM.

### ESN: Effects of Offline vs Online Learning:

To explore the biological plausibility of ESN, we compare the CSL task performance on three datasets between offline and online (FL and CL) learning methods. Tables. 7.1 (a). and 7.1 (b) report the CSL task performance of ESNs where the online learning method using FL yields better performance than online CL and offline learning, indicating the more biological plausibility of ESNs during online FL and the cognitive process of sentence comprehension. To investigate the internal states of ESN during online learning, we report the absolute variation of the activation of reservoir neurons during the processing of the sentence in Fig. 7.6 (a). Since we do not use any feedback in our reservoir, the states of the reservoir are fully determined by its initial random weights and the inputs received. In fact, the learning process happens by combining the useful activities given the random projections of the inputs done in the reservoir.

### STATISTICAL SIGNIFICANCE BETWEEN ESN AND LSTM:

For the *'Valid'* error, our findings indicate a statistically significant difference between the ESN and LSTM models on the complex corpora dataset, with p-values reaching as $1.62\text{x}10^{-9}$ ($< 0.05$) for Juven's CSL Data and $3.96\text{x}10^{-5}$ ($< 0.05$) for the GoLD Data. These results are reinforced by similar significant outcomes when comparing 'Exact' errors, where, for instance, the comparison on the Juven's dataset yielded a p-value of $1.79\text{x}10^{-11}$ ($< 0.05$), p-value of 0.0056 ($< 0.05$) for GoLD dataset and p-value of $7.02\text{x}10^{-8}$ ($< 0.05$) for Robot dataset, showcasing the ESN model's enhanced performance.

Moreover, our analysis extends to the exploration of FL within ESN models, where the statistical tests consistently underscore significant performance improvements over LSTM models across various datasets, as evidenced by p-values such as $4.75\text{x}10^{-4}$ ($< 0.05$)(GoLD Data for *'Valid'* error) and $7.02\text{x}10^{-8}$ ($< 0.05$) (Robot Data for *'Exact'* error). These statistically significant results, derived from careful application of the two-sample t-test, not only validate the performance differences highlighted in our study but also underline the robustness and reliability of ESN models in handling complex datasets.

For the reduced corpora, no statistically significant differences were observed between the ESN models employing FL or CL and the LSTM models, in terms of both *'Valid'* and *'Exact'* scores, across the reduced corpora from Juven's CSL and GoLD datasets. The relatively high p-values in all tests (all p-values > 0.05) suggest the null hypothesis of equal means between the compared groups cannot be rejected. This indicates that both ESN and LSTM models exhibit comparable performance on reduced corpora.

Conversely, the comparison between ESN models utilizing FL and those utilizing CL on the Juven's CSL Data demonstrated statistically significant differences in both *'Valid'* and *'Exact'* scores, with p-values of 0.032 and 0.046, respectively. This signifies notable performance disparities between ESN models using FL and CL approaches. However, for the GoLD data, the p-values exceeded 0.05, revealing no statistically significant differences between the FL and CL models in terms of both *'Valid'* and *'Exact'* scores. This analysis sheds light on the impact of different learning strategies (FL vs. CL) on ESN model performance and their comparative effectiveness to LSTM models in scenarios involving reduced corpora.

### ESN ONLINE LEARNING VS RANDOM LSTM:

In order to explore how RandLSTM learns to perform the CSL task, we compare the performance of RandLSTM with ESN Online models. Tables. 7.1 (a) and 7.1 (b) report the CSL task performance of RandLSTM where both input and LSTM layers are kept frozen, and training happens at the output layer similar to ESN models. From Tables. 7.1 (a) and 7.1 (b), we observe that the ESN online learning methods display supremacy over RandLSTM indicating that the more biological plausibility of ESNs compared to RandLSTMs. Further, we compare the computational complexity of ESNs with RandLSTM on complex corpora across three datasets. We observed the following insights from Table 7.2: (i) From a computational efficiency perspective, one of the major limitations of the RandLSTM model is that

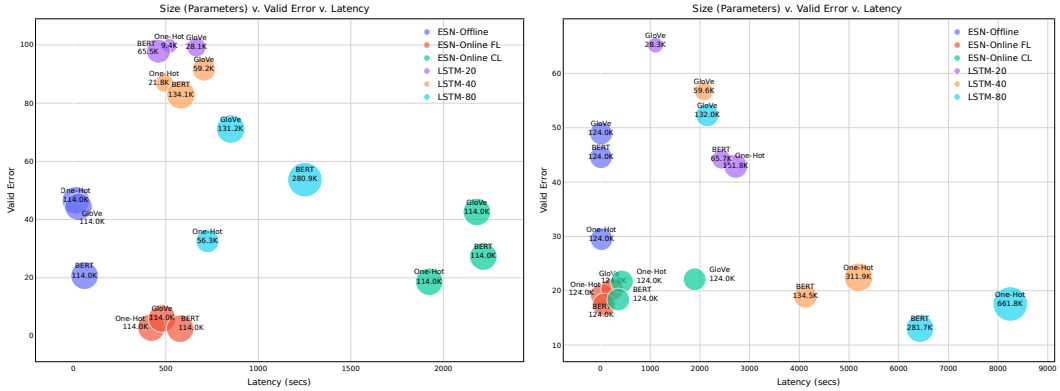
| | Latency (sec.), One-hot/GloVe/BERT | | |
|---|---|---|---|
| Model | Juven's (114K) | GoLD (124K) | Robotic (591K) |
| ESN Offline | 13.9 / 28.3 / 61 | 19.5 / 11.39 / 9.76 | 98 / 21 / 31 |
| ESN Online FL | 423 / 480 / 578 | 25 / 235 / 69 | 1,448 / 1,133 / 1,178 |
| ESN Online CL | 1,924 / 2,181 / 2,216 | 431 / 1,895 / 357 | 6,900 / 6,128 / 5,683 |
| RandLSTM | 3,025 / 4,211 / 1,541 | 19,602 / 11,324 / 18,304 | 12,261 / 11,974 / 16,092 |

Table 7.2: Training latency comparison for ESNs and Random-LSTMs. Each model has 1000 hidden recurrent units. Total number of trained parameters is provided for each dataset. *Fine-tuned BERT* has the same latency as *BERT* because dimensions are identical.

training time is computationally expensive, (ii) In contrast, ESN models are more efficient and require lower training time.

### Model size, Latency and Error Trade-off:

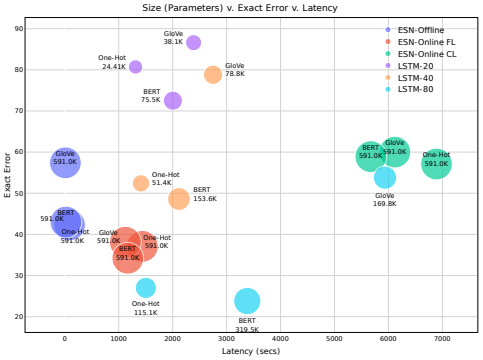
Our main goal is to build models that are efficient for human-robot interactions. Therefore, it is crucial to explore trade-offs between model size, latency, and error. For complex corpora datasets (Juven with 50 objects on *valid error*, GoLD with 47 objects on *valid error*, and Robot data with 25 objects on *exact error*), we analyze the model size, latency, and error score trade-off in Figs. 7.5 (a), 7.5 (b) and 7.5 (c) across two models: ESNs (offline, online + FL, online + CL) and LSTMs (20, 40, and 80 hidden units). Typically, ESN models have fewer parameters, where the number of parameters depends on the target vector dimension.
**Juven's Data:** From Fig. 7.5 (a), we observe that the ESN-online FL model showcases lower valid error using 114K parameters with a model training latency of 500 seconds compared to Offline, ESN-online CL, and LSTM with 20 and 40 hidden units (higher latency time for model training). It is clearly observed that the ESN models have better computational complexity in terms of latency and model size and report better performance.
**GoLD Data:** From Fig. 7.5 (b), we observe that the ESN-online FL model showcases lower valid error using 124K parameters with a model training latency of 64 seconds compared to Offline, ESN-online CL, and LSTM with 20 and 40 hidden units (higher latency time for model training). It is clearly observed that the ESN model has better computational complexity in terms of latency and model size.
**Robot Data:** Fig. 7.5 (c) shows the model size, latency and error trade-off on the Robot dataset. From the Fig. 7.5 (c), we observe that LSTM with 80 hidden units (115K parameters for one-hot and 319K parameters for GoLD) model showcases lower exact error compared to ESN with 591K parameters. Although the parameters of ESNs are much higher than LSTMs, the training of ESNs displays lower latency than LSTMs (higher latency time for model training). Since Robot data have a higher target dimension (591 binary vector), the ESN model parameters are much higher than LSTM. However, it does not affect the relative latency or error performance of ESNs much compared to LSTMs.
**Insights:** Hence, it is clearly observed from the Figs. 7.5 (a), 7.5 (b) and 7.5 (c) that ESNs display better generalizations than LSTMs for increasing the larger vocabularies. Although we compare the number of trained parameters for ESNs and LSTMS, they are not directly comparable given that they do not use the same theoretical computing principles (ESNs rely on the VC-dimension [Vapnik et al., 1994] like in Support Vector Machines).




(a) Juven (Large Data)




(b) GoLD (Large Data)




(c) Robot (Complex Corpora)

Figure 7.5: Parameters vs. Valid Error vs. Training Latency on complex corpora datasets: (a) Juven, (b) GoLD and (c) Robot. Here, the size of the bubble denotes the number of parameters.

## Qualitative Analysis

The challenge in applying simple neural network models to human-robot interaction research lies in the black-box nature of the process, where it is hard to decipher what the network learns while processing full sentences. Here, we discuss the inner working details of all the models and report the output activations of each model.

**Qualitative analysis of output units activation:** In order to understand the inner working details of both models, we plot the evolution of the output activation during the processing of a sentence across all the models (ESN Online CL, LSTM, ESN Offline, and ESN Online FL), as shown in Figs. 7.6, 7.7, 7.23, and 7.24. Observations from Figs. 7.6, 7.7, 7.23, and 7.24 that the intermediate output activations are much more meaningful and interpretable with the ESN-online CL and LSTM. However, for the ESN-online FL, the intermediate output activation cannot be interpreted with the default Final Learning (FL). As we can see in Fig. 7.24, the fluctuations seem unpredictable until the last word "END" is seen. For a correctly predicted output, the activation often "jumps" to the correct value when the last item "END" is inputted. This is due to the fact that we only apply the learning procedure at the final state, so there is no constraint on intermediate outputs. Similarly, the observations

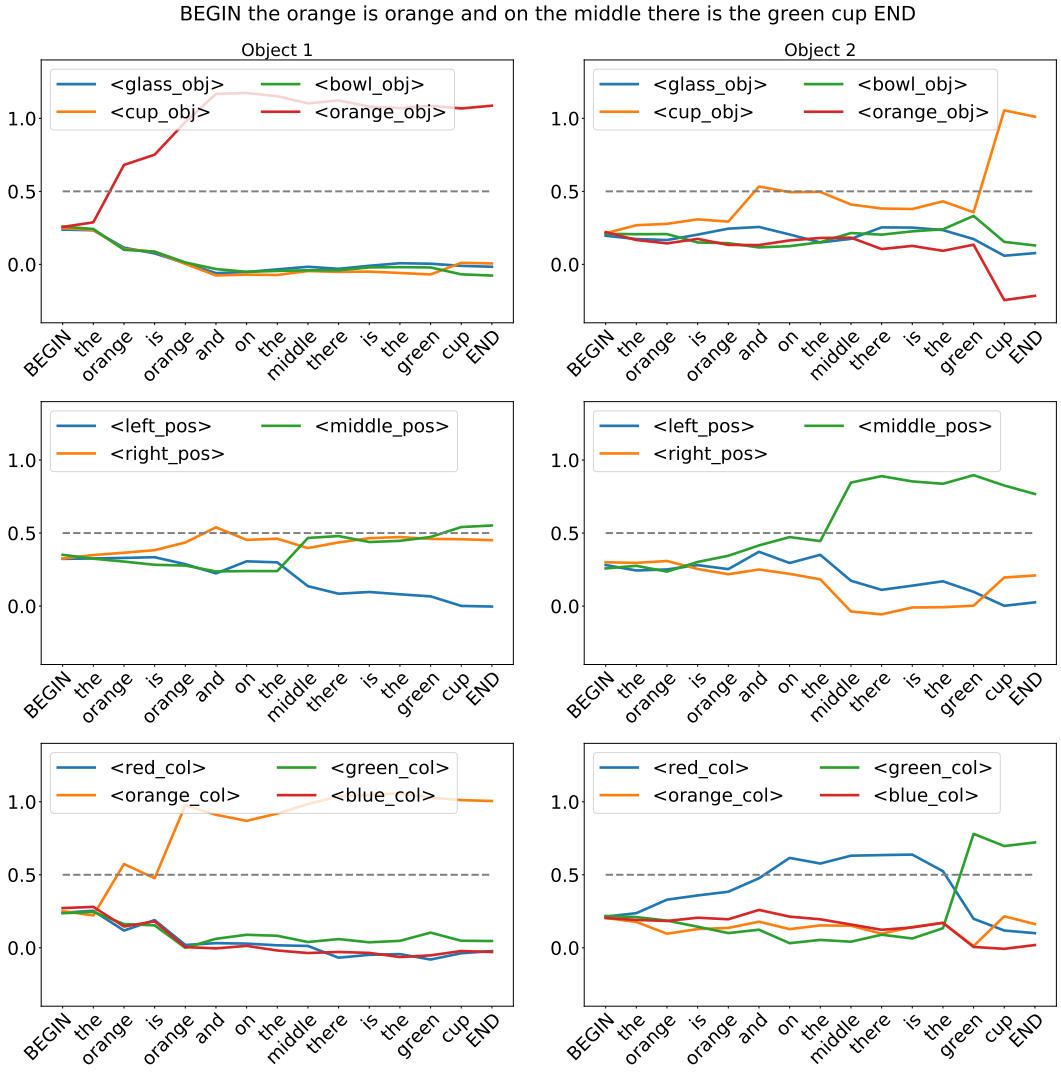



Figure 7.6: Juven's Data: Output activation of the ESN CL + fine-tuned BERT. After each word the model tries to predict the correct output. That's why we can see a jump in the correct characteristic after the related keyword is seen.

from Fig. 7.23 that the intermediate activations of the ESN-offline model cannot be interpreted due to its constant activation from the word "BEGIN" until the last word "END" is seen. Interestingly, when training the network ESN-online CL with both usual (whole sentence) and single-word sentences, the network outputs provide consistent predictions during the whole presentation of sentences, as shown in Fig. 7.6. This is because the final answer from the network can be predicted before the sentence is over, given its ongoing activations, i.e the output activity of a concept is activated once a word is pronounced.

Similar to ESN-online CL, during the training procedure of LSTM, the target outputs are given as a "whole" during all the timesteps (no particular label is given at a precise

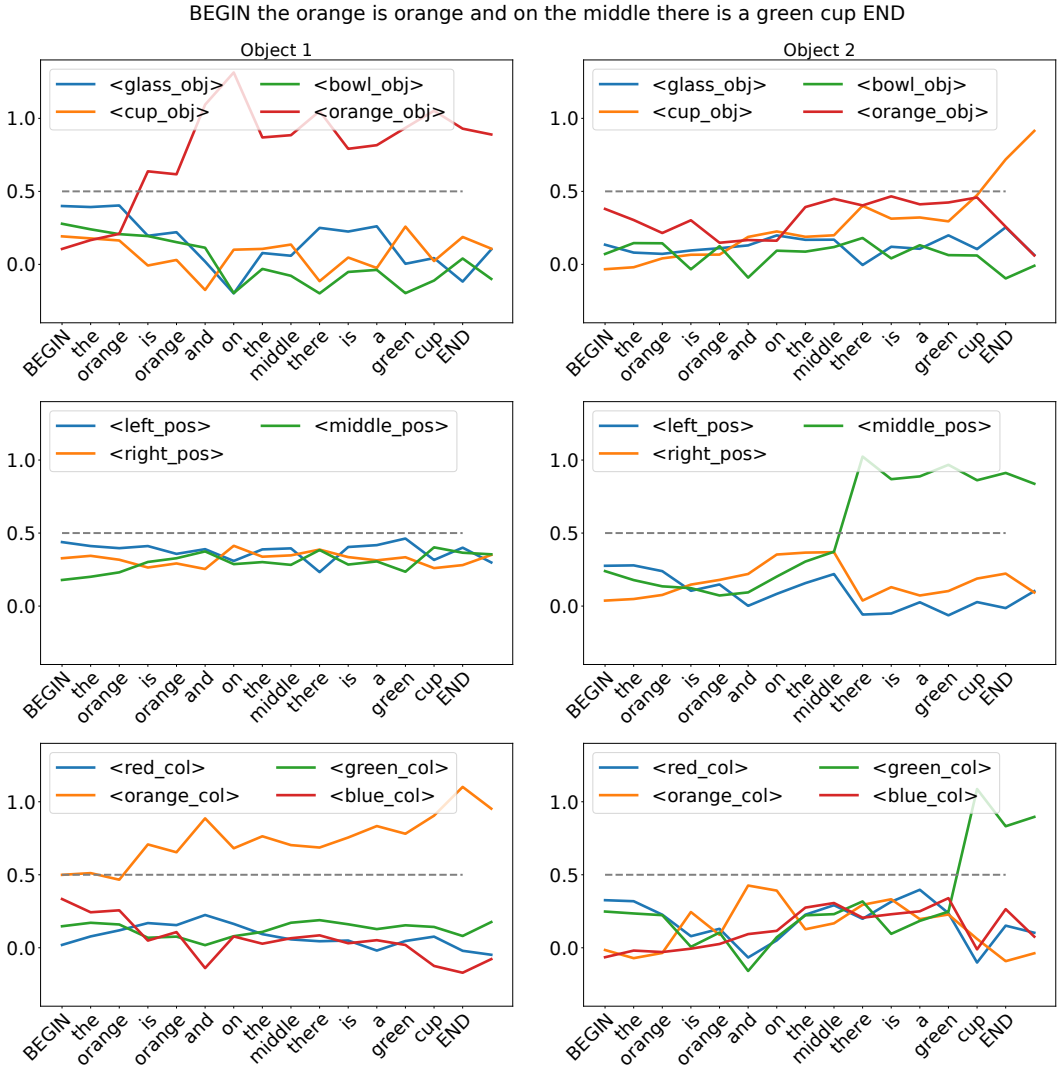



Figure 7.7: Juven's Data: Output activation of the LSTM + fine-tuned BERT. Even if the learning procedure is only applied at the end, because of its learning algorithm, intermediate states are also optimized. This is why we can also interpret these transitional steps: they behave similarly to the ESN online trained with CL.

time corresponding to a precise word). However, we can observe a spike in the activity of the concept as the model sees the corresponding keyword, as depicted in Fig. 7.7. For instance, we can observe a spike in the activity of the concept <Orange_obj> (i.e. the concept activated when the object 1 is Orange), and the spike is quickly inhibited when the following word "cup" is received <Cup_obj> (i.e. the concept activated when the object 2 is Cup) is seen. This phenomenon gives us a first hint on how both models are able to deal with polysemous meaning. Further, observations from Fig. 7.7 that the word "END" does not seem to affect the output of the network significantly.

For example, consider the sentence "BEGIN this orange is orange and on the middle there is the green cup END", activations are shown for the two models in the Fig. 7.7. Interestingly, the position is not mentioned for the first object <Orange_obj> in the sentence, and the position for the first object does not have any reference to the second object. Here, we can see that when the word "orange" appears twice in the first part of the sentence, the model has not yet the information that the word will be used as an adjective or noun. So, when this happens, for the LSTM, we can see a rise (i.e. a spike) in the activation of the <orange_object1> concept (i.e. the output neuron that should be activated when the first object is an orange). It is also clear that "orange" was an adjective and not a noun when it appeared a second time in the sentence. This gives a qualitative insight that fine-tuned BERT representations establish the reference from a word to an object when a full context is not provided.

## 7.5 Discussion

Grounded language acquisition is the process of learning a language - how infants can learn language by observing their environments, interacting with others, and understanding the concepts of a language as it relates to the world [Chen and Mooney, 2008, Thomason et al., 2018, Juven and Hinaut, 2020, Vanzo et al., 2020]. However, language acquisition becomes challenging when there are numerous possible meanings for a word in an utterance, introducing a high level of uncertainty. Traditional approaches for language grounding mainly focus on mapping natural language commands and task representations that are essentially sequences of primitive robot actions [Chen and Mooney, 2011, Matuszek et al., 2013, Tellex et al., 2011]. Moreover, existing robotic frameworks [Taniguchi et al., 2017, Roesler et al., 2018] do not model how children learn to understand directly from full sentences through cross-situational learning without providing specific cues such as visual cues [Roy, 2002], social cues [MacDonald et al., 2017], and auditory cues [Räsänen and Rasilo, 2015, Escudero et al., 2023], etc,. Overall, we take the language acquisition perspective to machine learning and robotics, where part of the problem is understanding how language models can perform grounded language acquisition through noisy supervision and discussing how they can account for brain learning dynamics. Our proposed framework, combining input featurization, dynamic memory, and learning modules, offers a flexible, biologically plausible architecture for investigating CSL tasks on diverse datasets.

In this paper, we investigate the ability of two sequence-based models, ESNs and LSTMs, to learn to parse sentences via noisy supervision (CSL) and compare different word representations (one-hot, GloVe, pretrained, and fine-tuned BERT). We evaluated our CSL task on three different datasets in five different settings: (i) ESN-Offline, (ii) ESN-Online FL, (iii) ESN-Online CL, (iv) RandLSTM, and (v) LSTM. These experiments yield the following insights: (1) fine-tuned BERT representation is the best representation among most models; (2) In general, ESNs display better prediction than LSTMs as the vocabulary size increases; (3a) For instance, in Juven's data, the trend of ESNs outperforming LSTMs in terms of generalization persists regardless of the sizes of LSTMs; (3b) The size of LSTMs needs to be increased to surpass the 1000-unit ESN that we took as a reference. LSTM with

20 units showcase higher performance for small datasets, but LSTM with more hidden units needs to perform reasonably well on larger corpora. (4) ESNs with online learning models are making better predictions during the processing of a sentence compared to other models. (5) The qualitative analysis reveals that both ESNs and LSTMs demonstrate better concept activation in the output during processing of a sentence. (6) ESNs have a better trade-off on all three datasets with better prediction error along with low latency.

Overall, our above proposed framework exhibits three key properties: interpretability, generalizability, and computational efficiency in the two sequence-based models. We discuss the details of these three properties below.

**Interpretability & Generalisability:** The challenge in applying simple neural network models to human-robot interaction research lies in the black-box nature of the process, where it is hard to decipher what the network learns while processing full sentences. In order to address this and to understand the model mechanisms, we devised the following: (i) visualising the output activations of all the models during the processing of full sentences and (ii) displaying the output value matrix of target concepts, as shown in Figs. 7.25, and 7.26 (please refer the supplementary). From Fig. 7.25, we observe that the ESN model captures the polysemous words. For instance, *"the orange on the right is green and there is a orange cup on the middle"*, the associated concepts are *orange, right, green* for the first object, and *cup, middle, orange* for the second object. The model learns the meaning of word "orange" as a Noun for the first object and as a color for the second object "cup", showcasing its ability to discern different meanings in context. Similarly, Fig. 7.26 captures the two colors "red" and "black" for the object "pliers". Since there is only one object present in the sentence, we do not see any activations for the second object.

From Tables. 7.1 (a) and (b), it is evident that LSTMs showcase higher performance than ESNs on both Valid and Exact errors on reduced corpora, such as Juven's (4 objects) and GoLD (10 objects) datasets. However, as the vocabulary size increases, ESNs demonstrate superior Valid and Exact error performance, suggesting that, in general, ESNs outperform LSTMs. Moreover, ESN displays better performance than RandLSTM while considering the same number of neurons in both models; these results demonstrate the biological plausibility of ESN compared to RandLSTM. Additionally, LSTMs perform worse for the Juven's dataset compared to other complex datasets, while the reverse is true for ESNs—they perform better on Juven's compared to more complex datasets with longer sentences..

**Computational Efficiency:** From a computational efficiency perspective, one of the major limitations of the LSTM model is that it uses BPTT to optimize the weights, which requires more training time and is computationally expensive. In contrast, ESN models have fewer parameters, and the number of parameters depends on the target vector dimension. Moreover, their computational complexity is more efficient as they employ ridge regression at the readout layer to learn the weights and require no training in the initial and reservoir layers. Although we compare the number of trained parameters for ESNs and LSTMs, they are not directly comparable given that they do not use the same theoretical computing principles (ESNs rely on the VC-dimension [Vapnik et al., 1994] like in Support Vector Machines). To overcome the above limitation, we compare the RandLSTM and ESNs with the same number of hidden units in both models. Observations from Table 7.2 indicate that ESN models are more efficient and require lower training time than RandLSTMs (about three orders of

magnitude in training CPU time). Moreover, we choose arbitrarily to have 1000 units in ESNs as it is a common number; this parameter can be increased to enhance performance.
**Limitations & Future work:** The performances on the GoLD and Robot data set could seem low compared to Juven's; however, it is crucial to consider that these datasets involve highly complex sentences that may contain unseen words. For example, the GoLD dataset includes sentences with many unseen words while describing a few concepts, as seen in the sentence: *"A single small red skinned potato is laying on its side with the pointier end pointing left and two dimpled eye facing me."*, the associated concepts are: *red, potato, small, left* for the first object, and *eye* for the second object). Similarly, the Robot dataset contains complex robotic commands with more actions and relations are described for few concepts, as illustrated in the sentence: e.g. *"pick up the gray block located on top of the blue tower near the left edge and place it on top of the red and green tower that is nearest to you"*, the associated concepts are: *pick, gray, top, blue, block, tower, near, left edge, place, red, green, nearest*.

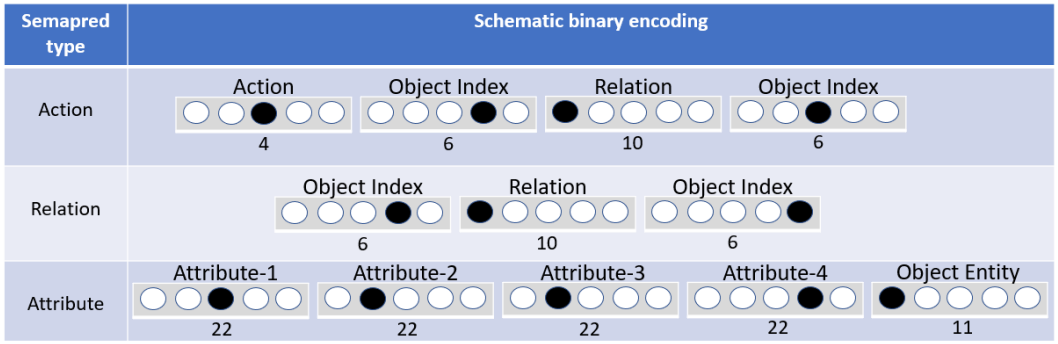
It's important to note that our experimental setup did not incorporate speech or image data. By limiting our analysis to text data, we aimed to isolate and examine the nuances of language grounding and incremental learning processes as they pertain to textual input, without the additional complexities that speech or image data might introduce. However, future research that might include speech, images, or other types of data to create a more holistic understanding of language acquisition and processing.

The major limitation of existing grounded language methods consider language that describes immediate and instantaneous actions (i.e. grasping language expressing concepts at a spatial level), primarily contributing to task learning. Recently, Karch et al. [2021] proposed a grounded language model that grasps concepts at both spatial and temporal level to learn the meaning of Spatio-temporal descriptions of behavioral traces of an embodied agent. In the future, we aim to train robots on more comprehensive grounding and diverse datasets, encompassing speech and multi-modal grounded language datasets while modeling infants' language acquisition. - In future work, we plan to bridge the gap by attempting to process sentences starting from speech (we have preliminary work showing that this is possible). Moreover, we aspire to make the architecture more grounded by using the images of datasets. The MSCoCo [Lin et al., 2014] is a good candidate dataset as it includes both segmented images and speech. ESNs show very good performance given the "light" training they use. Theoretically, longer sentences could become difficult for a limited-size ESN, while LSTMs should be better. In future work, we will look at how attention-like mechanisms could be integrated into these simple models, enhancing their ability to gate information [Strock et al., 2020].

# *Appendix for: CROSS-SITUATIONAL LEARNING TOWARDS LANGUAGE GROUNDING*

| | | | | | |
|---|---|---|---|---|---|
| Action 1-2 | move(prism_1, above, cube_2) | EEE(EEE, EEE, EEE) | | | |
| Relation 1-2 | is(cube_2, above, corner_3) | is(EEE, EEE, EEE) | | | |
| Attribute 1-5 | red(prism_1) | EEE(prism_1) | EEE(prism_1) | EEE(prism_1) | EEE(prism_1) |
| Attribute 6-10 | blue(cube_2) | EEE(cube_2) | EEE(cube_2) | EEE(cube_2) | EEE(cube_2) |
| Attribute 11-15 | back(corner_3) | left(corner_3) | EEE(corner_3) | EEE(corner_3) | EEE(corner_3) |
| Attribute 16-20 | EEE(EEE) | EEE(EEE) | EEE(EEE) | EEE(EEE) | EEE(EEE) |
| Attribute 21-25 | EEE(EEE) | EEE(EEE) | EEE(EEE) | EEE(EEE) | EEE(EEE) |

(a) The template for an example command in the Robot dataset.




(b) The SemaPred representation of binary encoding.

Figure 7.8: (a) The template for an example command in the Robot dataset. The missing predicates and slots have to be filled with empty tokens (EEE). Relation predicates always start with the word *is*. Here, template image is sourced from Twiefel [2020]. (b) The SemaPred representation of binary encoding: (i) To produce the whole output vector, all vectors are concatenated. (ii) It consists of 2 action vectors, 2 relation vectors and 5 attribute vectors in this. (iii) The output vector size is 2 * (4 + 6 + 10 + 6) + 2 * (6 + 10 + 6) + 5 * (22 + 22 + 22 + 22 + 11) = 591 for the given data set. Note: Unlike for Action and Relation SemaPreds, Attribute SemaPred does not encode the index of the entity but the entity itself.

## 7.6 Word Seen during Model Training

Fig. 7.10 displays the average number of times a word is seen during model training on three datasets. From Fig. 7.10, we can see that GoLD data contains more vocabulary (2417) words for 47 objects data), and the number of times a word is seen in model training is low compared to Juven's dataset. If an unseen word appears the corresponding concept outputs will be at 0, because corresponding weights would never be trained, i.e. all corresponding weights will be at 0. This makes the CSL task more difficult for big vocabularies.

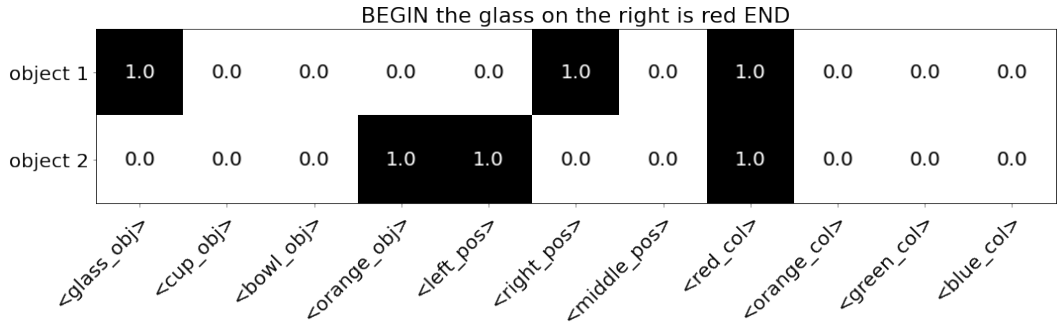



Figure 7.9: CSL Task Noisy Supervision Output: The target is a noisy supervision vector that contains additional concepts (<orange_obj>, <red_col>) that are not present in the input sentence.

| Dataset | #Objects | #Train Sentences | #Test Sentences | Vocabulary Size | Avg. seen word (Train) | Avg. seen word (Test) |
|---|---|---|---|---|---|---|
| Juven's | 50 | 1000 | 1000 | 66 | 217.28 | 216.65 |
| GoLD | 47 | 1000 | 7000 | 2417 | 7.75 | 18.89 |
| Robot | 11 | 1000 | 909 | 129 | 99.07 | 86.79 |

Figure 7.10: Corpus statistics for Juven's, GoLD and Robot datasets, including the average number of times a word is seen (Avg. seen word) during model training and testing.

| ↓Dataset | #Objects | #Colors | #Positions | #Objects Described | #Actions | #Relations |
|---|---|---|---|---|---|---|
| Juven's | 50 | 3 | 2 | 2 | NA | NA |
| GoLD | 47 | 7 | 6 | 2 | NA | NA |
| Robot | 11 | 8 | 9 | 5 | 4 | 3 |

Table 7.3: Dataset Statistics.

## 7.7 Quantitative Analysis: Varying the Objects in the Vocabulary

We compare the performance of our models with fine-tuned BERT while varying the number of objects from 4 to 50 for Juven's and 10 to 47 for GoLD datasets. The qualitative analysis for one-hot representations for two datasets is reported in the Appendix.

**Juven's CSL Results:** The results of the fine-tuned BERT feature representation is shown in Fig. 7.11. Observations from Fig. 7.11 that the 20-unit LSTM + fine-tuned BERT showcase an optimized performance compared to ESN in the 4-object dataset. However, we can see that the error explodes as soon as we increase the vocabulary size (i.e. number of objects) compared to the 4-object dataset for which it was designed. We then conducted another experiment with a 40-unit LSTM + fine-tuned BERT, applied dropout with a keep-probability of 0.2, and trained the model for a maximum of 70 epochs. It can be seen that the [*Valid*, *Exact*] errors for fine-tuned BERT [0.5, 6.58] perform better than One-Hot [3.94, 8.62], and BERT [0.1, 16.2] (Figs. 7.12 (a), and 7.12 (b)). To overcome the over-

fitting problem, we stop training if the validation loss does not decrease for five consecutive epochs. We found that this bigger model was able to outperform the ESN on exact error until the 12-objects dataset. After 12-objects, both *Valid* and *Exact* error began to rise higher than the ESN-FL model. Another LSTM + fine-tuned BERT model with 80 units was tested. We found that this model globally keeps the error lower than the two other LSTMs, especially for a high number of objects. Nonetheless, the ESN + fine-tuned BERT also outperformed it for all the range tested. In the end, Fig. 7.11 describe that the ESN with fine-tuned BER is able to keep the error low on challenging datasets despite having hyper-parameters optimized to perform well on a 4-object dataset. Whereas for the LSTMs, they can successfully learn these more featured datasets.

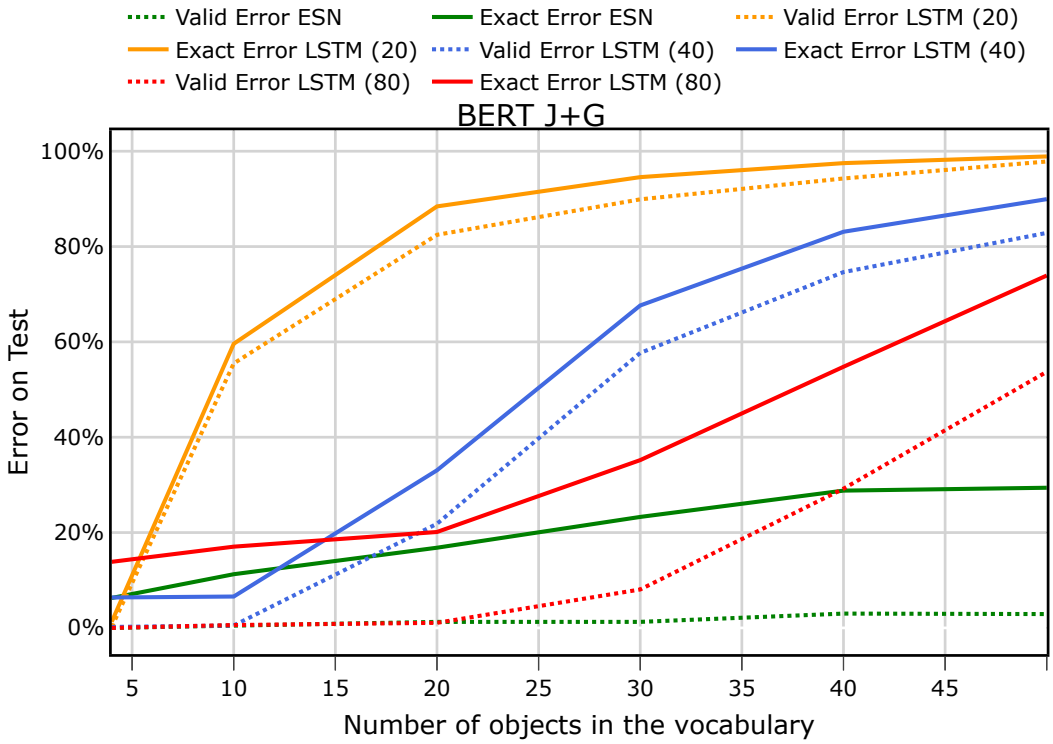
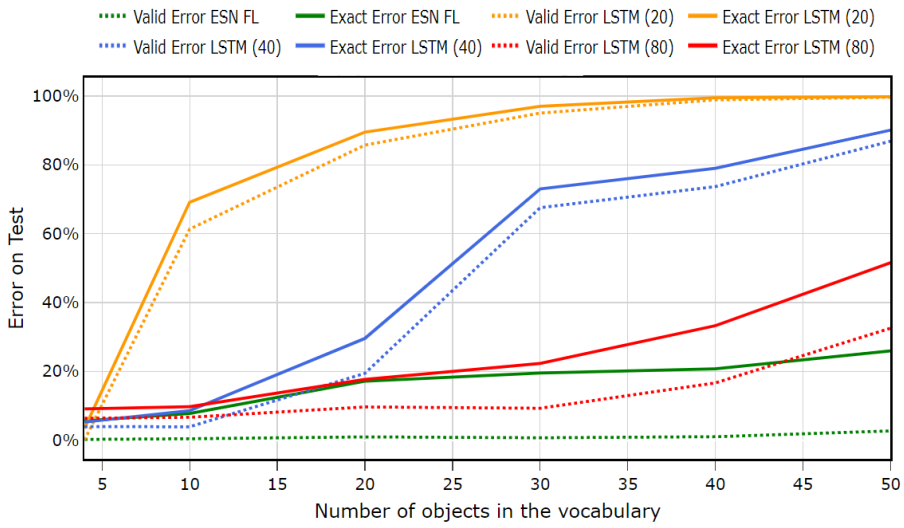


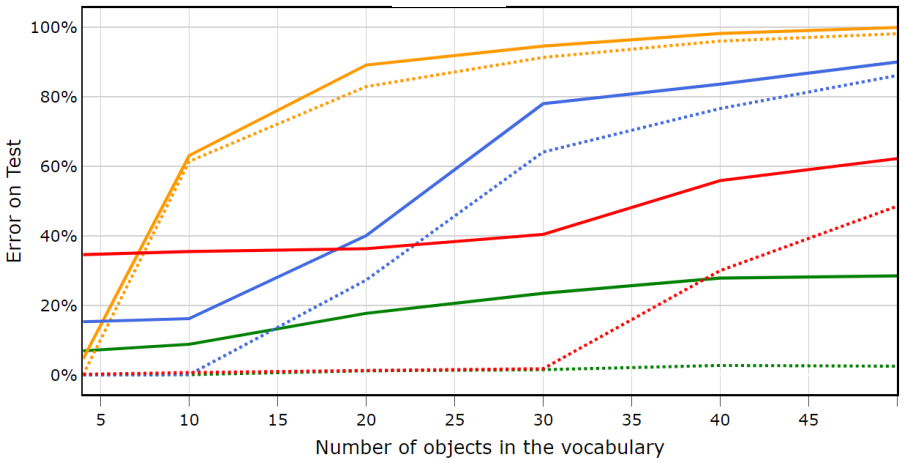

Figure 7.11: Juven's data: comparison of errors for ESN / LSTM with Fine-tuned BERT CSL. Full lines: Exact Errors. Dotted lines: Valid Errors.

**GoLD Results:** GoLD dataset was developed to reflect and provide data for domains in which dynamic human-robot teaming is a near-term interest area [Jenkins et al., 2020]. Compared to Juven's data, GoLD data provides grounded language learning in a human-centric environment: a robot talking to a person may have a partial view or understanding of an object, or vice versa.

Similar to Juven's data, we performed the experiments by varying the objects from 10 to 47 on GoLD data using fine-tuned BERT, as shown in Fig. 7.13. The results of the experiment for one-hot and pretrained BERT are shown in the Appendix (please refer Figs. 7.14 (a) and 7.14 (b). We make the following observations: (i) ESN + FL using fine-tuned BERT




(a) One-Hot Representations.




(b) pretrained BERT Representations.

Figure 7.12: Juven's data: comparison of errors for ESN / LSTM with (a) One-Hot and (b) pretrained BERT CSL representations. Full lines: Exact Errors. Dotted lines: Valid Errors.

CSL, we can see that the *Exact* error is better, but the *Valid* error explodes as soon as we increase the number of objects) compared to the 10-object dataset, as depicted in Fig. 7.13. (ii) With the model ESN + FL using fine-tuned BERT outperform the *Valid* and *Exact* errors while varying from 10 to 47-objects dataset than one-hot encoding and pretrained BERT representations as input. We then conducted two other experiments with LSTM by varying the hidden units from 20, 40, and 80, trained with a dropout of 0.2 on 70 epochs. By comparison on a test set, we found that LSTM + fine-tuned BERT model was able to outperform the ESN until 20-objects datasets. After that, valid error began to rise higher than the ESN model.

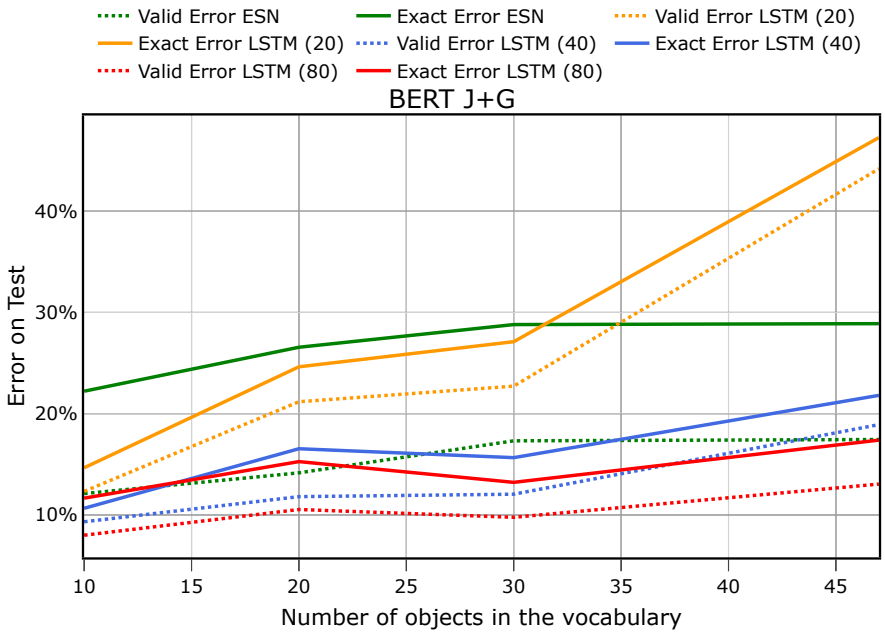
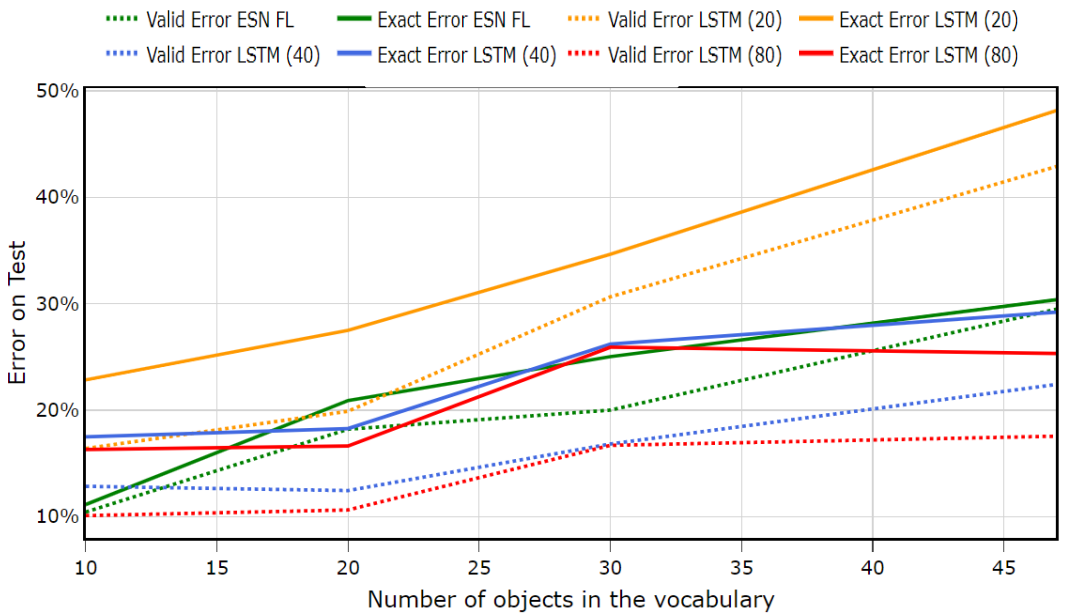
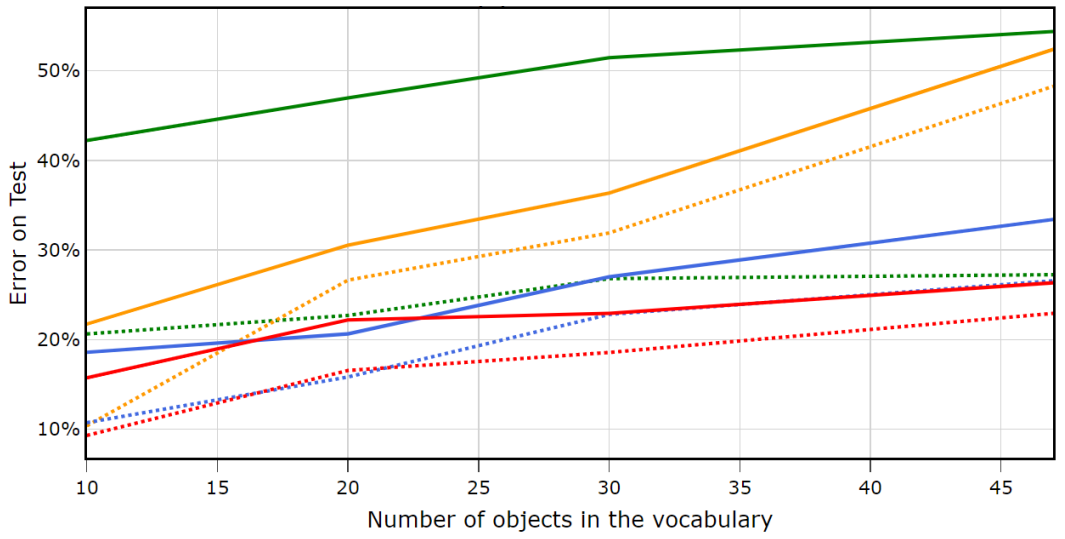



Figure 7.13: GoLD data: comparison of erros for ESN / LSTM with fine-tuned BERT CSL. Full lines: Exact Errors. Dotted lines: Valid Errors.




(a) One-Hot Representations.




(b) pretrained BERT Representations.

Figure 7.14: GoLD data: comparison of erros for ESN / LSTM with fine-tuned BERT CSL representations. Full lines: Exact Errors. Dotted lines: Valid Errors.

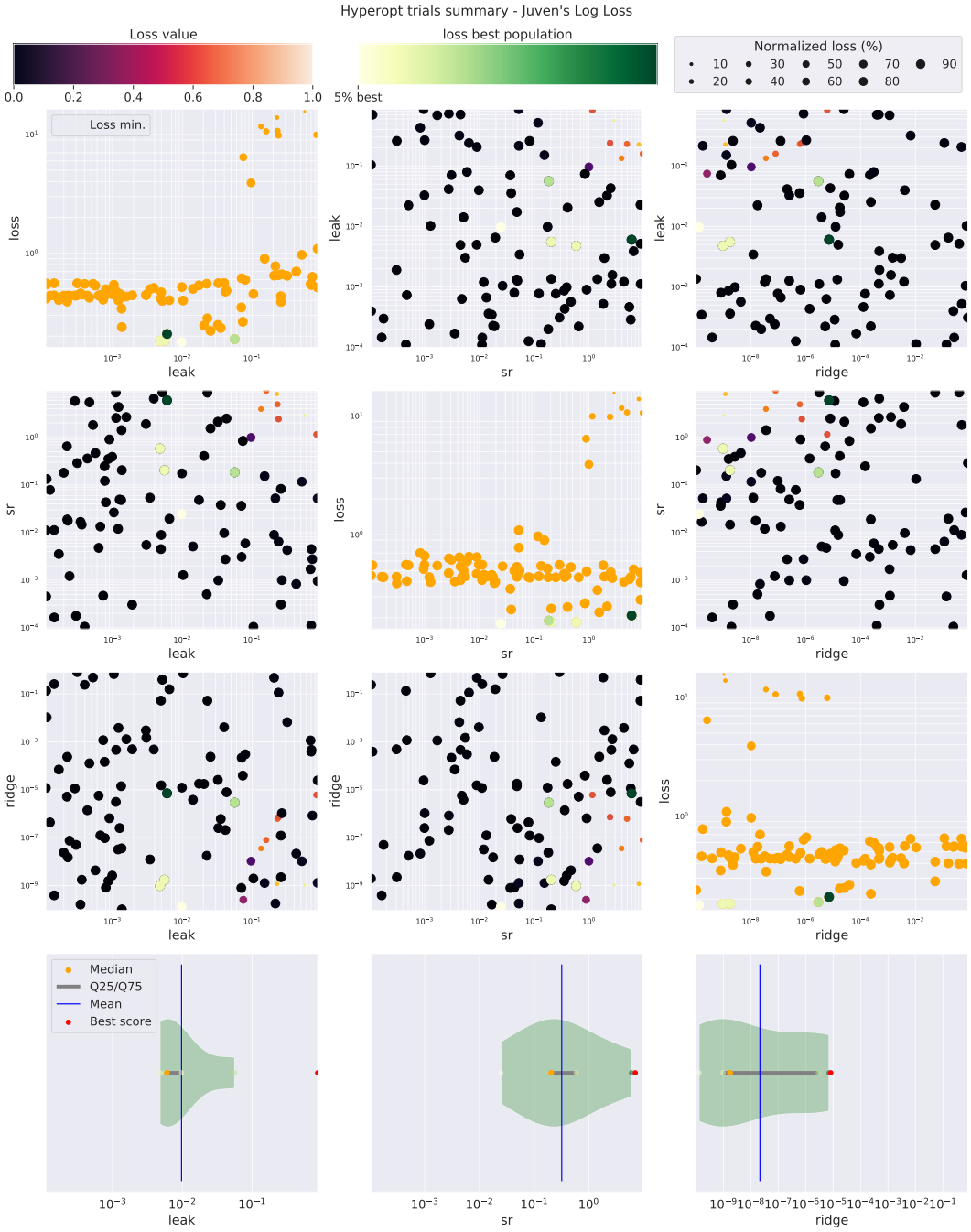



Figure 7.15: Hyper-parameter search dependence plot with **Cross-Entropy loss**, for ESN-Online FL with Juven's data.

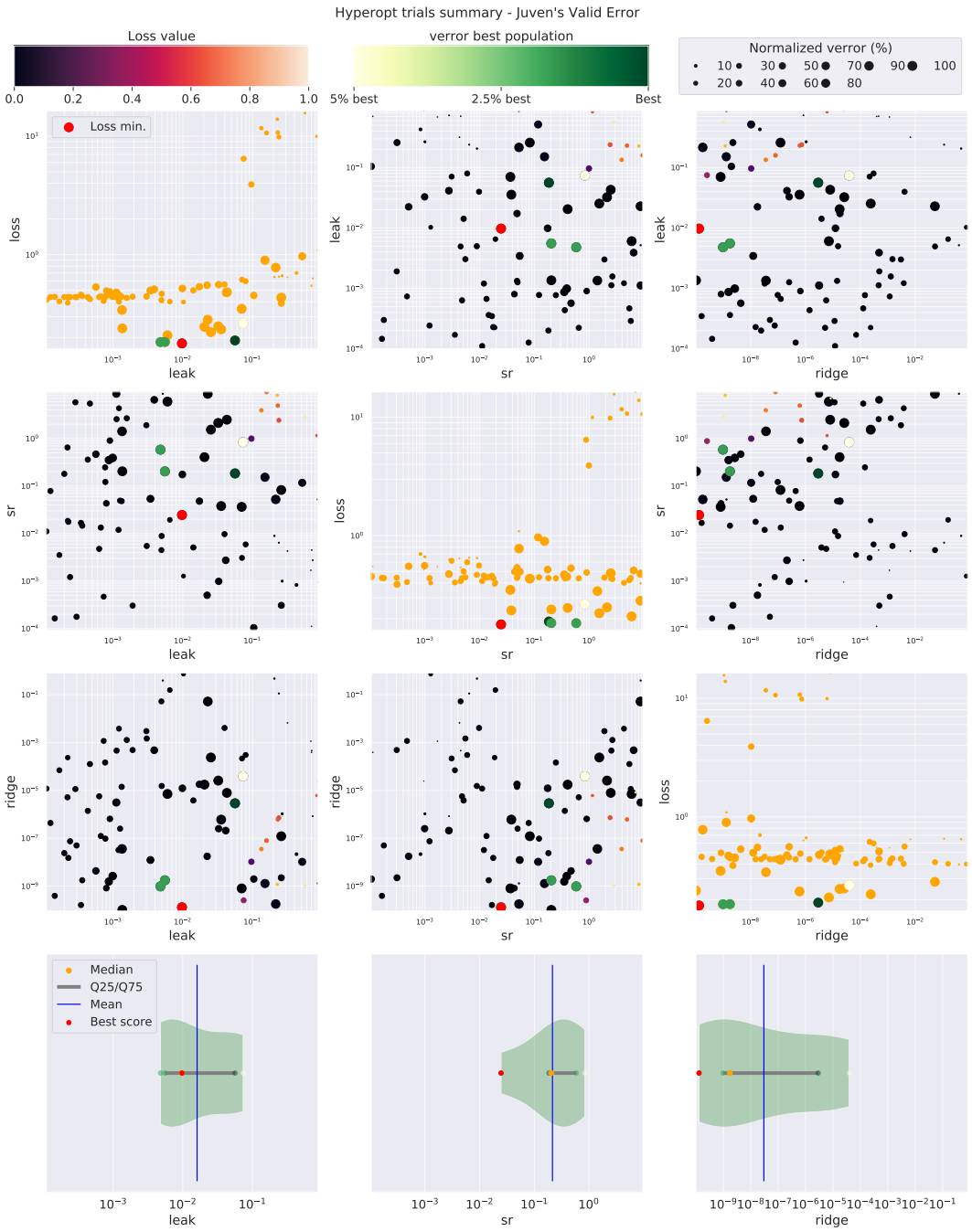



Figure 7.16: Hyper-parameter search dependence plot with **Valid Error**, for ESN-Online FL with Juven's data.

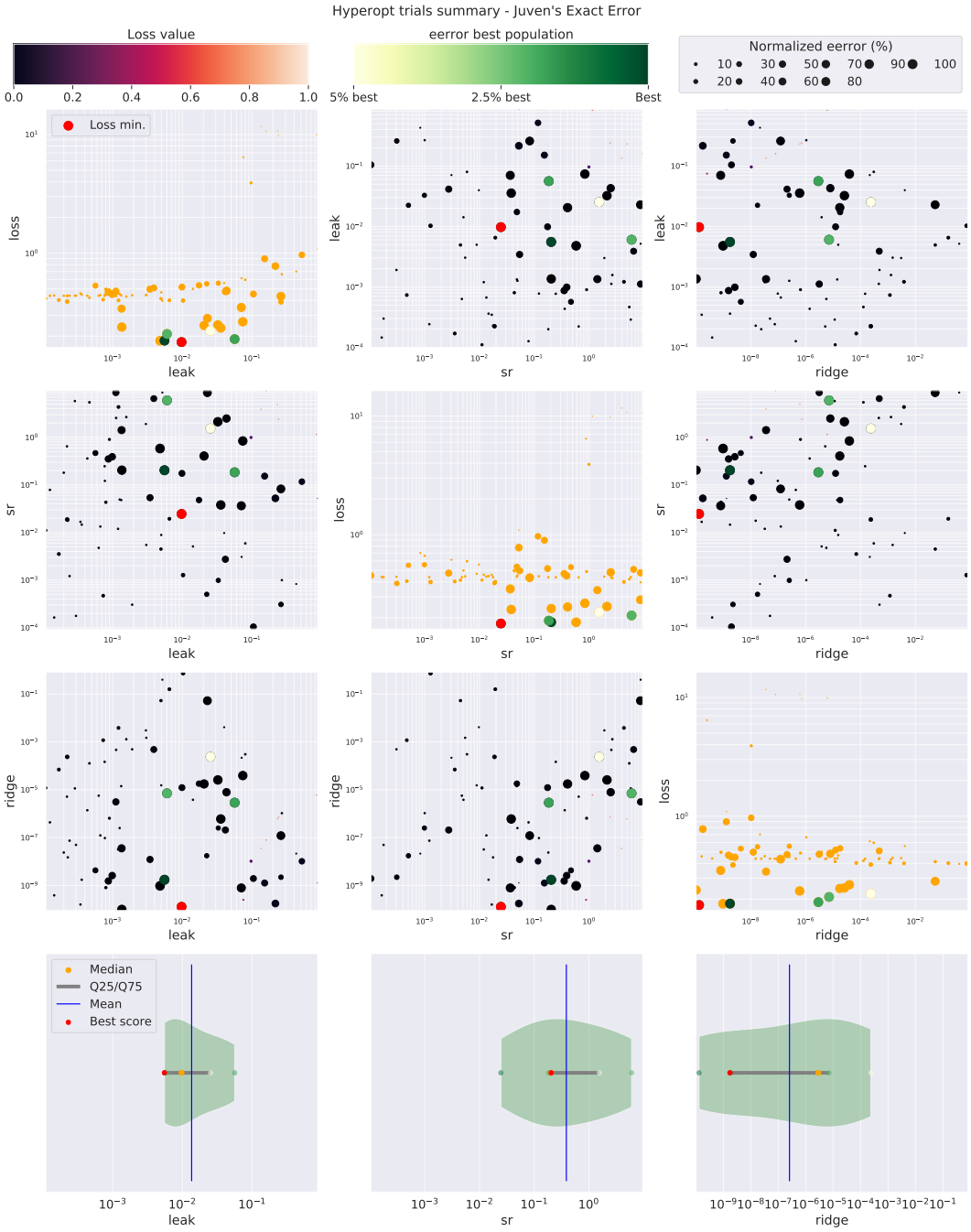



Figure 7.17: Hyper-parameter search dependence plot with **Exact Errorr**, for ESN-Online FL with Juven's data.

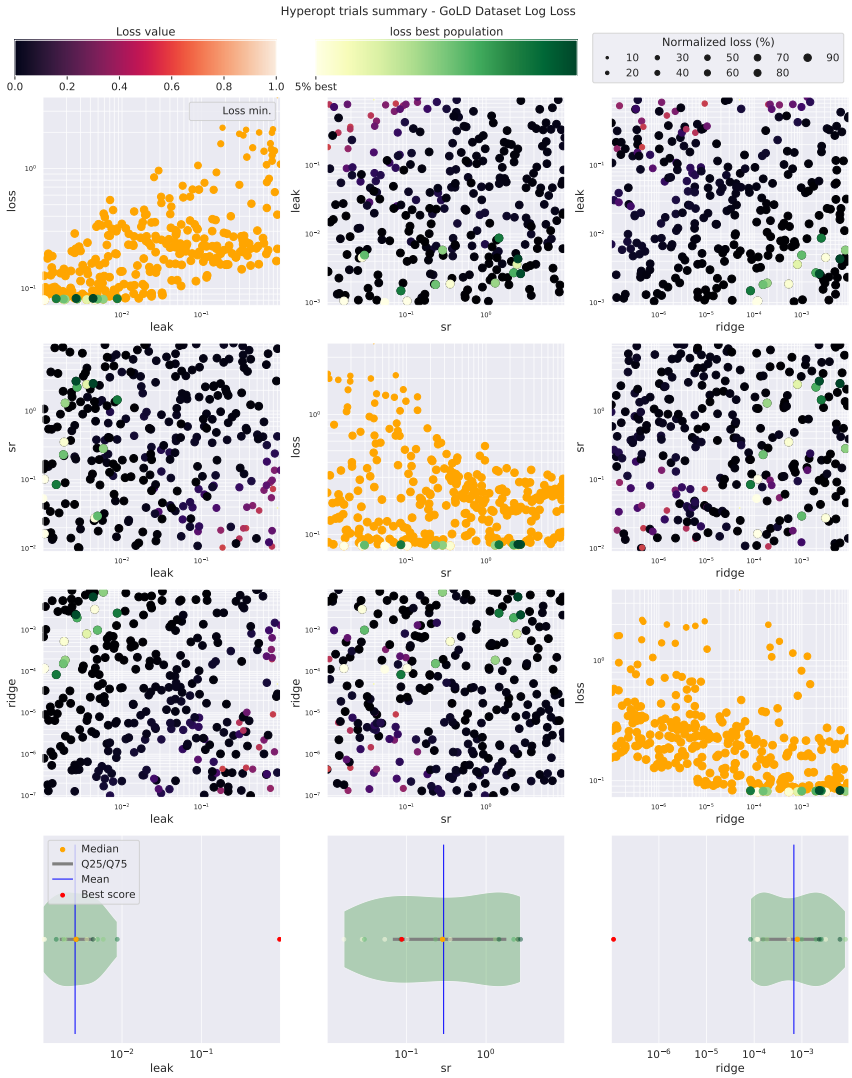



Figure 7.18: GoLD dataset Cross-entropy loss: Hyper-parameter search dependence plot for CSL task.

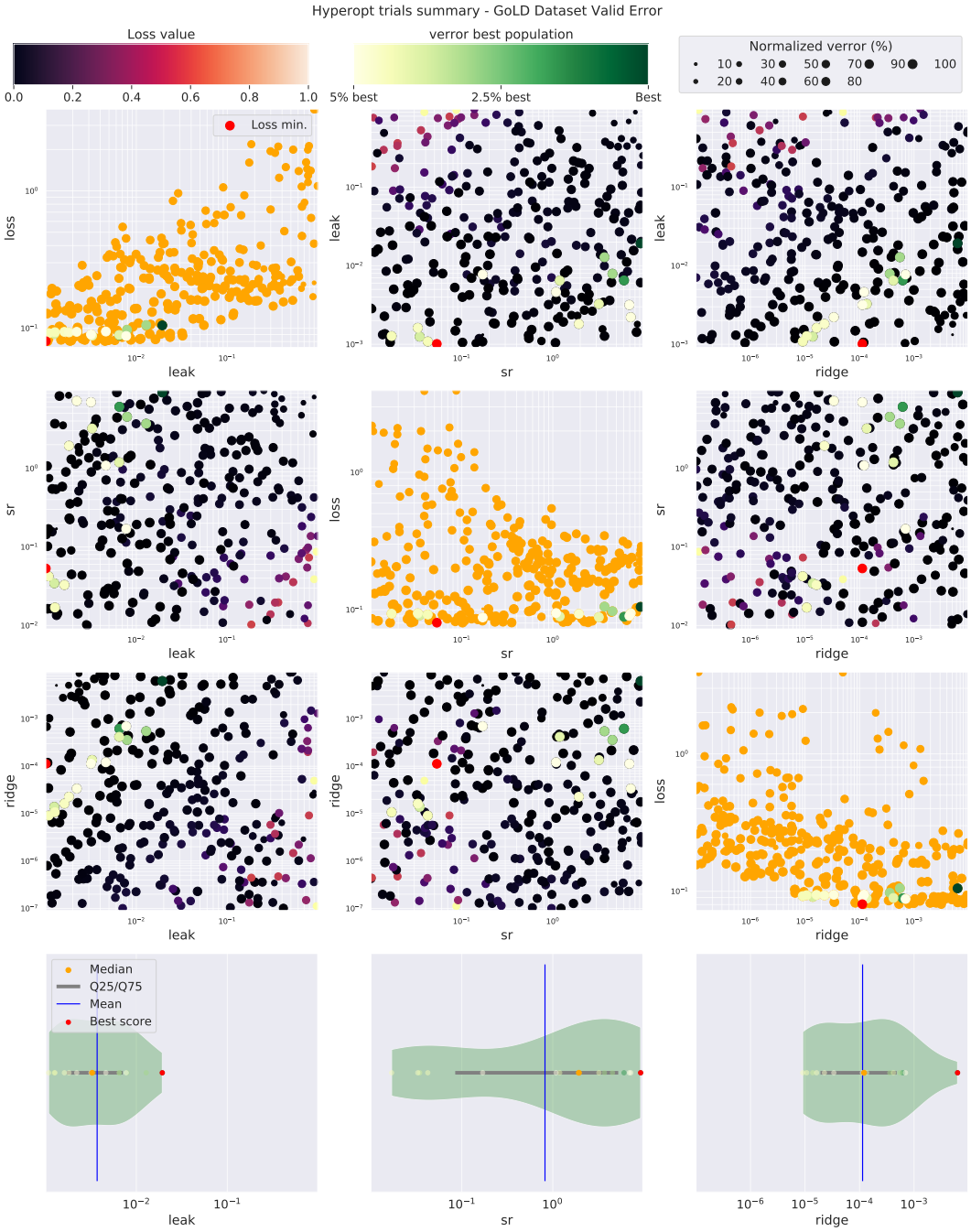



Figure 7.19: GoLD dataset Valid Error: Hyper-parameter search dependence plot for CSL task.

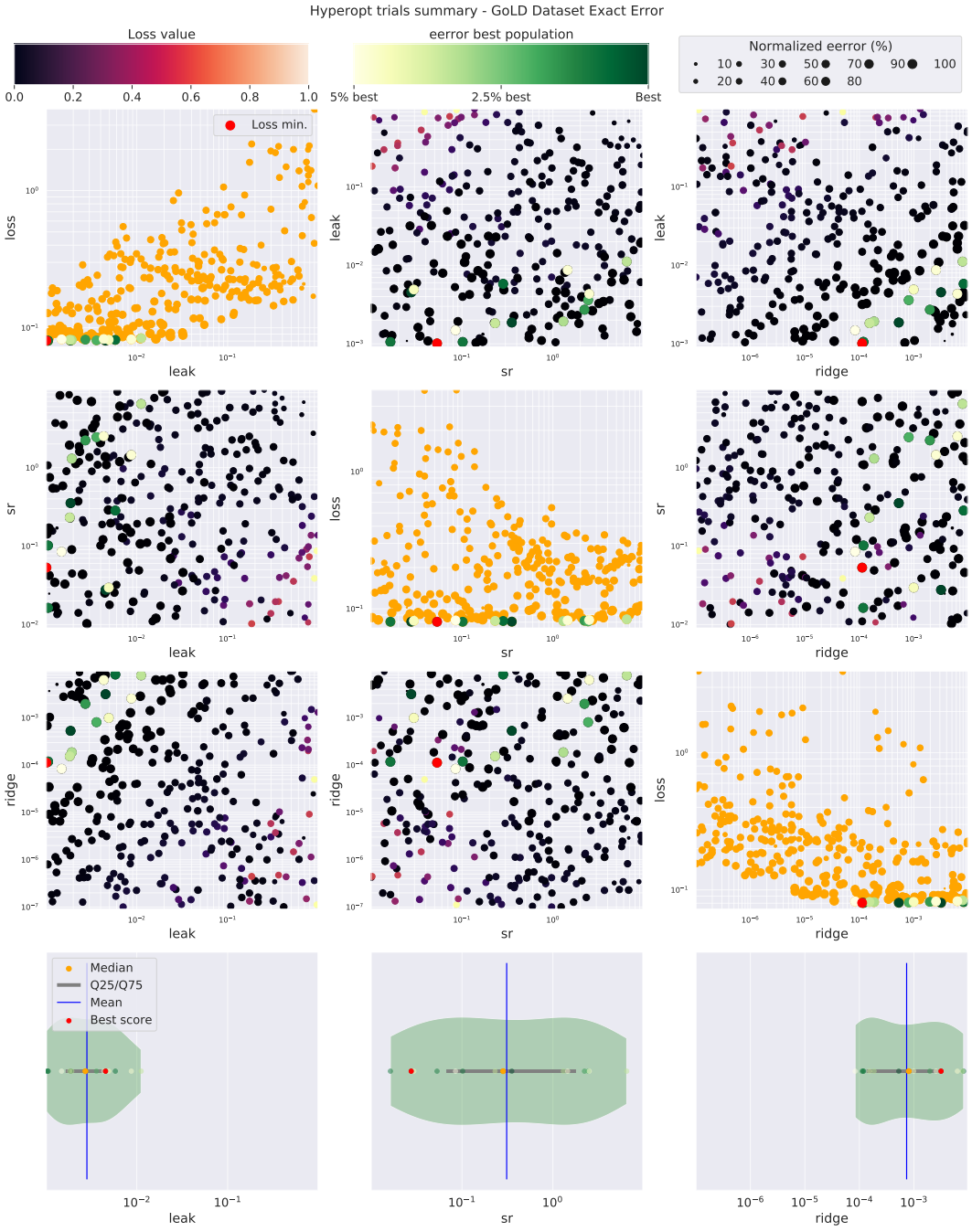



Figure 7.20: GoLD dataset Exact Error: Hyper-parameter search dependence plot for CSL task.

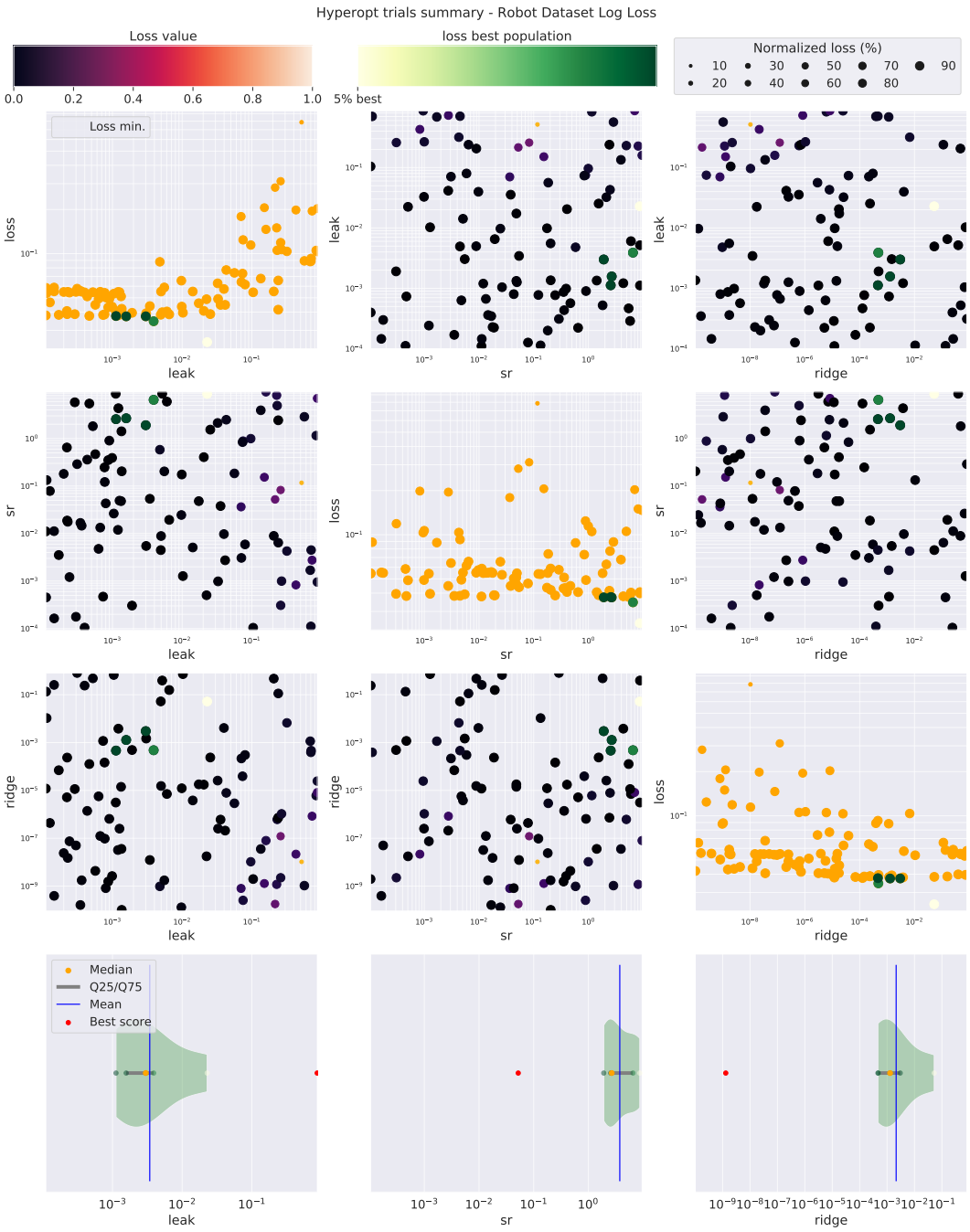



Figure 7.21: Robot dataset Cross-entropy loss: Hyper-parameter search dependence plot for CSL task.

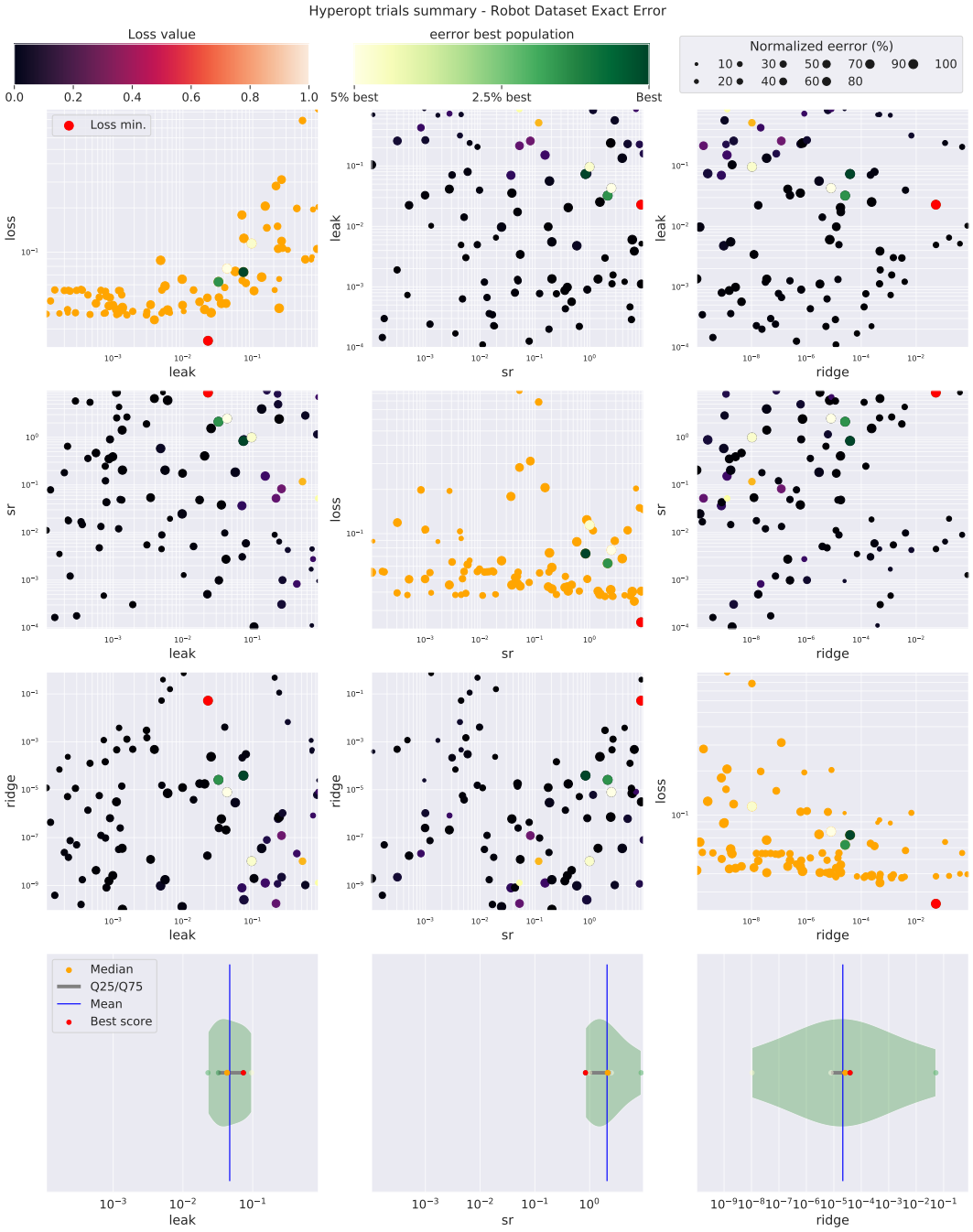



Figure 7.22: Robot dataset Exact Error: Hyper-parameter search dependence plot for CSL task.

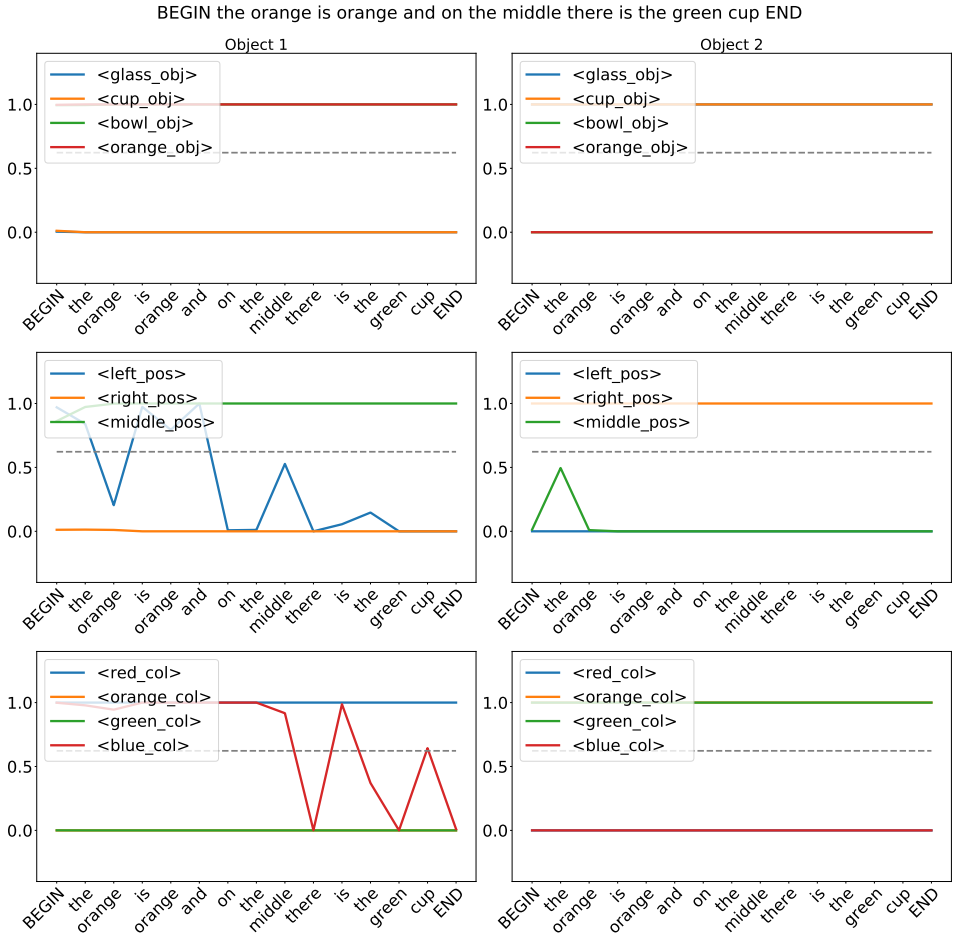



Figure 7.23: Juven's Data: Output activation of the ESN Offline + fine-tuned BERT. The activation are here shown after being transformed by the Sigmoid function.

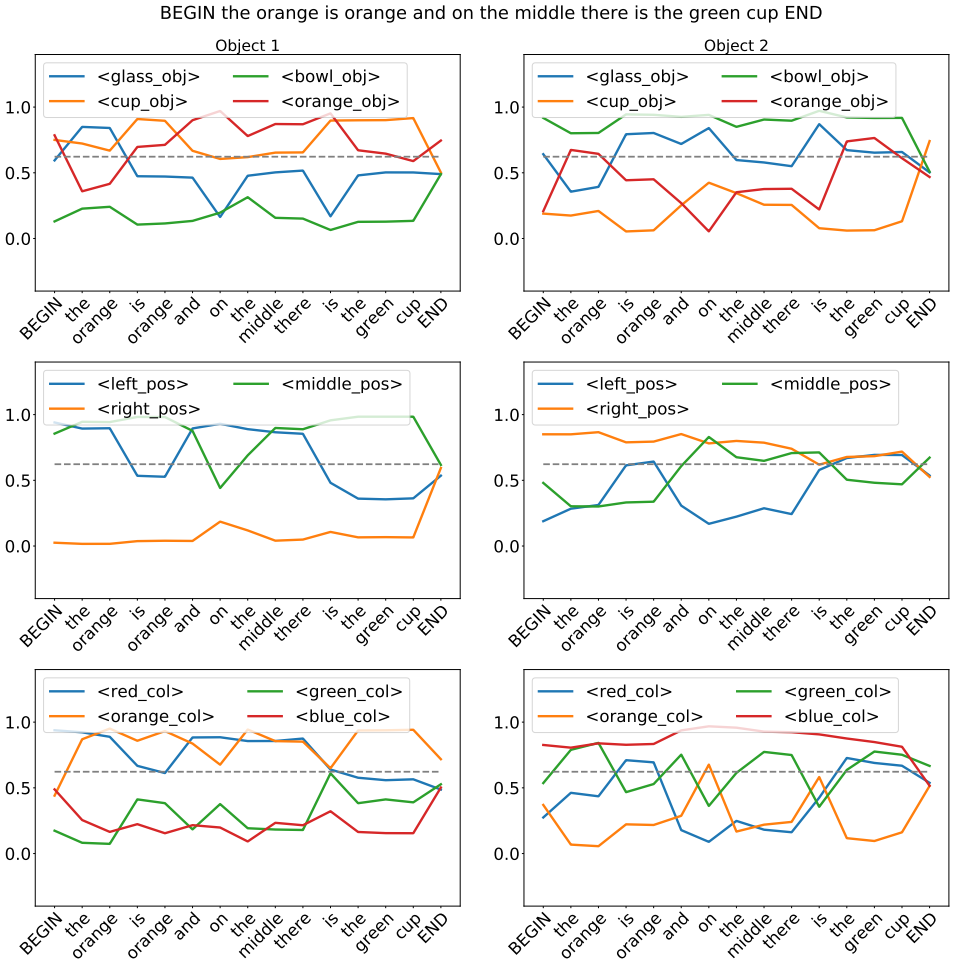



Figure 7.24: Juven's Data: Output activation of the ESN FL + fine-tuned BERT. The activation are here shown after being transformed by the Sigmoid function.

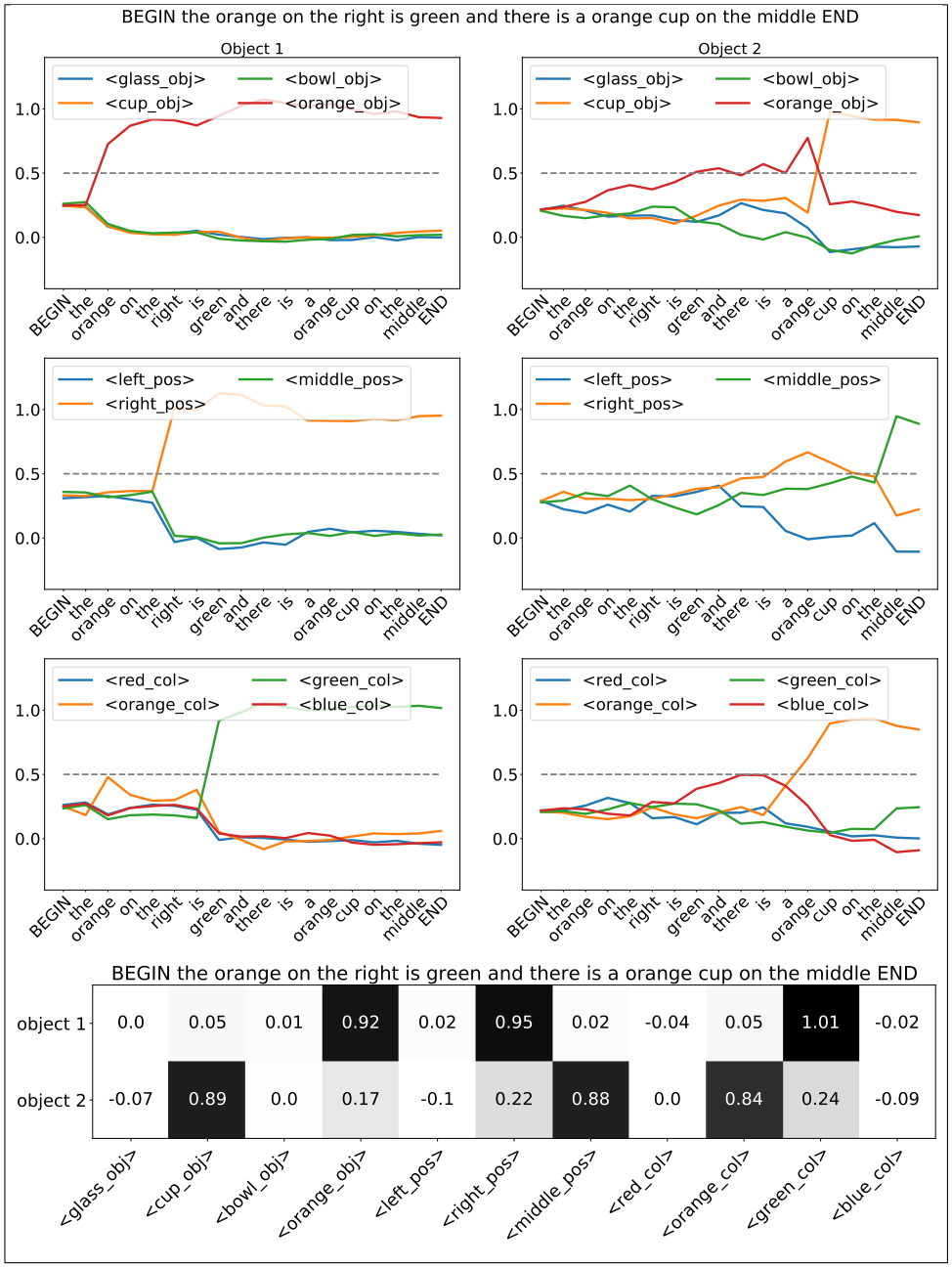



Figure 7.25: Capturing of polysemous words in Juven's Data: Output activation of the ESN CL + fine-tuned BERT. After each word the model tries to predict the correct output. That's why we can see a jump in the correct characteristic after the related keyword is seen.

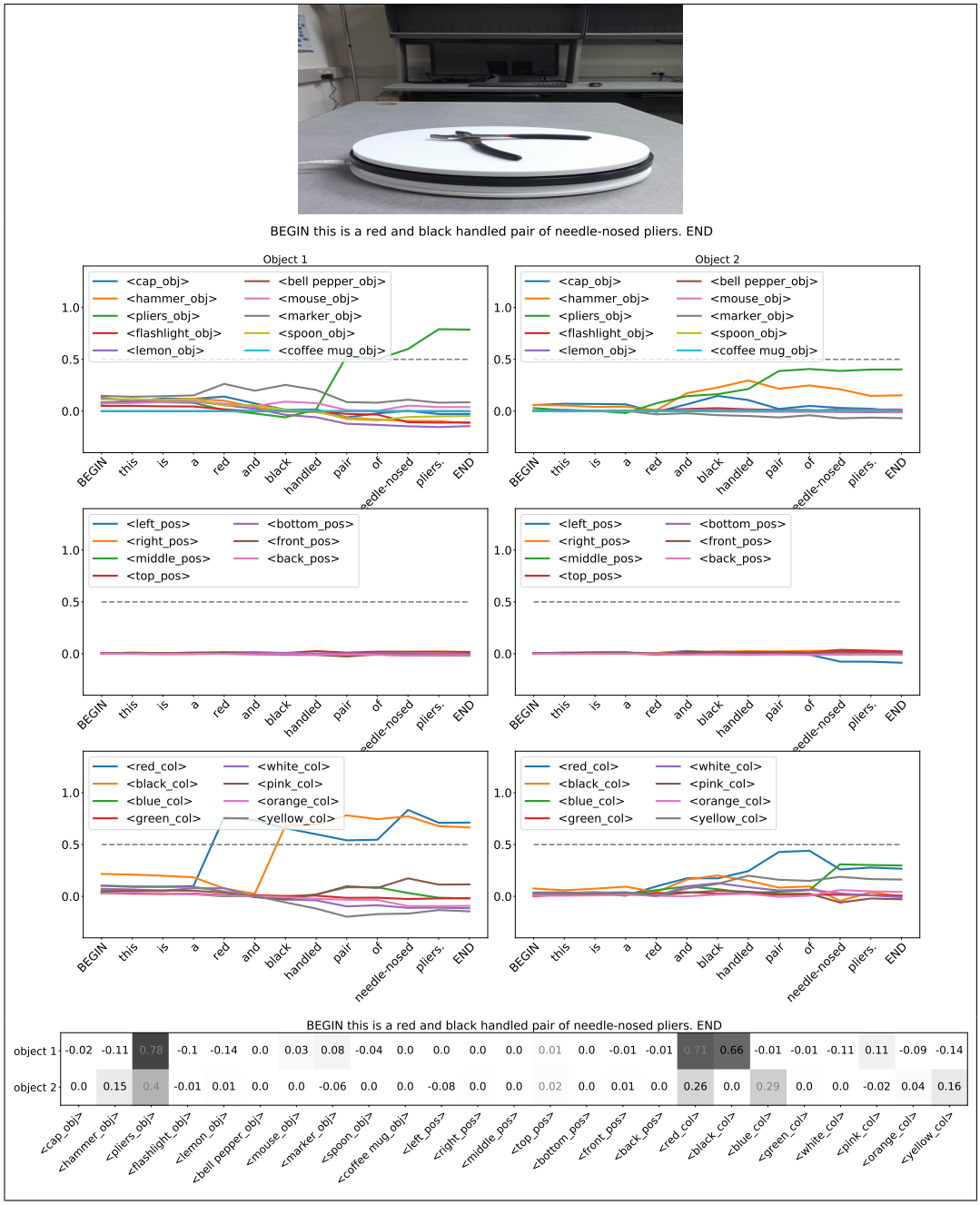



Figure 7.26: GoLD Data: Output activation of the ESN CL + fine-tuned BERT. After each word the model tries to predict the correct output. That's why we can see a jump in the correct characteristic after the related keyword is seen. The top image is source from GoLD dataset Jenkins et al. [2020].

# 8 CONCLUSION

This thesis introduces a data-driven framework bridging the gap between neurolinguistic processing observed in the human brain and the computational mechanisms of natural language processing (NLP) systems. By establishing a direct link between advanced imaging techniques and NLP processes, it conceptualizes brain information processing as a dynamic interplay of three critical components: "what," "where," and "when", offering insights into how the brain interprets language during engagement with naturalistic narratives. This study provides compelling evidence that enhancing the alignment between brain activity and NLP systems offers mutual benefits to the fields of neurolinguistics and NLP. The research showcases how these computational models can emulate the brain's natural language processing capabilities by harnessing cutting-edge neural network technologies across various modalities—language, vision, and speech. Specifically, the thesis highlights how modern pretrained language models achieve closer brain alignment during narrative comprehension. It investigates the differential processing of language across brain regions, the timing of responses (HRF delays), and the balance between syntactic and semantic information processing. Further, the exploration of how different linguistic features align with MEG brain responses over time and find that the alignment depends on the amount of past context, indicating that the brain encodes words slightly behind the current one, awaiting more future context. Furthermore, it highlights grounded language acquisition through noisy supervision and offers a biologically plausible architecture for investigating cross-situational learning, providing interpretability, generalizability, and computational efficiency in sequence-based models. Ultimately, this research contributes valuable insights into neurolinguistics, cognitive neuroscience, and NLP.

## 8.1 Summary of Contributions

### Enhancing Scientific Inference for Encoding Models by Utilizing Extensive Naturalistic Brain Datasets and Progress in Generative AI

The central aim of neuroscience is to unravel how the brain represents information and processes it to carry out various tasks (visual, linguistic, auditory, etc.). Deep neural networks (DNN) offer a computational medium to capture brain activity's unprecedented complexity and richness. *Encoding* and *decoding* stated as computational problems succinctly encapsulate this puzzle. The field is growing rapidly with the availability of large neuroimaging datasets when participants are processing stimuli in naturalistic settings. At the same time, there is tremendous progress in deep neural networks (DNNs) that process multimodal data robustly and efficiently. Drawing inspiration from the effectiveness of recent generative AI models for natural language processing, computer vision, and speech, we review popular deep learning based encoding and decoding architectures and note their benefits and limitations in the context of brain alignment. In Chapter 3, we summarize various encoding models in the form of a taxonomic survey tree. These models cater to vision, auditory, language, and multimodal domains. Given the abundance of recent publications in this area, Chapter 3 aims to facilitate contributions from the computational cognitive neuroscience community, thereby advancing the field of brain encoding and decoding.

### Unveiling the Neural Substrate: Language Models and Long-Term Dependencies in Brain Activation Prediction

Several popular sequence-based and pretrained language models have been found to be successful for text-driven prediction of brain activations [Jain and Huth, 2018, Toneva and Wehbe, 2019]. However, these models still lack long-term memory plausibility (i.e., how they deal with long-term dependencies and contextual information) and insights into the underlying neural substrate mechanisms. Also, the recent pretrained Transformer models like BERT and GPT-2 cannot handle the long-term dependencies (sequence length is fixed to 512 words) due to their self-attention operation. To overcome this limitation, recently, Beltagy et al. [2020] introduced *Longformer*, making it easy to process documents of thousands of tokens or longer and combining local windowed attention with global attention. Considering these challenges, Chapter 4 of this thesis aims to shed light on the relationship between fMRI voxel activations and representations generated by various language models. Our findings suggest that developing language models capable of handling more extensive contextual information and interpreting internal representations within these models can lead to a deeper understanding of how neural structures represent language information and maintain longer narrative memory.

### Unraveling the Interplay of Hemodynamic Response Delays and Language Processing in the Brain

The increasing availability of naturalistic fMRI datasets and large-scale neural models can enable a better understanding of the brain's response to natural stimuli. Just in the last few years, researchers have shown that brain responses of people comprehending language can be predicted well by text-based language models [Wehbe et al., 2014, Jain and Huth, 2018, Toneva and Wehbe, 2019, Deniz et al., 2019, Caucheteux and King, 2020, Schrimpf et al., 2021b, Caucheteux et al., 2021a, Toneva et al., 2022, Oota et al., 2022c, Antonello et al., 2021, Aw and Toneva, 2023, Merlin and Toneva, 2022]. However, existing studies on the alignment between language comprehension and the brain have been observed at constant hemodynamic response function (HRF) delay (around 7.5 to 8 seconds), there is still ongoing exploration into how language and the brain's processing mechanisms synchronize when faced with different HRF delays [Jain and Huth, 2018, Jain et al., 2020, Toneva and Wehbe, 2019, Deniz et al., 2019, Toneva et al., 2022, Aw and Toneva, 2023, Oota et al., 2022c, 2023c]. Further, the existing studies have mainly built brain encoding models by considering a fixed HRF delay and analyzing how different regions of interest (ROIs) involved in language processing influence the semantic and syntactic aspects of information processing in the brain [Jain and Huth, 2018, Jain et al., 2020, Toneva and Wehbe, 2019, Caucheteux et al., 2021a, Toneva et al., 2022, Merlin and Toneva, 2022, Aw and Toneva, 2023, Oota et al., 2022c, 2023c]. In this thesis, we systematically interplay between HRF delays and language processing is an area of investigation, aiming to comprehend how neural activity related to language tasks aligns with the subsequent hemodynamic response, and how this alignment may differ under varying conditions of HRF delays. Our findings suggest that the decomposition of representations into different linguistic features enables a fine-grained understanding of brain language processing across various delays, paving the way for more personalized and effective approaches in both linguistic and clinical applications.

### Exploring the Timing of Linguistic Feature Processing in the Brain with MEG

Over the past decade, Brain-Computer Interface (BCI) helped to make significant progress in understanding language processing in the brain using a popular computational paradigm: Brain encoding, the process aiming to map stimuli features to brain activity. There is a vast literature on linguistic brain encoding for functional MRI (fMRI) related to syntactic and semantic representations. Magnetoencephalography (MEG), with higher temporal resolution than fMRI, enables us to look more precisely at the timing of linguistic feature processing. Unlike MEG decoding, few studies on MEG encoding using natural stimuli exist. Existing ones on story listening focus on phoneme and simple word-based features, ignoring more abstract features such as context, syntactic, and semantic aspects. To understand when the brain processes linguistic structure in sentences, in this thesis, Chapter 5 leverages text representations using basic syntactic features and semantic features, with various context lengths, directions (past vs. future), and within-context relative importance.

### Noisy Supervision in Grounded Language Acquisition: A Language Model Perspective

Grounded language acquisition encompasses the process of acquiring a language, wherein infants learn by observing their surroundings, engaging in interactions with others, and grasping the concepts of a language within the context of the real world [Yu and Ballard, 2004a,b, 2007, Chen and Mooney, 2008, Thomason et al., 2018, Juven and Hinaut, 2020, Vanzo et al., 2020]. However, language acquisition becomes challenging. A single word in an utterance may carry multiple potential meanings, introducing high uncertainty. Traditional approaches to language grounding primarily center around mapping natural language commands to representations, often involving sequences of fundamental robotic actions [Chen and Mooney, 2011, Matuszek et al., 2013, Tellex et al., 2011]. Additionally, current robotic frameworks [Taniguchi et al., 2017, Roesler et al., 2018] do not address how children naturally learn to comprehend complete sentences through cross-situational learning without specific cues. Given these challenges, Chapter 7 of this thesis delves into an investigation of how language models can undertake grounded language acquisition under conditions of noisy supervision. It also explores how these models can account for the dynamics of learning in the brain.

### 8.1.1 NLP →Neurolinguistics

**The Role of Long-Term Context in Brain Encoding: Insights from Language Models** In Chapter 4 of our research, we explore long-term contextual information in language models concerning brain encoding. Using fMRI recordings, we unveil that pretrained models, which incorporate more extensive contextual information, exhibit higher correlation during narrative story listening tasks. Our investigation examines the performance of encoding across different layers within regions of interest (ROIs) associated with language processing in the brain. Our findings indicate that intermediate layers align better with brain activity patterns, highlighting their importance in understanding language comprehension. In LSTM, we observe that cell state representations, responsible for long-term memory, outperform hidden state representations associated with short-term memory. This insight suggests that the internal dynamics of LSTMs may yield more cognitively plausible activations than traditional LSTM activations. This comprehensive investigation was greatly facilitated by leveraging NLP models as model organisms for language comprehension. This approach enabled us to generate contextual numerical representations that offer deeper insights into brain encoding processes.

**Language Model Behavior Across HRF Delays: A Comprehensive Brain Encoding Study** In Chapter 5 of our research, we explored how various language regions in the human brain process word-level syntactic features at different HRF delays. Using fMRI recordings, we observed that word-level syntactic information, including dependency tags (DEP Tags), is notably encoded at early delays (6 secs) in specific regions like IFG and IFGOrb, known for syntactic processing. We also investigated constituent syntactic embeddings, revealing significant encoding of hierarchical syntax information in the MFG region at early delays. Moreover, complex syntax information was found to be encoded in the IFGOrb region.

Additionally, when we examined pretrained language model representations, we discovered that longer context significantly increased HRF delays. For instance, BERT with a context length 20 exhibited higher predictivity within language regions like AG, IFG, ATL, and PTL, particularly for delays ranging from 9 to 12 seconds. These findings suggest that syntax and semantics are distributed across language ROIs.

**Temporal Dynamics of Linguistic Features and Brain Responses: Insights from MEG** Chapter 6 of our research investigated how different linguistic features align with MEG brain responses over time. Pretrained language models outperformed other features in predicting brain alignment, particularly between 50-550ms (or 250ms to 750ms with word onset at 200ms). This alignment depends on the amount of past context, indicating that the brain encodes words slightly before the current one, awaiting more future context. We hypothesize that such "word encoding center of mass" (is a few words before the current word) lying in the past is also what is happening in the speaker's brain, suggesting that past events are retained in memory to disambiguate future events.

**Comparing ESNs and LSTMs: A Study in Computational Efficiency and Performance** Chapter 7 compares the performance of two sequence-based models, ESNs and LSTMs, in learning to parse sentences with noisy supervision (CSL) while examining different word representations. Our findings show that ESNs outperform LSTMs on all three datasets, including one simple and two complex datasets, achieving better prediction accuracy and lower latency. Notably, ESNs demonstrate better generalization than LSTM models, especially when dealing with increasingly large vocabularies, and are more efficient with about a three orders of magnitude reduction in training CPU time. Even when considering the same number of neurons in both models, ESNs outshine RandLSTMs in terms of performance, highlighting the biological plausibility of learning the reservoir states in ESNs. Our study offers three key advantages: interpretability, generalizability, and computational efficiency in sequence-based models.

### 8.1.2 NEUROLINGUISTICS →NLP

**Neurocomputational Perspectives in NLP: Hierarchies, Timing, and Context** In the realm of NLP, Chapter 5 insights suggest a multifaceted research agenda focusing on the nuanced encoding of linguistic information by the brain. Key areas include the exploration of how timing (via variable HRF delays) impacts language processing, the hierarchical nature of linguistic encoding from syntactic to semantic layers in neural models like BERT, and the dynamics of how different types of linguistic information are encoded at various stages. Additionally, the role of context length in language comprehension and its alignment with brain activity presents a promising avenue, particularly in models that leverage both recurrence and self-attention mechanisms.

**Enhancing NLP with Cross-Situational Learning: Insights from Grounded Language Datasets** The cross-situational learning (CSL) paradigm applied to various grounded language datasets shows that fine-tuned BERT representations capture complex word relationships more effectively than other word representations. This strongly indicates that current pretrained language models acquire conceptual meanings through unsupervised or semi-supervised methods. Given the advancements in grounded language models that understand

concepts across both spatial and temporal dimensions, thereby learning the spatio-temporal descriptions of an embodied agent's behavior [Karch et al., 2021], we anticipate the development of more cognitively plausible NLP systems. These systems are expected to benefit significantly from insights into human language processing.

## 8.2 Future Research Directions

These contributions serve as a source of inspiration for various avenues of research, which we anticipate will have a future impact on both neurolinguistics and natural language processing. The following is a summary of the directions that hold a specific interest for the author.

**Exploring the Future of Brain Language Processing with Time-Based Disentanglement for a Deeper Understanding** Recent studies have focused on understanding the complexities of brain language processing, employing techniques to isolate syntactic elements from language models and analyze semantic differences [Reddy and Wehbe, 2021, Caucheteux et al., 2021a, Oota et al., 2023d]. While current research maintains a uniform delay in analysis, future efforts aim to adopt variable delays to enhance comprehension of language processing. This approach intends to dissect language model representations into distinct syntactic and semantic components, such as discourse and emotion, to improve the interpretability of models and provide deeper insights into brain function, suggesting a promising avenue for upcoming research endeavors.

**Comparing Reading and Listening: How Future Words Impact Language Models and Brain Activity** Recent research has indicated an augmented correlation between language model representations and brain activity when exposed to both current and future words, underscoring a connection between brain function and the anticipation of forthcoming words [Caucheteux and King, 2020]. Furthermore, our findings in Chapter 6 emphasize that, in the realm of narrative story listening, the predictive efficacy of past context surpasses that of future context. This raises the question: are there disparities in how the brain processes information during reading compared to listening? Exploring potential distinctions between these two modalities of language comprehension could deepen our understanding of how the human brain adapts its processing strategies based on the mode of information intake, whether through written text or spoken discourse. Such insights have the potential to advance our knowledge in both neurocognitive research and practical applications, including natural language processing and storytelling platforms for enhanced engagement.

**Long Contexts and Language Models: Towards Enhanced fMRI Response Predictions** The performance of language models trained on text, particularly in predicting fMRI responses during reading and listening, has demonstrated its impressiveness, as elaborated in Chapter 4. However, the current level of brain alignment between these models and the human brain does not reach the estimated noise ceiling. Inducing brain-relevant bias can be one way to enhance the alignment of these models with the human brain [Schwartz et al., 2019]. For the advancement of text-based language models, an intriguing question that arises is whether we can elevate the performance of these models by incorporating them with the capability to retain and utilize information from longer contexts.

**Cross-Linguistic Brain Research: Unveiling Language-Dependent Insights** In the scope of this thesis, our research relies on brain recordings collected from individuals who speak English as their primary language. Additionally, we utilize experimental stimuli that are presented in the English language. As a result, our approach predominantly leverages language models and neural models that have been trained extensively on English text data and brain responses elicited by text or speech in English. However, it is essential to acknowledge the potential variability in our study outcomes when extrapolated to languages other than English. The intricate interplay between language-specific nuances and neural responses may introduce distinctions in the results. Therefore, it becomes imperative for future research endeavors to delve into this aspect further and investigate how these factors might influence the generalizability of our findings across diverse linguistic contexts.

**Integrating Information from Multiple Modalities in the Brain** The human brain seamlessly integrates data from various sensory modalities, utilizing its internal memory to influence behavior. Conversely, machine learning models often struggle to create representations that can be applied universally within the same modality, let alone across different modalities. One promising avenue for in-depth brain research involves the investigation of memory mechanisms, specifically understanding how the brain encodes memory representations capable of generalization and retrieves them based on the current sensory context. This represents a long-term research direction with the potential to shed light on fundamental computational processes underlying memory formation and utilization in the brain. On a more immediate note, there is a compelling short-term research opportunity to capitalize on multi-modal brain recording experiments, such as the Courtois NeuroMod dataset [Boyle et al., 2020], where participants watch movies. Instead of analyzing individual modalities separately, this approach advocates modeling all aspects concurrently, including language, non-language auditory cues, and visual stimuli. This shift in focus can offer valuable insights into the holistic functioning of the brain across multiple sensory domains.

# BIBLIOGRAPHY


Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, 2016.

Mostafa Abdou, Ana Valeria González, Mariya Toneva, Daniel Hershcovich, and Anders Søgaard. Does injecting linguistic structure into language models lead to better alignment with brain recordings? *arXiv preprint arXiv:2101.12608*, 2021.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*. International Conference on Learning Representations, ICLR, 2017.

Nicolas Affolter, Beni Egressy, Damian Pascual, and Roger Wattenhofer. Brain2word: Decoding brain activity for language generation. *arXiv preprint arXiv:2009.04765*, 2020.

Nameera Akhtar and Lisa Montague. Early lexical acquisition: The role of cross-situational learning. *First Language*, 19(57):347–358, 1999.

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Mark Allen, William Badecker, and Lee Osterhout. Morphological analysis in sentence processing: An erp study. *Language and Cognitive Processes*, 18(4):405–430, 2003.

Andrew J Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *TACL*, 5:17–30, 2017a.

Andrew James Anderson, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev DS Raizada. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex*, 27(9):4379–4395, 2017b.

Andrew James Anderson, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Rajeev DS Raizada, Feng Lin, and Edmund C Lalor. An integrated neural decoder of linguistic and experiential meaning. *Journal of Neuroscience*, 39(45): 8969–8987, 2019.

Andrew James Anderson, Kelsey McDermott, Brian Rooks, Kathi L Heffner, David Dodell-Feder, and Feng V Lin. Decoding individual identity from brain activity elicited in imagining common experiences. *Nature communications*, 11(1):1–14, 2020.

Richard Antonello, Javier S Turek, Vy Vo, and Alexander Huth. Low-dimensional structure in the space of language representations is reflected in brain responses. *NeurIPS*, 34: 8332–8344, 2021.

Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36, 2024.

Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*, 2017.

Khai Loong Aw and Mariya Toneva. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*, 2023.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 33:12449–12460, 2020.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, pages 1298–1312. PMLR, 2022.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *Universal Language Model Fine-tuning for Text Classification*, 2018.

Cordell M Baker, Joshua D Burks, Robert G Briggs, Andrew K Conner, Chad A Glenn, Kathleen N Taylor, Goksel Sali, Tressie M McCoy, James D Battiste, Daniel L O'Donoghue, et al. A connectomic atlas of the human cerebrum—chapter 7: the lateral parietal lobe. *Operative Neurosurgery*, 15(suppl_1):S295–S349, 2018.

Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W Pillow, Uri Hasson, and Kenneth A Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721, 2017.

Marco Baroni. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307, 2020.

Lawrence W Barsalou. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4): 577–660, 1999.

Lawrence W Barsalou, Ava Santos, W Kyle Simmons, and Christine D Wilson. Language and simulation in conceptual processing. *Symbols, embodiment, and meaning*, pages 245–283, 2008.

Roman Bartusiak, Łukasz Augustyniak, Tomasz Kajdanowicz, Przemysław Kazienko, and Maciej Piasecki. Wordnet2vec: Corpora agnostic word vectorization method. *Neurocomputing*, 326:141–150, 2019.

Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. Multimodal grounding for language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2325–2339, 2018.

Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, 32, 2019.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Douglas K Bemis and Liina Pylkkänen. Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *Journal of Neuroscience*, 31(8):2801–2814, 2011.

Yohann Benchetrit, Hubert Banville, and Jean-Remi King. Brain decoding: toward real-time reconstruction of visual perception. In *The Twelfth International Conference on Learning Representations*, 2023.

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

Julia Berezutskaya, Zachary V Freudenburg, Luca Ambrogioni, Umut Güçlü, Marcel AJ van Gerven, and Nick F Ramsey. Cortical network responses map onto data-driven features that capture visual semantics of movie fragments. *Scientific reports*, 10(1):1–21, 2020.

Shohini Bhattasali, Murielle Fabre, Wen-Ming Luh, Hazem Al Saied, Mathieu Constant, Christophe Pallier, Jonathan R Brennan, R Nathan Spreng, and John Hale. Localising memory retrieval and syntactic composition: an fmri study of naturalistic language comprehension. *Language, Cognition and Neuroscience*, 34(4):491–510, 2019.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

Julie A Boyle, Basile Pinsard, A Boukhdhir, S Belleville, S Bram-batti, J Chen, J Cohen-Adad, A Cyr, A Fuente, P Rainville, et al. The courtois project on neuronal modelling: 2020 data release. In *OHBM*, 2020.

Jonathan R Brennan and John T Hale. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one*, 14(1):e0207741, 2019.

Peter Bright, Helen E Moss, Emmanuel A Stamatakis, and Lorraine K Tyler. Longitudinal studies of semantic dementia: the relationship between structural and functional changes over time. *Neuropsychologia*, 46(8):2177–2188, 2008.

Dean V Buonomano and Michael M Merzenich. Temporal information transformed into a spatial code by a neural network with realistic properties. *Science*, 267(5200):1028–1030, 1995.

Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Computational Biology*, 10(12):e1003963, 2014.

Lu Cao, Dandan Huang, Yue Zhang, Xiaowei Jiang, and Yanan Chen. Brain decoding using fnirs. In *AAAI*, volume 35, pages 12602–12611, 2021.

Alfonso Caramazza and Edgar B Zurif. Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and language*, 3(4):572–582, 1976.

Charlotte Caucheteux and Jean-Rémi King. Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv*, 2020.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Disentangling syntax and semantics in the brain with deep networks. In *International Conference on Machine Learning*, pages 1336–1348. PMLR, 2021a.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. In *EMNLP 2021-Conference on Empirical Methods in Natural Language Processing*, 2021b.

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 for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174, 2018.

Joshua S Cetron, Andrew C Connolly, Solomon G Diamond, Vicki V May, and James V Haxby. Decoding individual differences in stem learning from functional mri data. *Nature communications*, 10(1):1–10, 2019.

Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):1–18, 2019.

Tyler A Chang and Benjamin K Bergen. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16, 2022.

Linda L Chao, James V Haxby, and Alex Martin. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature neuroscience*, 2(10): 913–919, 1999.

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. *Interspeech 2014*, 2014.

David Chen and Raymond Mooney. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th International Conference on Machine Learning*, pages 128–135, 2008.

David Chen and Raymond Mooney. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, 2011.

Xuhang Chen, Baiying Lei, Chi-Man Pun, and Shuqiang Wang. Brain diffuser: An end-to-end brain image to brain network pipeline. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 16–26. Springer, 2023.

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems*, 36, 2024.

Noam Chomsky. Logical structure in language. *Journal of the American Society for Information Science*, 8(4):284, 1957.

Yu-An Chung, Hao Tang, and James Glass. Vector-quantized autoregressive predictive coding. *Interspeech*, pages 3760–3764, 2020.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6 (1):27755, 2016.

Radoslaw Martin Cichy, Gemma Roig, Alex Andonian, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Yalda Mohsenzadeh, Kandan Ramakrishnan, and Aude Oliva. The algonauts project: A platform for communication between the sciences of biological and artificial intelligence. In *2019 Conference on Cognitive Computational Neuroscience*. Cognitive Computational Neuroscience, 2019.

Radoslaw Martin Cichy, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Polina Iamshchinina, M Graumann, A Andonian, NAR Murty, K Kay, Gemma Roig, et al. The algonauts project 2021 challenge: How the human brain makes sense of a world in motion. *arXiv preprint arXiv:2104.13714*, 2021.

Yarden Cohen, Jun Shen, Dawit Semu, Daniel P Leman, William A Liberti III, L Nathan Perkins, Derek C Liberti, Darrell N Kotton, and Timothy J Gardner. Hidden neural states underlie canary song syntax. *Nature*, 582(7813):539–544, 2020.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, 2017.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single &!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, 2018.

R Todd Constable, Kenneth R Pugh, Ella Berroya, W Einar Mencl, Michael Westerveld, Weijia Ni, and Donald Shankweiler. Sentence complexity and input modality effects in sentence comprehension: an fmri study. *NeuroImage*, 22(1):11–21, 2004.

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*, 2023. doi: 10.1101/2022.03.28.485868.

Seana Coulson, Jonathan W King, and Marta Kutas. Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and cognitive processes*, 13(1): 21–58, 1998.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019.

Hanna Damasio, Thomas J Grabowski, Daniel Tranel, Richard D Hichwa, and Antonio R Damasio. A neural basis for lexical retrieval. *Nature*, 380(6574):499–505, 1996.

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. *Journal of Neuroscience*, 37(27):6539–6557, 2017.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, 2023.

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. *Journal of Neuroscience*, 39(39):7722–7736, 2019.

Timo I Denk, Yu Takagi, Takuya Matsuyama, Andrea Agostinelli, Tomoya Nakai, Christian Frank, and Shinji Nishimoto. Brain2music: Reconstructing music from human brain activity. *arXiv preprint arXiv:2307.11078*, 2023.

Rutvik Desai, Usha Tadimeti, and Nicholas Riccardi. Proper and common names in the semantic system, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

Thanh Trung Dinh and Xavier Hinaut. Language acquisition with echo state networks: Towards unsupervised learning. In *2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 1–6. IEEE, 2020.

Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24 (7):431–450, 2023.

Peter Ford Dominey and Jean-David Boucher. Developmental stages of perception and language acquisition in a perceptually grounded robot. *Cognitive Systems Research*, 6 (3):243–259, 2005.

Peter Ford Dominey, Michel Hoen, and Toshio Inui. A neurolinguistic model of grammatical construction processing. *Journal of Cognitive Neuroscience*, 18(12):2088–2107, 2006.

Dota Tianai Dong and Mariya Toneva. Interpreting multimodal video transformers using brain recordings. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023.

Changde Du, Changying Du, Lijie Huang, and Huiguang He. Conditional generative neural decoding with structured cnn feature prediction. In *AAAI*, pages 2629–2636, 2020.

Kshitij Dwivedi, Michael F Bonner, Radoslaw Martin Cichy, and Gemma Roig. Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS computational biology*, 17(8):e1009267, 2021.

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

Jeffrey L Elman. Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, 7:195–225, 1991.

Pierre Enel, Emmanuel Procyk, René Quilodran, and Peter Ford Dominey. Reservoir computing properties of neural dynamics in prefrontal cortex. *PLoS computational biology*, 12(6):e1004967, 2016.

Paola Escudero, Eline A Smit, and Anthony J Angwin. Investigating orthographic versus auditory cross-situational word learning with online and laboratory-based testing. *Language Learning*, 73(2):543–577, 2023.

Tao Fang, Yu Qi, and Gang Pan. Reconstructing perceptive images from brain activity by shape-semantic gan. *NeurIPS*, 33:13038–13048, 2020.

Evelina Fedorenko, Idan Asher Blank, Matthew Siegelman, and Zachary Mineroff. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, 203:104348, 2020.

Graham Flick and Liina Pylkkänen. Isolating syntax in natural language: Meg evidence for an early contribution of left posterior temporal cortex. *Cortex*, 127:42–57, 2020.

Michael C Frank, Noah D Goodman, and Joshua B Tenenbaum. Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, 20(5): 578–585, 2009.

Angela D Friederici. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392, 2011.

Angela D Friederici. The cortical language circuit: from auditory perception to sentence comprehension. *Trends in cognitive sciences*, 16(5):262–268, 2012.

Angela D Friederici, Shirley-Ann Rüschemeyer, Anja Hahne, and Christian J Fiebach. The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. *Cerebral cortex*, 13(2):170–177, 2003.

Angela D Friederici, Christian J Fiebach, Matthias Schlesewsky, Ina D Bornkessel, and D Yves Von Cramon. Processing linguistic complexity and grammaticality in the left frontal cortex. *Cerebral Cortex*, 16(12):1709–1717, 2006.

Kunihiko Fukushima and Sei Miyake. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern recognition*, 15(6):455–469, 1982.

Jianxiong Gao, Yuqian Fu, Yun Wang, Xuelin Qian, Jianfeng Feng, and Yanwei Fu. Mind-3d: Reconstruct high-quality 3d objects in human brain. *arXiv preprint arXiv:2312.07485*, 2023.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew E 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)*, pages 1–6, 2018.

Isabel Gauthier, Thomas W James, Kim M Curby, and Michael J Tarr. The influence of conceptual knowledge on visual discrimination. *Cognitive Neuropsychology*, 20(3-6): 507–523, 2003.

Jon Gauthier and Roger Levy. Linking artificial and human neural representations of language. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Guy Gaziv, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254:119121, 2022.

Christopher R Genovese. A bayesian time-course model for functional magnetic resonance imaging data. *Journal of the American Statistical Association*, 95(451):691–703, 2000.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 1999.

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. Random walks and neural network language models on knowledge bases. In *Proceedings of the 2015 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 1434–1439, 2015.

RF Goldberg, CA Perfetti, and W Schneider. Distinct and common cortical activations for multimodal semantic categories. *Cognitive, Affective, & Behavioral Neuroscience*, 6: 214–222, 2006a.

RF Goldberg, CA Perfetti, and W Schneider. Perceptual knowledge retrieval activates sensory brain regions. *Journal of Neuroscience*, 26(18):4917–4921, 2006b.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and

Matti S. Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13, 2013. doi: 10.3389/fnins.2013.00267.

Alex Graves and Alex Graves. *Supervised sequence labelling*. Springer, 2012.

Mandy Guo, Joshua Ainslie, David C Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. Longt5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, 2022.

Laura Gwilliams, Jean-Remi King, Alec Marantz, and David Poeppel. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature communications*, 13(1):6606, 2022.

Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkänen, David Poeppel, and Jean-Rémi King. Introducing meg-masc a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Scientific Data*, 10(1):862, 2023a.

Laura Gwilliams, Alec Marantz, David Poeppel, and Jean-Remi King. Top-down information shapes lexical processing when listening to continuous speech. *Language, Cognition and Neuroscience*, pages 1–14, 2023b.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. Finding syntax in human encephalography with beam search. In *ACL*, pages 2727–2736, 2018.

James Hampton. Conceptual combination 1. In *Knowledge Concepts and Categories*, pages 133–159. Psychology Press, 2013.

Giacomo Handjaras, Emiliano Ricciardi, Andrea Leo, Alessandro Lenci, Luca Cecchetti, Mirco Cosottini, and Giovanna Marotta. How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge. *Neuroimage*, 135:232–242, 2016.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

Felix Hill, Sona Mokra, Nathaniel Wong, and Tim Harley. Human instruction-following with deep reinforcement learning via transfer-learning from text. *arXiv preprint arXiv:2005.09382*, 2020.

Xavier Hinaut. Which input abstraction is better for a robot syntax acquisition model? phonemes, words or grammatical constructions? In *2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 281–286. IEEE, 2018.

Xavier Hinaut and Peter F Dominey. On-line processing of grammatical structure using reservoir computing. In *Artificial Neural Networks and Machine Learning–ICANN 2012: 22nd International Conference on Artificial Neural Networks, Lausanne, Switzerland, September 11-14, 2012, Proceedings, Part I 22*, pages 596–603. Springer, 2012.

Xavier Hinaut and Peter Ford Dominey. Real-time parallel processing of grammatical structure in the fronto-striatal system: A recurrent network simulation study using reservoir computing. *PLoS ONE*, 8(2):e52946, 2013.

Xavier Hinaut and Nathan Trouvain. Which hype for my new task? hints and random search for reservoir computing hyperparameters. In *ICANN 2021-30th International Conference on Artificial Neural Networks*, 2021.

Xavier Hinaut and Johannes Twiefel. Teach your robot your language! trainable neural parser for modeling human sentence processing: Examples for 15 languages. *IEEE Transactions on Cognitive and Developmental Systems*, 12(2):179–188, 2019.

Xavier Hinaut, Maxime Petit, Gregoire Pointeau, and Peter Ford Dominey. Exploring the acquisition and production of grammatical constructions through human-robot interaction with echo state networks. *Frontiers in Neurorobotics*, 8:16, 2014.

Xavier Hinaut, Johannes Twiefel, Maxime Petit, Peter Dominey, and Stefan Wermter. A recurrent neural network for multiple language acquisition: Starting with english and french. In *Proceedings of the NIPS Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (CoCo 2015)*, 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13, 2018.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. Cognival: A framework for cognitive word embedding evaluation. In *CoNLL*, pages 538–549, 2019.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):1–15, 2017.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *TASLP*, 29:3451–3460, 2021.

David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

Colin Humphries, Jeffrey R Binder, David A Medler, and Einat Liebenthal. Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *Journal of cognitive neuroscience*, 18(4):665–679, 2006.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926, 2018.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

Alexander G Huth, Shinji Nishimoto, An T Vu, Dupre la Tour T, and Gallant JL. Gallant lab natural short clips 3t fmri data. *G-Node*, 2022. URL https://doi.org/10.12751/g-node.vy1zjd.

Anna A Ivanova, Martin Schrimpf, Stefano Anzellotti, Noga Zaslavsky, Evelina Fedorenko, and Leyla Isik. Beyond linear regression: mapping models in cognitive neuroscience should align with research goals. *Neurons, Behavior, Data analysis, and Theory*, 5, 2022. doi: https://doi.org/10.51628/001c.37507.

Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148(34):13, 2001.

Herbert Jaeger. Adaptive nonlinear system identification with echo state networks. *Advances in neural information processing systems*, 15, 2002.

Herbert Jaeger. Echo state network. *scholarpedia*, 2(9):2330, 2007.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In *NIPS*, pages 6629–6638, 2018.

Shailee Jain, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S Turek, and Alexander Huth. Interpretable multi-timescale models for predicting fmri responses to continuous natural speech. *NeurIPS*, 33:13738–13749, 2020.

S Jat, H Tang, P Talukdar, and T Mitchel. Relating simple sentence representations in deep neural networks and the brain. In *ACL*, pages 5137–5154, 2020.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Patrick Jenkins, Rishabh Sachdeva, Gaoussou Youssouf Kebe, Padraig Higgins, Kasra Darvish, Edward Raff, Don Engel, John Winder, Francisco Ferraro, and Cynthia Matuszek. Presentation and analysis of a multimodal dataset for grounded languagelearning. *arXiv preprint arXiv:2007.14987*, 2020.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *EACL 2017*, page 427, 2017.

Marcel Adam Just, Vladimir L Cherkassky, Sandesh Aryal, and Tom M Mitchell. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS one*, 5(1):e8622, 2010.

Alexis Juven and Xavier Hinaut. Cross-situational learning with reservoir computing for language acquisition modelling. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

George Kachergis, Chen Yu, and Richard M Shiffrin. An associative model of adaptive inference for learning word–referent mappings. *Psychonomic bulletin & review*, 19(2): 317–324, 2012.

Antonia Karamolegkou, Mostafa Abdou, and Anders Søgaard. Mapping brains with language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9748–9762, 2023.

Tristan Karch, Laetitia Teodorescu, Katja Hofmann, Clément Moulin-Frier, and Pierre-Yves Oudeyer. Grounding spatio-temporal language with transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and Evelina Fedorenko. Lexical-semantic content, not syntactic structure, is the main contributor to ann-brain similarity of fmri responses in the language network. *Neurobiology of Language*, 5(1):7–42, 2024.

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

David Kemmerer. Word classes in the brain: Implications of linguistic typology for cognitive neuroscience. *Cortex*, 58:27–51, 2014.

David Kemmerer. *Cognitive neuroscience of language*. Routledge, 2022.

Meenakshi Khosla and Leila Wehbe. High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv*, 2022.

DP Kingma. Adam: a method for stochastic optimization. In *ICLR*, 2014.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. *Advances in neural information processing systems*, 28, 2015.

Arnold R Kochari, Ashley G Lewis, Jan-Mathijs Schoffelen, and Herbert Schriefers. Semantic and syntactic composition of minimal adjective-noun phrases in dutch: An meg study. *Neuropsychologia*, 155:107754, 2021.

Naoko Koide-Majima, Shinji Nishimoto, and Kei Majima. Mental image reconstruction from human brain activity: Neural decoding of mental imagery via deep neural network-based bayesian estimation. *Neural Networks*, 170:349–363, 2024.

Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Grounding verbs of motion in natural language commands to robots. In *Experimental robotics: The 12th international symposium on experimental robotics*, pages 31–47. Springer, 2014.

Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1):491, 2022.

Brigitte Krenn, Martin Trapp, Stephanie Gross, and Friedrich Neubarth. Crossmodal cross-situational learning with attention. In *Seventh Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics: Workshop on Computational Models for Crossmodal Learning, Lisbon, Portugal, Sep. IEEE*, 2017.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *NIPS*, 32:12805–12816, 2019.

Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *BioRxiv*, pages 2022–06, 2022.

Gina R Kuperberg. Neural mechanisms of language comprehension: Challenges to syntax. *Brain research*, 1146:23–49, 2007.

Ganit Kupershmidt, Roman Beliy, Guy Gaziv, and Michal Irani. A penny for your (visual) thoughts: Self-supervised reconstruction of natural movies from brain activity. *arXiv preprint arXiv:2206.03544*, 2022.

Marta Kutas and Kara D Federmeier. Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, 62:621–647, 2011.

Marta Kutas, Cyma K Van Petten, and Robert Kluender. Psycholinguistics electrified ii (1994–2005). In *Handbook of psycholinguistics*, pages 659–724. Elsevier, 2006.

Tom Dupré la Tour, Michael Eickenberg, Anwar O Nunez-Elizalde, and Jack L Gallant. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728, 2022.

Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N Apurva Ratan Murty, et al. Bold moments: modeling short visual events through a video fmri dataset and metadata. *bioRxiv*, pages 2023–03, 2023.

Ryan Law and Liina Pylkkänen. Lists with and without syntax: A new approach to measuring the neural processing of syntax. *Journal of Neuroscience*, 41(10):2186–2196, 2021.

Lynn Le, Luca Ambrogioni, Katja Seeliger, Yağmur Güçlütürk, Marcel van Gerven, and Umut Güçlü. Brain2pix: Fully convolutional naturalistic video reconstruction from brain activity. *BioRxiv*, pages 2021–02, 2021.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R Nathan Spreng, Jonathan R Brennan, Yiming Yang, Christophe Pallier, and John Hale. Le petit prince: A multilingual fmri corpus using ecological stimuli. *Scientific Data*, pages 2021–10, 2021.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems*, 35: 29624–29636, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yulong Liu, Yongqiang Ma, Wei Zhou, Guibo Zhu, and Nanning Zheng. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding from fmri. *arXiv preprint arXiv:2302.12971*, 2023.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns llms to the human brain. *arXiv e-prints*, pages arXiv–2312, 2023.

Alessandro Lopopolo, Stefan L Frank, Antal Van den Bosch, and Roel M Willems. Using stochastic language models (slm) to map lexical, syntactic, and phonological information processing in the brain. *PloS one*, 12(5):e0177794, 2017.

Alessandro Lopopolo, Stefan L Frank, Antal Van den Bosch, Annabel Nijhof, and Roel M Willems. The narrative brain dataset (nbd), an fmri dataset for the study of natural language processing in the brain. *Linguistic and Neuro-Cognitive Resources (LiNCR)*, 2018.

Haoyu Lu, Qiongyi Zhou, Nanyi Fei, Zhiwu Lu, Mingyu Ding, Jingyuan Wen, Changde Du, Xin Zhao, Hao Sun, Huiguang He, et al. Multimodal foundation models are better simulators of the human brain. *arXiv preprint arXiv:2208.08263*, 2022.

Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.

Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, 14(11):2531–2560, 2002.

Kyle MacDonald, Daniel Yurovsky, and Michael C Frank. Social cues modulate the representations underlying cross-situational learning. *Cognitive psychology*, 94:67–84, 2017.

Christian K Machens, Ranulfo Romo, and Carlos D Brody. Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. *Journal of Neuroscience*, 30(1): 350–360, 2010.

Alana Marzoev, Samuel Madden, M Frans Kaashoek, Michael Cafarella, and Jacob Andreas. Unnatural language processing: Bridging the gap between synthetic and natural language data. *arXiv preprint arXiv:2004.13645*, 2020.

William Matchin and Gregory Hickok. The cortical organization of syntax. *Cerebral Cortex*, 30(3):1481–1498, 2020.

Takuya Matsuyama, Kota S Sasaki, and Shinji Nishimoto. Applicability of scaling laws to vision encoding models. *arXiv preprint arXiv:2308.00678*, 2023.

Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. Learning to parse natural language commands to a robot control system. In *Experimental Robotics*, pages 403–415. Springer, 2013.

James L McClelland and Nigel H Goddard. Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus*, 6(6):654–665, 1996.

Michael McCloskey. Networks and theories: The place of connectionism in cognitive science. *Psychological science*, 2(6):387–395, 1991.

Bob McMurray, Jessica S Horst, and Larissa K Samuelson. Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological review*, 119(4):831, 2012.

SJ McNaughton. Compensatory plant growth as a response to herbivory. *Oikos*, pages 329–336, 1983.

Tamara Nicol Medina, Jesse Snedeker, John C Trueswell, and Lila R Gleitman. How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22):9014–9019, 2011.

Gabriele Merlin and Mariya Toneva. Language models and brain alignment: beyond word-level semantics and prediction. *arXiv preprint arXiv:2212.00596*, 2022.

Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013b.

Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, Jean-Remi King, et al. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, 35:33428–33443, 2022.

Camille K Milton, Vukshitha Dhanaraj, Isabella M Young, Hugh M Taylor, Peter J Nicholas, Robert G Briggs, Michael Y Bai, Rannulu D Fonseka, Jorge Hormovas, Yueh-Hsin Lin, et al. Parcellation-based anatomic model of the semantic network. *Brain and behavior*, 11(4):e02065, 2021.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, and Robert A Mason. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Hosein Mohebbi, Ali Modarressi, and Mohammad Taher Pilehvar. Exploring the role of bert token representations to explain sentence probing results. In *The 2021 Conference on Empirical Methods in Natural Language Processing*, pages 792–806. Association for Computational Linguistics, 2021.

Gregory L Murphy. Noun phrase interpretation and conceptual combination. *Journal of memory and language*, 29(3):259–288, 1990.

Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. The brain tells a story: Unveiling distinct representations of semantic content in speech, objects, and stories in the human brain with large language models. *bioRxiv*, pages 2024–02, 2024.

Tomoya Nakai, Naoko Koide-Majima, and Shinji Nishimoto. Music genre neuroimaging dataset. *Data in Brief*, 40:107675, 2022.

Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6): 902–915, 2009.

Samuel A Nastase, Yun-Fei Liu, Hanna Hillman, Kenneth A Norman, and Uri Hasson. Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space. *NeuroImage*, 217:116865, 2020a.

Samuel A Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J Honey, Yaara Yeshurun, Mor Regev, et al. Narratives: fmri data for evaluating models of naturalistic language comprehension. preprint. *Neuroscience, December*, pages 2020–06, 2020b.

Lance Edward Nathan. *On the interpretation of concealed questions*. PhD thesis, Massachusetts Institute of Technology, 2006.

Satoshi Nishida and Shinji Nishimoto. Decoding naturalistic experiences from human brain activity via distributed representations of words. *Neuroimage*, 180:232–242, 2018.

Satoshi Nishida, Yusuke Nakano, Antoine Blanc, Naoya Maeda, Masataka Kado, and Shinji Nishimoto. Brain-mediated transfer learning of convolutional neural networks. In *AAAI*, pages 5281–5288, 2020.

Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.

Sam V Norman-Haignere and Josh H McDermott. Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS biology*, 16(12):e2005127, 2018.

Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage*, 197:482–492, 2019.

Tom O'Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi, et al. Kerastuner. https://github.com/keras-team/keras-tuner, 2019.

Subba Reddy Oota, Naresh Manwani, and Raju S Bapi. fMRI Semantic Category Decoding Using Linguistic Encoding of Word Embeddings. In *ICONIP*, pages 3–15. Springer, 2018.

Subba Reddy Oota, Vijay Rowtula, Manish Gupta, and Raju S Bapi. Stepencog: A convolutional lstm autoencoder for near-perfect fmri encoding. In *IJCNN*, pages 1–8. IEEE, 2019.

Subba Reddy Oota, Frédéric Alexandre, and Xavier Hinaut. Cross-situational learning towards robot grounding. 2022a.

Subba Reddy Oota, Frederic Alexandre, and Xavier Hinaut. Long-term plausibility of language models and neural dynamics during narrative listening. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022b.

Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Surampudi. Neural language taskonomy: Which nlp tasks are the most predictive of fmri brain activity? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3220–3237, 2022c.

Subba Reddy Oota, Jashn Arora, Manish Gupta, and Raju S Bapi. Multi-view and cross-view brain decoding. In *COLING*, pages 105–115, 2022d.

Subba Reddy Oota, Jashn Arora, Manish Gupta, Raju Surampudi Bapi, and Mariya Toneva. Deep learning for brain encoding and decoding. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022e.

Subba Reddy Oota, Jashn Arora, Vijay Rowtula, Manish Gupta, and Raju S Bapi. Visio-linguistic brain encoding. In *COLING*, pages 116–133, 2022f.

Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. Speech language models lack important brain-relevant semantics. *arXiv preprint arXiv:2311.04664*, 2023a.

Subba Reddy Oota, Manish Gupta, Raju S Bapi, Gael Jobard, Frédéric Alexandre, and Xavier Hinaut. Deep neural networks and brain alignment: Brain encoding and decoding (survey). *arXiv preprint arXiv:2307.10246*, 2023b.

Subba Reddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *NeurIPS*, 2023c.

Subba Reddy Oota, Mounika Marreddy, Manish Gupta, and Raju Bapi. How does the brain process syntactic structure while listening? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6624–6647, 2023d.

Subba Reddy Oota, Trouvain Nathan, Frederic Alexandre, and Xavier Hinaut. Meg encoding using word context semantics in listening stories. In *Interspeech*, 2023e.

Subba Reddy Oota, Khushbu Pahwa, Mounika Marreddy, Manish Gupta, and Raju Surampudi Bapi. Neural architecture of speech. In *ICASSP*, 2023f.

Subba Reddy Oota, Agarwal Veeral, Marreddy Mounika, Gupta Manish, and Raju Surampudi Bapi. Speech taskonomy: Which speech tasks are the most predictive of fmri brain activity? In *24th INTERSPEECH Conference*, 2023g.

Randall C O'Reilly and Michael J Frank. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, 18(2): 283–328, 2006.

Alexander G Ororbia, Ankur Mali, Matthew A Kelly, and David Reitter. Like a baby: Visually situated neural language acquisition. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 5127–5136. Association for Computational Linguistics (ACL), 2020.

Yohei Oseki and M Asahara. Design of bccwj-eeg: Balanced corpus with human electroencephalography. In *LREC*, pages 189–194, 2020.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

Pankaj Pandey, Gulshan Sharma, Krishna P Miyapuram, Ramanathan Subramanian, and Derek Lomas. Music identification using brain responses to initial snippets. In *ICASSP*, pages 1246–1250, 2022.

Barbara H Partee and Vladimir Borschev. Genitives, relational nouns, and argument-modifier ambiguity. *Modifying adjuncts*, 4:67–112, 2003.

Alexandre Pasquiou, Yair Lakretz, John T Hale, Bertrand Thirion, and Christophe Pallier. Neural language models are not born equal to fit brain data, but training helps. In *International Conference on Machine Learning*, pages 17499–17516. PMLR, 2022.

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. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Luca Pedrelli and Xavier Hinaut. Hierarchical-task reservoir for online semantic analysis from continuous speech. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Francisco Pereira, Matthew Botvinick, and Greg Detre. Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial intelligence*, 194:240–252, 2013.

Francisco Pereira, Bin Lou, Brianna Pritchett, Nancy Kanwisher, Matthew Botvinick, and Ev Fedorenko. Decoding of generic mental representations from functional mri data using word embeddings. *bioRxiv*, page 057216, 2016.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):1–13, 2018.

Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, 2018.

Gorana Pobric, Elizabeth Jefferies, and Matthew A Lambon Ralph. Anterior temporal lobes mediate semantic representation: mimicking semantic dementia by using rtms in normal participants. *Proceedings of the National Academy of Sciences*, 104(50):20137–20141, 2007.

Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature neuroscience*, 24(11): 1628–1636, 2021.

Liina Pylkkänen. Neural basis of basic composition: what we have learned from the red–boat studies and their extensions. *Philosophical Transactions of the Royal Society B*, 375 (1791):20190299, 2020.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21 (140):1–67, 2020.

Matthew A Lambon Ralph, Gorana Pobric, and Elizabeth Jefferies. Conceptual knowledge is underpinned by the temporal pole bilaterally: convergent evidence from rtms. *Cerebral cortex*, 19(4):832–838, 2009.

Okko Räsänen and Heikki Rasilo. A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological review*, 122(4):792, 2015.

Aniketh Janardhan Reddy and Leila Wehbe. Can fmri reveal the representation of syntactic structure in the brain? *NeurIPS*, 34:9843–9856, 2021.

Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013.

Brian Roark. Probabilistic top-down parsing and language modeling. *Computational linguistics*, 27(2):249–276, 2001.

Tanja C Roembke, Matilde E Simonetti, Iring Koch, and Andrea M Philipp. What have we learned from 15 years of research on cross-situational word learning? a focused review. *Frontiers in Psychology*, 14, 2023.

Oliver Roesler, Amir Aly, Tadahiro Taniguchi, and Yoshikatsu Hayashi. A probabilistic framework for comparing syntactic and semantic grounding of synonyms through cross-situational learning. In *ICRA-2018 Workshop on" Representing a Complex World: Perception, Inference, and Learning for Joint Semantic, Geometric, and Physical Understanding"*, 2018.

Corianne Rogalsky and Gregory Hickok. Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. *Cerebral Cortex*, 19(4):786–796, 2009.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8: 842–866, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Alexa Romberg and Chen Yu. Integration and inference: Cross-situational word learning involves more than simple co-occurrences. In *Proceedings of the annual meeting of the cognitive science society*, volume 35, 2013.

Deb K Roy. Learning visually grounded words and syntax for a scene description task. *Computer speech & language*, 16(3-4):353–385, 2002.

Graeme D Ruxton and Guy Beauchamp. Time for some a priori thinking about post hoc testing. *Behavioral ecology*, 19(3):690–693, 2008.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. Statistical learning by 8-month-old infants. *Science*, pages 1926–1928, 1996.

Riitta Salmelin. Clinical neurophysiology of language: the meg approach. *Clinical Neurophysiology*, 118(2):237–254, 2007.

Gabriel Sarch, Michael Tarr, Katerina Fragkiadaki, and Leila Wehbe. Brain dissection: fmri-trained networks reveal spatial selectivity in the processing of natural images. *Advances in Neural Information Processing Systems*, 36, 2024.

Robert Scholz, Arno Villringer, and Mauricio JD Martins. Distinct hippocampal and cortical contributions in the representation of hierarchies. *bioRxiv*, pages 2022–06, 2022.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2020.

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *PNAS*, Vol:To appear, 2021a.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *PNAS*, 118(45), 2021b.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. Inducing brain-relevant bias in natural language processing models. *NIPS*, 32:14123–14133, 2019.

Paul S Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *arXiv preprint arXiv:2305.18274*, 2023.

K Seeliger, RP Sommers, Umut Güçlü, Sander E Bosch, and MAJ Van Gerven. A large single-participant fmri dataset for probing brain responses to naturalistic stimuli in space and time. *bioRxiv*, page 687681, 2019.

Jyun Senda, Mai Tanaka, Keiya Iijima, Masato Sugino, Fumina Mori, Yasuhiko Jimbo, Masaki Iwasaki, and Kiyoshi Kotani. Auditory stimulus reconstruction from ecog with dnn and self-attention modules. *Biomedical Signal Processing and Control*, 89:105761, 2024.

Elisabet Service, Päivi Helenius, Sini Maury, and Riitta Salmelin. Localization of syntactic and semantic brain responses using magnetoencephalography. *Journal of Cognitive Neuroscience*, 19(7):1193–1205, 2007.

Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*, 2019.

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. *Neuropsychologia*, 45(12):2802–2810, 2007.

K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

Chandan Singh, Aliyah R Hsu, Richard Antonello, Shailee Jain, Alexander Huth, Bin Yu, and Jianfeng Gao. Explaining black box text modules in natural language with language models. *arXiv preprint arXiv:2305.09863*, 2023.

Vishwajeet Singh, Krishna P. Miyapuram, and Raju S. Bapi. Detection of cognitive states from fmri data using machine learning techniques. In Manuela M. Veloso, editor, *IJCAI*, pages 587–592, 2007.

Jonathan Smallwood and Jonathan W Schooler. The science of mind wandering: empirically navigating the stream of consciousness. *Annual review of psychology*, 66:487–518, 2015.

Edward E Smith and Daniel N Osherson. Conceptual combination with prototype concepts. *Cognitive science*, 8(4):337–361, 1984.

Kerri Smith. Reading minds. *Nature*, 502(7472):428, 2013.

Linda Smith and Chen Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. In *The Twelfth International Conference on Learning Representations*, 2023.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Jon Scott Stevens, Lila R Gleitman, John C Trueswell, and Charles Yang. The pursuit of word meanings. *Cognitive science*, 41:638–676, 2017.

Anthony Strock, Xavier Hinaut, and Nicolas P Rougier. A robust model of gated working memory. *Neural computation*, 32(1):153–181, 2020.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*, 2018.

Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62(1):451–463, 2012.

Jingyuan Sun and Marie-Francine Moens. Fine-tuned vs. prompt-tuned supervised representations: which better account for brain language representations? In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5197–5205, 2023.

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Towards sentence-level brain decoding with distributed representations. In *AAAI*, pages 7047–7054, 2019.

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Neural encoding and decoding with distributed sentence representations. *IEEE TNNLS*, 32(2):589–603, 2020.

Jingyuan Sun, Xiaohan Zhang, and Marie-Francine Moens. Tuning in to neural encoding: Linking human brain and artificial supervised representations of language. In *ECAI 2023*, pages 2258–2265. IOS Press, 2023.

David Sussillo and Larry F Abbott. Generating coherent patterns of activity from chaotic neural networks. *Neuron*, 63(4):544–557, 2009.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023a.

Yu Takagi and Shinji Nishimoto. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. *arXiv preprint arXiv:2306.11536*, 2023b.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. 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)*, pages 5100–5111, 2019.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5): 858–866, 2023.

Jerry Tang, Meng Du, Vy Vo, Vasudev Lal, and Alexander Huth. Brain encoding models based on multimodal transformers can transfer across language and vision. *Advances in Neural Information Processing Systems*, 36, 2024.

Akira Taniguchi, Tadahiro Taniguchi, and Angelo Cangelosi. Cross-situational learning with bayesian generative models for multimodal category and word learning in robots. *Frontiers in Neurorobotics*, 11:66, 2017.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, 2011.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*, 2018.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, 2019.

Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4):1104–1116, 2006.

Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. Learning multi-modal grounded linguistic semantics by playing" i spy". In *IJCAI*, pages 3477–3483, 2016.

Jesse Thomason, Jivko Sinapov, Raymond Mooney, and Peter Stone. Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Andrea Lockerd Thomaz, Cynthia Breazeal, et al. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *Aaai*, volume 6, pages 1000–1005. Boston, MA, 2006.

Andreĭ Nikolaevich Tikhonov, Vasilij Ja Arsenin, and Vasiliĭ Arsenin. *Solutions of ill-posed problems*. V H Winston, 1977.

Michael Tomasello. *Constructing a language*. Harvard university press, 2009.

Mariya Toneva. *Bridging Language in Machines with Language in the Brain*. PhD thesis, Carnegie Mellon University, 2021.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, 32, 2019.

Mariya Toneva, Otilia Stretcu, Barnabás Póczos, Leila Wehbe, and Tom M Mitchell. Modeling task effects on meaning representation in the brain via zero-shot meg prediction. *NIPS*, 33, 2020.

Mariya Toneva, Jennifer Williams, Anand B, Christoph Dann, and Leila Wehbe. Same cause; different effects in the brain. In *CLeaR*, 2021.

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2 (11):745–757, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

John C Trueswell, Tamara Nicol Medina, Alon Hafri, and Lila R Gleitman. Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, 66 (1):126–156, 2013.

Natalie M Trumpp, Daniel Kliese, Klaus Hoenig, Thomas Haarmeier, and Markus Kiefer. Losing the sound of concepts: Damage to auditory association cortex impairs the processing of sound-related concepts. *Cortex*, 49(2):474–486, 2013.

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, 21(12):e3002366, 2023.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, pages 1–18, 2024.

Johannes Twiefel. *Robust Bidirectional Processing for Speech-controlled Robotic Scenarios*. PhD thesis, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky, 2020.

Johannes Twiefel, Xavier Hinaut, and Stefan Wermter. Semantic role labelling for robot instructions using echo state networks. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2016.

Aditya Vaidya, Shailee Jain, and Alexander Huth. Self-supervised models of audio effectively explain human cortical responses to speech. In *International Conference on Machine Learning*, pages 21927–21944. PMLR, 2022.

Marcel Van Gerven, Jason Farquhar, Rebecca Schaefer, Rutger Vlek, Jeroen Geuze, Anton Nijholt, Nick Ramsey, Pim Haselager, Louis Vuurpijl, Stan Gielen, et al. The brain–computer interface cycle. *Journal of neural engineering*, 6(4):041001, 2009.

Andrea Vanzo, Danilo Croce, Emanuele Bastianelli, Roberto Basili, and Daniele Nardi. Grounded language interpretation of robotic commands through structured learning. *Artificial Intelligence*, 278:103181, 2020.

Vladimir Vapnik, Esther Levin, and Yann Le Cun. Measuring the vc-dimension of a learning machine. *Neural computation*, 6(5):851–876, 1994.

Alexandre Variengien and Xavier Hinaut. A journey in esn and lstm visualisations on a language task. *arXiv preprint arXiv:2012.01748*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

Jonathan H Venezia, Steven M Thurman, Virginia M Richards, and Gregory Hickok. Hierarchy of speech-driven spectrotemporal receptive fields in human auditory cortex. *Neuroimage*, 186:647–666, 2019.

Maya Visser, Elizabeth Jefferies, and MA Lambon Ralph. Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. *Journal of cognitive neuroscience*, 22(6):1083–1094, 2010.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.

Aria Wang, Michael Tarr, and Leila Wehbe. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *NeurIPS*, 32:15501–15511, 2019.

Aria Y Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. *BioRxiv*, pages 2022–09, 2022.

Jing Wang, Vladimir L Cherkassky, and M Adam Just. Predicting the brain activation pattern associated with the propositional content of a sentence: Modeling neural representations of events and states. *HBM*, 10:4865–4881, 2017.

Shaonan Wang, Jiajun Zhang, Nan Lin, and Chengqing Zong. Probing brain activation patterns by dissociating semantics and syntax in sentences. In *AAAI*, volume 34, pages 9201–9208, 2020a.

Shaonan Wang, Jiajun Zhang, Haiyan Wang, Nan Lin, and Chengqing Zong. Fine-grained neural decoding with distributed word representations. *Information Sciences*, 507:256–272, 2020b.

David E Warren, Tanja C Roembke, Natalie V Covington, Bob McMurray, and Melissa C Duff. Cross-situational statistical learning of new words despite bilateral hippocampal damage and severe amnesia. *Frontiers in Human Neuroscience*, 13:448, 2020.

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

Masha Westerlund and Liina Pylkkänen. The role of the left anterior temporal lobe in semantic composition vs. semantic memory. *Neuropsychologia*, 57:59–70, 2014.

John Wieting and Douwe Kiela. No training required: Exploring random encoders for sentence classification. In *International Conference on Learning Representations*, 2018.

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. *PNAS*, 111(23):8619–8624, 2014.

Chen Yu and Dana H Ballard. On the integration of grounding language and learning objects. In *AAAI*, volume 4, pages 488–493, 2004a.

Chen Yu and Dana H Ballard. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception (TAP)*, 1(1): 57–80, 2004b.

Chen Yu and Dana H Ballard. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165, 2007.

Chen Yu and Linda B Smith. Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, 18(5):414–420, 2007.

Daniel Yurovsky and Michael C Frank. An integrative account of constraints on cross-situational learning. *Cognition*, 145:53–62, 2015.

Emiliano Zaccarella and Angela D Friederici. Merge in the human brain: A sub-region based functional investigation in the left pars opercularis. *Frontiers in psychology*, 6: 1818, 2015.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

Linmin Zhang and Liina Pylkkänen. The interplay of composition and concept specificity in the left anterior temporal lobe: An meg study. *NeuroImage*, 111:228–240, 2015.

Xiaohan Zhang, Shaonan Wang, Nan Lin, Jiajun Zhang, and Chengqing Zong. Probing word syntactic representations in the brain by a feature elimination method. *AAAI*, 2022a.

Xiaohan Zhang, Shaonan Wang, Nan Lin, and Chengqing Zong. Is the brain mechanism for hierarchical structure building universal across languages? an fmri study of chinese and english. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7852–7861, 2022b.

Yizhen Zhang, Kuan Han, Robert Worth, and Zhongming Liu. Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature communications*, 11(1):1–13, 2020a.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635, 2020b.

Junpei Zhong, Angelo Cangelosi, and Tetsuya Ogata. Toward abstraction from multi-modal data: empirical studies on multiple time-scale recurrent models. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3625–3632. IEEE, 2017.

Jayden Ziegler and Liina Pylkkänen. Scalar adjectives and the temporal unfolding of semantic composition: An meg investigation. *Neuropsychologia*, 89:161–171, 2016.

Benjamin D Zinszer, Laurie Bayet, Lauren L Emberson, Rajeev DS Raizada, and Richard N Aslin. Decoding semantic representations from functional near-infrared spectroscopy signals. *Neurophotonics*, 5(1):011003–011003, 2018.