Deciphering the neural bases of language comprehension using latent linguistic representations

Alexandre Pasquiou

université PARIS-SACLAY

Deciphering the Neural Bases of Language Comprehension Using Latent Linguistic Representations

Déchiffrer les bases neurales de la compréhension du langage à l'aide de représentations linguistiques latentes

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580, Sciences et Technologies de l'Information et de la Communication (ED STIC)
Spécialité de doctorat: Informatique
Graduate School : Informatique et sciences du numérique. Référent : Faculté des sciences d'Orsay

Thèse préparée dans les unités de recherche **Inria Saclay Île-de-France (Université Paris-Saclay, Inria), Neuroimagerie cognitive (Université Paris-Saclay, Inserm, CEA).**, sous la direction de **Bertrand THIRION**, directeur de recherche, le co-encadrement de **Christophe PALLIER**, directeur de recherche.

Thèse soutenue à Paris-Saclay, le 15 Juin 2023, par

Alexandre PASQUIOU

Composition du jury

Membres du jury avec voix délibérative

Pierre ZWEIGENBAUM Directeur de recherche, Université Paris-Saclay, CNRS, LISN	Président	
Alexander HUTH Professeur, University of Texas	Rapporteur &	Examinateur
Frédéric ALEXANDRE Directeur de recherche, Inria, Bordeaux	Rapporteur &	Examinateur
Claire NEDELLEC Directrice de recherche, Université Paris-Saclay, INRAE	Examinatrice	
Leila WEHBE Professeure, Carnegie Mellon University	Examinatrice	
Yair LAKRETZ Chargé de recherche, ENS ULM	Examinateur	

THESE DE DOCTORAT

NNT : 2023UPASG041

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Déchiffrer les bases neurales de la compréhension du langage à l'aide de représentations linguistiques latentes

Mots clés: IRMf, Transformers, Traitement du Langage Naturel, Modèles linguistiques neuronaux, Modèles d'encodage, Apprentissage Profond

Résumé: Au cours des dernières décennies, les modèles de langage (MLs) ont atteint des performances équivalentes à celles de l'homme sur plusieurs tâches. Ces modèles peuvent générer des représentations vectorielles qui capturent diverses propriétés linguistiques des mots d'un texte, telles que la sémantique ou la syntaxe. Les neuroscientifiques ont donc mis à profit ces progrès et ont commencé à utiliser ces modèles pour explorer les bases neurales de la compréhension du langage. Plus précisément, les représentations des ML calculées à partir d'une histoire sont utilisées pour modéliser les données cérébrales d'humains écoutant la même histoire, ce qui permet l'examen de plusieurs niveaux de traitement du langage dans le cerveau. Si les représentations du ML s'alignent étroitement avec une région cérébrale, il est probable que le modèle et la région codent la même information.

En utilisant les données cérébrales d'IRMf de participants américains écoutant l'histoire du Petit Prince, cette thèse 1) examine les facteurs influant l'alignement entre les représentations des MLs et celles du cerveau, ainsi que 2) les limites de telles alignements. La comparaison de plusieurs MLs pré-entraînés et personnalisés (GloVe, LSTM, GPT-2 et BERT) a révélé que les Transformers s'alignent mieux aux données d'IRMf que LSTM et GloVe. Cependant, aucun d'entre eux n'est capable d'expliquer tout le signal IRMf, suggérant des limites liées au paradigme d'encodage ou aux MLs. En étudiant l'architecture des Transformers, nous avons constaté qu'aucune région cérébrale n'est mieux expliquée par une couche ou une tête d'attention spécifique. Nos résultats montrent que la nature et la quantité de données d'entraînement affectent l'alignement. Ainsi, les modèles pré-entraînés sur de petits ensembles de données ne sont pas efficaces pour capturer les activations cérébrales. Nous avons aussi montré que l'entraînement des MLs influence leur capacité à s'aligner aux données IRMf et que la perplexité n'est pas un bon prédicteur de leur capacité à s'aligner. Cependant, entraîner les MLs améliore particulièrement leur performance d'alignement dans les régions coeur de la sémantique, indépendamment de l'architecture et des données d'entraînement. Nous avons également montré que les représentations du cerveau et des MLs convergent d'abord pendant l'entraînement du modèle avant de diverger l'une de l'autre.

Cette thèse examine en outre les bases neurales de la syntaxe, de la sémantique et de la sensibilité au contexte en développant une méthode qui peut sonder des dimensions linguistiques spécifiques. Cette méthode utilise des MLs restreints en information, c'est-à-dire des architectures entraînées sur des espaces de représentations contenant un type spécifique d'information. Tout d'abord, l'entraînement de MLs sur des représentations sémantiques et syntaxiques a révélé un bon alignement dans la plupart du cortex mais avec des degrés relatifs variables. La quantification de cette sensibilité relative à la syntaxe et à la sémantique a montré que les régions cérébrales les plus sensibles à la syntaxe sont plus localisées, contrairement au traitement de la sémantique qui reste largement distribué dans le cortex. Une découverte notable de cette thèse est que l'étendue des régions cérébrales sensibles à la syntaxe et à la sémantique est similaire dans les deux hémisphères. Cependant, l'hémisphère gauche a une plus grande tendance à distinguer le traitement syntaxique et sémantique par rapport à l'hémisphère droit.

Dans un dernier ensemble d'expériences, nous avons conçu une méthode qui contrôle les mécanismes d'attention dans les Transformers afin de générer des représentations qui utilisent un contexte de taille fixe. Cette approche fournit des preuves de la sensibilité au contexte dans la plupart du cortex. De plus, cette analyse a révélé que les hémisphères gauche et droit avaient tendance à traiter respectivement des informations contextuelles plus courtes et plus longues.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title: Deciphering the Neural Bases of Language Comprehension Using Latent Linguistics Representations
Keywords: fMRI, Transformers, Natural Language Processing, Neural Language Models, Encoding Models, Deep Learning

Abstract: In the last decades, language models (LMs) have reached human level performance on several tasks. They can generate rich representations (features) that capture various linguistic properties such as semantics or syntax. Following these improvements, neuroscientists have increasingly used them to explore the neural bases of language comprehension. Specifically, LM's features computed from a story are used to fit the brain data of humans listening to the same story, allowing the examination of multiple levels of language processing in the brain. If LM's features closely align with a specific brain region, then it suggests that both the model and the region are encoding the same information. LM-brain comparisons can then teach us about language processing in the brain.

Using the fMRI brain data of fifty US participants listening to *The Little Prince* story, this thesis 1) investigates the reasons why LMs' features fit brain activity and 2) examines the limitations of such comparisons. The comparison of several pre-trained and custom-trained LMs (GloVe, LSTM, GPT-2 and BERT) revealed that Transformers better fit fMRI brain data than LSTM and GloVe. Yet, none are able to explain all the fMRI signal, suggesting either limitations related to the encoding paradigm or to the LMs. Focusing specifically on Transformers, we found that no brain region is better fitted by specific attentional head or layer. Our results caution that the nature and the amount of training data greatly affects the outcome, indicating that using off-the-shelf models trained on small datasets is not effective in capturing brain activations. We showed that LMs' training influences their ability to fit fMRI brain data, and that perplexity was not a good predictor of brain score. Still, training LMs particularly improves their fitting performance in core semantic regions, irrespective of the architecture and training data. Moreover, we showed a partial convergence between brain's and LM's representations. Specifically, they first converge during model training before diverging from one another.

This thesis further investigates the neural bases of syntax, semantics and context-sensitivity by developing a method that can probe specific linguistic dimensions. This method makes use of *information-restricted LMs*, that are customized LMs architectures trained on feature spaces containing a specific type of information, in order to fit brain data. First, training LMs on semantic and syntactic features revealed a good fitting performance in a widespread network, albeit with varying relative degrees. The quantification of this relative sensitivity to syntax and semantics showed that brain regions most attuned to syntax tend to be more localized, while semantic processing remain widely distributed over the cortex. One notable finding from this analysis was that the extent of semantic and syntactic sensitive brain regions was similar across hemispheres. However, the left hemisphere had a greater tendency to distinguish between syntactic and semantic processing compared to the right hemisphere.

In a last set of experiments we designed *masked-attention generation*, a method that controls the attention mechanisms in transformers, in order to generate latent representations that leverage fixed-size context. This approach provides evidence of context-sensitivity across most of the cortex. Moreover, this analysis found that the left and right hemispheres tend to process shorter and longer contextual information respectively.

Acknowledgments

I would like to thank Frédéric Alexandre and Alexander Huth, as well as Claire Nédellec, Leila Wehbe, Pierre Zweigenbaum and Yair Lakretz for accepting to review this manuscript and be part of my defense jury.

First, I want to thank Stanislas Dehaene for creating the spark that led me to this thesis, with his courses at Collège de France. Thanks to Christophe Pallier who introduced me to the wonders of Neurosciences, and with whom I have worked for the last 4 years. Thanks to Yair who helped me make my first steps in research, and with whom I have worked on many projects. Thanks to Bertrand, who made this thesis possible, bringing salvation (fundings) when all hope was lost. Thanks again to Christophe and Bertrand without whom this thesis would not have been possible, for their dedicated and kind guidance, their rigor, their willingness to help, their patience regarding my stubbornness and their teaching in general. I am profundly grateful for the opportunity that they granted me.

I have a grateful thought for all research friends that have welcomed me at Neurospin, for their valuable comments, discussions and technical advice: Thomas Bazeille, Théo Desbordes, Christos Nikolaos, Minye Zhan, Antonio Moreno, Isabelle Denghien, Swetha Shankar, Ana Luisa Pinho, Olivier Grisel.

Thanks to the friends that gave me the strength to go to Saclay and with whom I shared so many coffees, lab meetings, team meetings, meetings, jokes and so on. I will keep beautiful memories from the time spent at neurospin, climbing walls, around drinks, surfing or during trips abroad. Thanks to Thomas Chapalain and Himanshu Aggarwal for always bringing a smile and a good joke to share. Thanks to Tiffany Bounmy for introducing me to caffeine. Thanks to Alexis Thual for bringing so much enthousiasm in our discussions and projects. Thanks to Corentin Bel, Théo Morfoisse and Yvan Nedelec for the nice climbing sessions.

And more broadly, I want to thank all UNICOG and Parietal/Mind friends especially: Alexander Paunov, Audrey Mazencieux, Caroline Bévalot, Cédric Foucault, Lorenzo Ciccione, Maëva l'Hôtellier, Fernanda Ponce, Raphael Meudec. I also want to thank Philippe Ciuciu, Alexandre Gramfort and Demian Wasserman for interesting discussions. Not forgetting Pierre Bellec, François Paugam, Pravish Sainath, Julie Boyle and Isil Bilgin for their warm welcome in Montréal and the great time I had collaborating with them.

This thesis project was conducted both at Neurospin (CEA) and at INRIA. For that, I thank all the PIs and the administration. Thanks to Arnaud Martel, Jean-Marc Le Failler and the GIPSI, without whom my models would still be running for a few centuries. Thanks to the former Parietal team assistant, Corinne Petitot, and to the ED STIC assistant, Stéphanie Druetta, without whom, I would have been lost in the meanders of administration.

My thoughts go to my parents, my brother, and my sister who supported me all along, trusted me, and gave me the strength to move forward. I know they will take great pleasure in reading my thesis. A special thought goes to my grandmothers who left too early but would have dreamed of seeing the end of this project. They inspired me and taught me the willingness to stay strong and positive when facing tough times, and to never let go.

And thank *you*, if you are reading these lines (and the rest).

Contents

List of Figures

List of Tables

Abbreviations

Brain Regions

- STG: superior Temporal Gyrus
- STS: superior Temporal Sulcus
- TP: Temporal Pole
- IFG: inferior Frontal Gyrus
- IFS: inferior Frontal Sulcus
- dmPFC: Dorso-Medial Prefrontal Cortex
- pMTG: posterior Middel Temporal Gyrus
- TPJ: temporo-parietal junction
- pCC: posterior Cingulate Cortex
- AG: Angular Gyrus
- SMA: Supplementary Motor Area

Other

- NLP: Natural Language Processing
- NLM: Neural Language Model
- LM: Language Model
- ANN: Artificial Neural Network

Part I

Background

1 - Introduction: Probing the neural bases of language comprehension with artificial language models

This introductory chapter presents the approach that we followed in the quest to understand language comprehension in the human brain. *'Language processing'* is first described under the prism of Marr's Tri-Level Hypothesis (Marr, 1982), which decomposes any cognitive system into 3 hierarchical levels of understanding. From these three levels of understanding, emphasis is given to the deepest one: neurolinguistics. Specifically, motivations are given to explore the neural bases of language comprehension, which is a particular area of interest within neurolinguistics. The description starts with a brief review of the methods used to probe brain activity during language processing as well as the on-going debates in the field. Then we highlight the opportunity represented by artificial language models in the investigation of the neural bases of language comprehension, and presents the thesis outline.

1.1 . Exploring Language through Marr's Three Levels of Analysis

1.1.1 . Computational level: linguistics

The highest level in Marr's hierarchy, the *Computational Level*, addresses *what* the system does and *why*, that is, the overall goal and purpose of the system. This level describes language as a system of rules and procedures for transforming and manipulating symbols. These symbols can represent sounds, words, phrases, or entire sentences, and they are used to communicate ideas, convey information, or express emotions. These unique features have laid the foundations of human society, allowing to accumulate and transmit knowledge through time and space. They allowed humans to express themselves and understand each other, playing a crucial role in social interaction, in the transmission of cultural features, or more generally in passing down knowledge from one individual to the next.

Despite language ubiquity, it remains a complex and dynamic system that varies across communities and contexts. The formal scientific study of language and its structure is called *linguistics*. It revolves around understanding the patterns, rules, and mechanisms underlying human language, including its sounds, grammar, meaning, and use. Linguistics can be divided into several sub-fields, including phonetics, phonology, morphology, syntax, semantics, and pragmatics.

Phonetics is the study of the physical properties of speech sounds, including their production, perception, and acoustic properties. Phonology is the study of the patterns of sounds and how they are used to form words and convey meaning. Morphology is the study of the structure of words, including the rules for forming new words and the meaning of word parts (such as prefixes and suffixes). Syntax is the study of the structure of sentences, including the rules for combining words into phrases and clauses. Semantics is the study of the meaning of words and sentences. Finally, pragmatics is the study of how meaning is related to context and culture, including how people use language in social situations,

that is, how language is used to convey meaning beyond the literal interpretation of words. Taken together, these sub-fields provide a comprehensive framework for understanding the structure of language and how it is used to convey meaning.

One of the key findings of linguistics is that all languages share certain fundamental properties, such as the ability to express tense, aspect, and modality, and to create complex structures through recursion (see Hauser et al. (2002), but see Everett (2005) for counter arguments). However, languages are also subject to variation in their sounds, grammar, and vocabulary, reflecting the specific cultural and historical contexts in which they have evolved. In addition to its theoretical insights, linguistics has practical applications in fields such as education, translation, and language technology. By understanding the structure and mechanisms of language, linguists can help to develop more effective language learning strategies, improve machine translation systems, and design more efficient natural language processing algorithms. Overall, the computational level of analysis seeks to answer the "*what*" and "*why*" questions of language, and provides a high-level view of language and its role in human communication and cognition.

1.1.2 . Algorithmic level: psycholinguistics

The next level is the *Algorithmic Level*, that is, *how* does the system achieve its goal? It involves identifying the steps and processes used by the system to perform the desired function. For language, the algorithmic level seeks to answer questions such as: how is language *processed by the mind*? The investigation of the mental processes that underlie language is called *psycholinguistics*.

More precisely, psycholinguistics is a sub-field of linguistics that focuses on the psychological mechanisms involved in the production, comprehension, and acquisition of language. This includes understanding how individuals identify and parse speech sounds, build meaning from words and sentences, and generate language output. Psycholinguistics relies on behavioral methods, such as reaction time experiments, eye-tracking studies, and brain imaging techniques such as functional Magnetic Resonance Imaging (fMRI), to investigate the mental processes involved in language.

Research in psycholinguistics has covered many topics and has built our current understanding of language processing in the human mind. For examples, for language acquisition, psycholinguists have identified different stages of language development: phonological, syntactic, and semantic. They have also discovered that infants are able to distinguish between different sounds and can recognize patterns in language before they are able to produce words (Goodman, 1997; Werker and Tees, 1984). For language comprehension, psycholinguists have shown that the processing of language involves a serie of steps, such as phonological processing, syntactic parsing, and semantic integration (Just and Carpenter, 1986). They have also demonstrated a complex interplay between bottom-up and top-down processing. Bottom-up processing refers to the processing pathway going from sensory input, such as the sounds and letters of speech, to more abstract representations. Conversely, top-down processing refers to the use of prior knowledge, expectations, and context to guide the processing of lower-level linguistic features. For example, when listening to a sentence, the listener's brain uses bottom-up processing to analyze the sounds and structure of the speech, but also relies on top-down processing to interpret the meaning of the sentence based on prior knowledge and context. Thus, the understanding of some sounds might be modified depending on the individual's prior on what he might hear, and the sound he hears might modify his understanding of the context. For language produc-

tion, psycholinguists have studied how people plan speech, select words, and formulate sentences (Dell, 1986; Levelt, 1989). They also looked into how bilinguals were able to switch between languages, how language disorders affect language production and perception, or the factors that influence language proficiency (Bialystok et al., 2012; Kroll and Bialystok, 2013). Psycholinguists have also investigated the relationship between language and other cognitive processes, such as memory (Craik and Lockhart, 1972), attention, and problem-solving (Boroditsky, 2001). They have shown that language can influence these processes and that cognitive factors can also affect language processing.

Overall, psycholinguistics is a rich and complex field that seeks to unravel the mysteries of language processing and how it is accomplished by the human mind. It provides insights into how we acquire, use, and understand language, and has practical applications in fields such as language education and language therapy.

1.1.3 . Implementational level: neurolinguistics

Finally, the lowest level in Marr's hierarchy, the *Implementational Level*, covers the physical realization of the system. It requires to understand *how the system is implemented* in the brain or other physical medium. In the case of language, this means understanding the neural and cognitive processes involved in producing and comprehending language, as well as the physical and technological infrastructure that supports language use in the world. The identification of the brain regions involved in language processing, the neural pathways that connect them, and the cellular and molecular mechanisms that enable neural communication form what is called *neurolinguistics*.

Neurolinguistics addresses questions such as: what are the neural networks involved in different aspects of language processing, such as speech perception, syntax, and semantics? The seminal works of Paul Broca (Broca, 1861) and Carl Wernicke (Wernicke, 1874) in the 19th century lay the foundation of current neurolinguistics, identifying respectively the left inferior Frontal Gyrus (IFG) and left Temporal regions as the respective epicenters of language production and comprehension. Since then, studies have shown that the Broca's and Wernicke's areas are just small parts of a larger network of brain regions involved in language processing (Hickok and Poeppel, 2007; Price, 2012). One of the key insights of neurolinguistics is that language is not localized in a single brain region, but rather involves the coordinated activity of multiple networks of brain regions whose function differ. For example, the parsing of speech occurs bilaterally in the Superior Temporal Gyri (STG) (Hickok and Poeppel, 2007), the neural bases of syntax processing seem to involve the IFG and posterior superior Temporal regions (Friederici, 2012; Friederici et al., 2017), while the cortical regions involved in semantic processing are more distributed, involving posterior multi-modal and hetero-modal association cortices, hetero-modal prefrontal cortex, and medial limbic regions (Binder et al., 2009). Despite multiple investigations, the neural bases of language comprehension and especially the extent to which the regions that process semantics and syntax are separated or intertwined remains highly discussed. Subsequent chapters will delve into the matter more extensively.

Neurolinguistics has also shed light on the neuroplasticity of the brain in relation to language processing. Studies have shown that the brain can reorganize its neural networks to compensate for damage or to adapt to new language environments, such as in the case of bilingualism (Mechelli et al., 2004).

By understanding how language is implemented in the brain, neurolinguistics has the potential to contribute to the development of more effective language therapies and lan-

guage technologies.

Overall, the implementational level of analysis seeks to answer the "*how exactly*" questions of language, providing an in-depth view of the mechanisms and processes underlying language processing.

1.2 . Investigating the Neural Bases of Language Comprehension

1.2.1 . Why study the Neural Bases of Language Comprehension?

Understanding the lowest level of Marr's hierarchy, and more specifically, the neural bases of language comprehension, is important for several reasons. By understanding how the brain processes language, we can link brain mechanisms and processes to the working of the human mind and how it relates to our social interactions and daily life. Secondly, it can also provide insights into the general inner working of the brain: how does it process information and perform complex cognitive tasks? Thirdly, understanding the neural bases of language comprehension can help us understand and treat language-related disorders, such as dyslexia, aphasia, and other communication disorders. Finally, it can help us design better language technologies in fields such as speech recognition, language translation, and natural language processing. Understanding the processes underlying cognition can have various implications in fields such as artificial intelligence and robotics.

1.2.2 . Searching for the Neural Bases of Language Comprehension: a brief history

The investigation of the neural bases of language comprehension, and of cognition in general, started with lesion studies. Scientists inferred the cortical localisation of brain functions by correlating impairments with diseases- or traumas-related brain lesions. Famous examples of such individuals include Phineas Gage whose personality was dramatically altered after damaging his frontal lobe, while keeping much of his cognitive abilities, Paul Broca's "Tan" who lost his language skills after suffering damage to his left frontal cortex, or even H.M. whose bilateral medial temporal lobotomy led to global amnesia for new material. These early studies identified left-hemisphere regions, namely Broca's and Wernicke's areas, as playing a critical role in language processing, with Broca's area (IFG[1]) being associated with speech production and Wernicke's area (STG/STS) with language comprehension. More generally, patients undergoing local brain tissue stimulation or removal (during a surgical intervention) provide direct evidence of brain function localisation. The function of a brain region can also be determined from the recording of cortical activity prior to surgical intervention thanks to subdural or intra-cortical electrodes placed directly into the patient's brain (Halgren et al., 1978; Vignal et al., 2000). These studies have been used to investigate the temporal dynamics of language processing, and have identified the precise timing and sequence of neural activity involved in different aspects of language comprehension. For example, Edwards et al. (2010) uses electrocorticography (ECoG) to investigate the neural activation patterns associated with verb generation and picture naming tasks in the human brain, showing distinct stages of perception, semantic association/selection, and speech production. However, these studies are "case-studies", very rare and highly dependent on the medical needs of the patients.

While the previously described techniques are highly invasive, some alternative have

[1]Brain regions' abbreviations are listed in Appendix A.1

been developed to create *temporary* lesions or stimulations of the brain by disrupting the electrical currents between neurons. These methods, named TMS and TES, respectively create a magnetic and electrical current at the surface of the scalp that leads to a perturbation of the normal electrical activity during a cognitive task. The impairment elicited gives insights on the functional role of the stimulated brain region. For example, a study using TMS found that disrupting activity in the left IFG damaged participants' ability to produce grammatically correct sentences, suggesting that this region is critical for syntactic processing (Sakai et al., 2002). However, the precise localisation of the perturbation is difficult to assess, leading to broad functional mapping (Pascual-Leone et al., 1999).

Beyond brain lesions and stimulations, functional brain mapping can be done by identifying the brain regions eliciting neural activity while performing a certain task. There are several ways to probe neural activity. The oldest one, called electroencephalography (Berger, 1938), records the electrical activity of the human brain using electrodes placed at the scalp surface (see Swartz and Goldensohn, 1998, for a brief history). A more recent one, called magnetoencephalography (MEG) (Cohen, 1968), records neural activity by capturing the magnetic field generated by the cortical currents. It detects magnetic field gradients using superconducting quantum interference devices (SQUID, Clarke, 1996). Both MEG and EEG have good temporal resolutions on the order of milliseconds, allowing them to capture neural activity dynamics throughout the brain. For example, an event-related brain potential response, the N400, has been linked to meaning processing in EEG data, see Kutas and Federmeier (2011) for a review of the N400 characterization and evolving use. EEG and MEG devices respectively capture electrical or magnetic current outside the scalp and have to solve the inverse problem to identify the source of the signal. This ill-posed problem leads to poor spatial resolution, making both acquisition modalities uneffective at mapping brain function to precise brain regions.

Positron Emission Tomography (PET), MRI, and especially fMRI, appeared later in history (see Raichle, 2000, for a summary). Although previous methods attempted to directly measure the electromagnetic activity of neurons, PET and fMRI use proxies of brain activity. PET captures changes in blood flow by detecting gamma rays, released from the decay of radioactive atoms. While fMRI detects changes in magnetic properties caused by variations in blood oxygenation levels resulting from neural activity. PET has been an important tool for mapping both physiological and cognitive brain functions. A seminal study using PET-imaging is the one of Petersen et al. (1988), who revealed the cortical areas involved in single-word processing. However, PET has several important limitations such as its use of ionising radiations, the acquisition time of an image (> 30s, assuming that the brain is in a "steady-state" during this time) and finally its spatial resolution ($> 1 \times 1 \times 1\text{cm}^3$).

MRI allows the imaging of anatomical brain structures by detecting the magnetic properties of soft tissues in a non-invasive and safe way. To map brain function, that is, to capture brain activity, a variant of MRI called fMRI has been developed. fMRI detects changes in brain activity by means of the blood oxygenation level variations. A deeper description of MRI and fMRI are given in Chapter 2. It has a very good spatial resolution on the order of millimeters and no risk for the participant (if security measures are followed). FMRI has revolutionized the search for the neural bases of cognition and especially for language comprehension (see, e.g., Binder et al., 2009; Friederici, 2011; Hickok and Poeppel, 2007; Price, 2012, for reviews). Examples include Bookheimer (2002), that

reviewed fMRI studies to understand the cortical organization of semantic processing, or Hickok and Poeppel (2000) who reviewed fMRI studies to show task-induced biases in the characterization of the functional neuroanatomy of speech perception.

Overall, studies found that the previously identified left-lateralized regions, namely the STG/STS and IFG, were part of a larger network of areas, including the middle temporal gyrus (MTG), and the superior temporal gyrus (STG), extending up to the temporo-parietal junction. These brain regions where found to be involved in various aspects of language processing, such as phonological processing as well as syntactic and semantic processing. Taken all together, neuroimaging studies have provided insights into the neural bases of language comprehension, and have helped to identify the critical brain regions and networks involved in this process.

1.2.3 . A review of debates on the neural bases of language comprehension

Even after decades of investigation the neural bases of language comprehension are not fully understood. As a matter of fact, the complex nature of language makes it difficult to discern how the various processes underlying language processing are topographically and dynamically organized in the human brain, and therefore many questions remain open to this date (see Poeppel et al. (2012) for a description of issues encountered when investigating the neurobiology of language). The following presents a non-exhaustive list of these open debates and questions that remain unanswered.

Modularity of language One central open question in neurolinguistics concerns the modularity of language processing: Is language processing modular or distributed across multiple brain regions? Many studies have tried to elucidate this puzzle, but came to contradictory conclusions. Modularity can be defined accordingly to Jerry Fodor's theory (Fodor, 1983). Fodor defined a module as a cognitive system that is domain-specific, informationally encapsulated, automatic, and mandatory. According to him, modules are specialized systems that process information in a highly efficient and automatic manner, without interference from other cognitive processes (Fodor, 1983). Neuronal modularity of language processing gained support from early lesion studies suggesting, for example, that syntactic processing takes place in localized and specialized brain regions such as Broca's area, showing double dissociations between syntactic and semantic processing (Caramazza and Zurif, 1976; Goodglass, 1993). Neuroimaging studies (Embick, 2000; Friederici et al., 2006, 2017; Garrard et al., 2004; Grodzinsky and Santi, 2008; Hagoort, 2014; Hashimoto and Sakai, 2002; Matchin and Hickok, 2020; Pallier et al., 2011; Shetreet and Friedmann, 2014; Vigliocco, 2000) as well as simulation work on language acquisition and processing in artificial neural language models (Lakretz et al., 2019, 2021; O'Reilly and Frank, 2006; Russin et al., 2019; Ullman, 2004) have provided further support to this view since then.

However, in parallel, an opposing view has argued that semantics and syntax are processed in a common distributed language processing system (Bates and Dick, 2002; Bates and MacWhinney, 1989; Dick et al., 2001). Recent work in support of this view has raised concerns regarding the replicability of some of the early results from the modular view (Siegelman et al., 2019) and provided evidence that semantic and syntactic processing in the language network might not be so easily dissociated from one another (Fedorenko et al., 2020; Mollica et al., 2018).

Taken all together, these findings suggest that language processing involves a network of interconnected brain regions that take part in various processing. Indeed, functional neuroimaging studies have identified distinct brain regions involved in different aspects of language processing, such as phonology, syntax, and semantics. However, these regions show significant overlap and inter-connectivity, suggesting a distributed system for language processing, and discrediting the neural modularity in favor of a functional modularity, i.e. different functional activation patterns instead of distinct brain regions associated with different processing.

The neural bases of syntax and semantics Another question that has divided the scientific community concerns the neural bases of semantic and syntactic processing. What are the neural bases of syntactic processing? Similarly, how are word meanings and concepts represented in the brain?

Language comprehension requires to access word meanings (lexical semantics), but also to compose these meanings to construct the meaning of entire sentences. In languages such as English, semantic composition strongly depends on word order in the sentence - for example, 'The boy kissed the girl' has a different meaning than that of 'The girl kissed the boy' although both sentences contain the exact same words. The brain constructs these different meanings conditionally on words order, which is the backbone of sentence processing, indicating how to combine the lexical meanings of its sub-parts. Importantly, meaning construction of new sentences would be roughly done in the same way if only the structure of the sentences remains the same ('The X kissed the Y'), independently of the lexical meanings of the single nouns in the sentences ('boy' and 'girl'). This combinatorial property of language allows to construct meanings of sentences that we have never heard before and suggests that it might be computationally advantageous for the brain to have developed neural mechanisms for composition that are separate from those dedicated to the processing of lexico-semantic content. Such neural mechanisms for composition would be sensitive to only the abstract structure of sentences and would implement the syntactic rules according to which sentence parts should be composed.

Following related considerations, the neuroimaging studies that have probed the neural bases of syntax and semantics mostly relied on controlled experimental paradigms that manipulate the words or sentences (Bottini et al., 1995; Caplan et al., 1998; Mazoyer et al., 1993; Pallier et al., 2011; Stromswold et al., 1996), trying to isolate specific aspects of language processing. This approach probes linguistic dimensions in one of the following ways: varying the presence or absence of syntactic or semantic information (Friederici et al., 2003, 2009a) or varying the syntactic structure difficulty or the semantic interpretation difficulty (Cooke et al., 2001; Friederici et al., 2009b; Kinno et al., 2007; Newman et al., 2010; Santi and Grodzinsky, 2010). However, the conclusions from such studies may be bounded to the peculiarity of the task and setup used in the experiment (Nastase et al., 2020). To overcome these shortcomings, over the last years, researchers have become increasingly interested in data using "Ecological Paradigms", in which participants are engaged in more natural tasks, such as conversation or story listening (LeBel et al., 2022; Lerner et al., 2011; Nastase et al., 2021; Pasquiou et al., 2022; Regev et al., 2013; Wehbe et al., 2014a). This avoids any task-induced bias and takes into consideration both lexical and supra-lexical levels of syntax and semantic processing.

Overall, the neural bases of both syntactic and semantic conditions remain under in-

vestigation, even if they have been studied to a great extent, see (Binder and Desai, 2011a; Binder et al., 2009; Ralph et al., 2017) for reviews on semantics and (Friederici, 2012; Grodzinsky and Friederici, 2006; Vigneau et al., 2006) for reviews on syntax.

Hemispheric specialization To what extent are different aspects of language processing localized in the left hemisphere of the brain, and what role does the right hemisphere play? The study of the brain hemispheric lateralization, when processing language, overlaps the previous problematics, but still has an interest of its own.

Since the lesion studies of Broca (Broca, 1861) and Wernicke (Wernicke, 1874), the left hemisphere was considered to group most of the regions involved in language processing. Additional studies of patients with brain damage have shown that damage to the left hemisphere, particularly in Broca's or Wernicke's areas, can result in language impairments such as aphasia (Dronkers, 1996). These results suggest that these areas are critical for language processing. It has been discovered later on that this hemispheric lateralization was dependent on one's handedness: most right-handed people presented this left-dominant hemisphere for language (96%), while left-handed people presented this left-dominance in only 73% of cases (Knecht, 2000). Most studies investigating core components of language like phonology, syntax and semantics with controlled experimental stimuli only found activations in the left hemisphere, enforcing its ubiquity in language processing.

However, while the left hemisphere is (most of the time) dominant for language processing, it has been shown that the right hemisphere is also involved in language comprehension. In particular regarding aspects such as prosody (intonation, stress, and rhythm) and discourse processing (processing the meaning of sentences in context) (Bookheimer, 2002; Hickok and Poeppel, 2007). Additional studies have shown that the right hemisphere is involved in the processing of emotional prosody, or the emotional tone of speech (Ross, 1979), and that damage to the right hemisphere can result in impairments in processing prosody. Brain activations in the right hemisphere have also been reported in more recent studies using naturalistic paradigms, that is, stimuli closer to what the participant could find in a daily environment. Such studies reported widely distributed activations in both hemispheres (Binder et al., 2009; Huth et al., 2016), confirming the role of the right hemisphere in language comprehension.

In conclusion, while different aspects of language processing are primarily localized in the left hemisphere of the brain, the right hemisphere also plays a role in language comprehension, particularly in aspects such as prosody and discourse processing.

The neural bases of compositionality Finally, one last example of open debates in neurolinguistics concerns the neural bases of compositionality. Compositionality is the idea that combinatorial operations from a finite set of building blocks can generate novel expressions. In other words, the meaning of a phrase or sentence is a function of the meanings of its constituent parts and the way they are put together.

The neural bases of compositionality have been studied using a variety of neuroimaging methods. Studies investigating basic composition have shown that the left inferior frontal gyrus (Left IFG) is activated during simple rule-based syntactic computation (Zaccarella et al., 2015). Additional works have focused on the neural mechanisms underlying the processing of semantic compositionality. For example, some studies have shown that the left anterior temporal lobes (left ATL) is involved in the representation of concepts and

their semantic relationships, which are essential for semantic compositionality (Bemis and Pylkkänen, 2013, 2011; Chang et al., 2022; Lambon Ralph et al., 2010). Going beyond basic composition, Lerner et al. (2011) investigated various levels of compositionality, from the merging of phonemes into words, up to the grouping of sentences into paragraphs, finding distributed network of brain regions that work together to support the integration of meaning across different levels of linguistic representation.

Taken together, these findings suggest that the brain has specialized mechanisms underlying compositionality, allowing it to build bigger linguistics structures from smaller constituent parts of language.

1.3 . Probing the brain with Artificial Neural Networks (ANNs)

Recently, thanks to advances in natural language processing (NLP), neural language models (NLMs)[2] have been increasingly employed in the quest for understanding language comprehension in the human brain. Neural language models are models based on neural networks, trained to capture joint probability distributions of words in sentences using, for example, next-word, or masked-word prediction tasks (e.g. Devlin et al., 2019; Elman, 1991; Pennington et al., 2014; Radford et al., 2019). By doing so, these models learn semantic and syntactic relations among word tokens in the language, which allow them to generate coherent sentences and perform language-related tasks such as sentiment analysis and machine translation. Over the last decades, there has been a proliferation of neural language models. From Word2Vec and GloVe co-occurrences models, to Recurrent Neural Networks like GRU and LSTM, to the recent Transformers like BERT and GPT-2, the search for better and better language models is blooming (see Chapter 3 for more details). Assessing the quality of a language model typically involves evaluating its performance on a range of natural language processing tasks, including next-token prediction, named entity recognition, question answering, sentiment analysis, and others. Although neural networks are theoretically capable of approximating any function[3], designing an effective language model requires finding the optimal combination of model architecture, size, and training data quality and quantity. Achieving this balance can be a challenging task.

Because these models learn the statistical structure of natural language, researchers have used them to simulate the neural activity of the human brain during language processing. By comparing the models' behavior to the neural activity of human participants, they have been able to gain insights into the neural mechanisms underlying language comprehension and production. These models have proven especially useful in the analysis of ecological paradigms, that is, stimuli that are not constrained by any tasks, and that are similar to what you could find in a daily environment. Such stimuli include, among others, story listening or movie watching.

1.3.1 . Comparing Neural Language Models' behavior to neural activity

To study brain data collected from story listening or reading, NLMs are presented with the same sentence stimuli, and their activations (a.k.a. embeddings) are extracted and used to fit and predict the brain data (Caucheteux and King, 2022; Huth et al., 2016; Pasquiou et al., 2022; Schrimpf et al., 2020; Wehbe et al., 2014a). The embeddings derived

[2]that is Artificial Neural Networks applied to language

[3]According to the Universal Approximation Theorem

from NLMs are generally the stacked outputs of a subset of the units of the model.

The seminal work of Mitchell et al. (2008) is one of the first to introduce fMRI brain data prediction using features derived from computational models. Given a stimulus word w, they began by analyzing the occurrences of w within a large text corpus. Using this information, they created a vector of intermediate semantic features that represented the meaning of w. Then, they predicted the fMRI activation of each voxel, independently, as a weighted sum of the neural activations contributed by the intermediate semantic features. Fig. 1.1 illustrates this last step. They found that the statistics learnt from words co-occurrences could predict the fMRI activations resulting from the same words stimuli. This work has marked the beginning of a shift, from studies that have cataloged the patterns of fMRI activity associated with specific categories of words and pictures to the systematic prediction of arbitrary-word related fMRI activations using computational models.

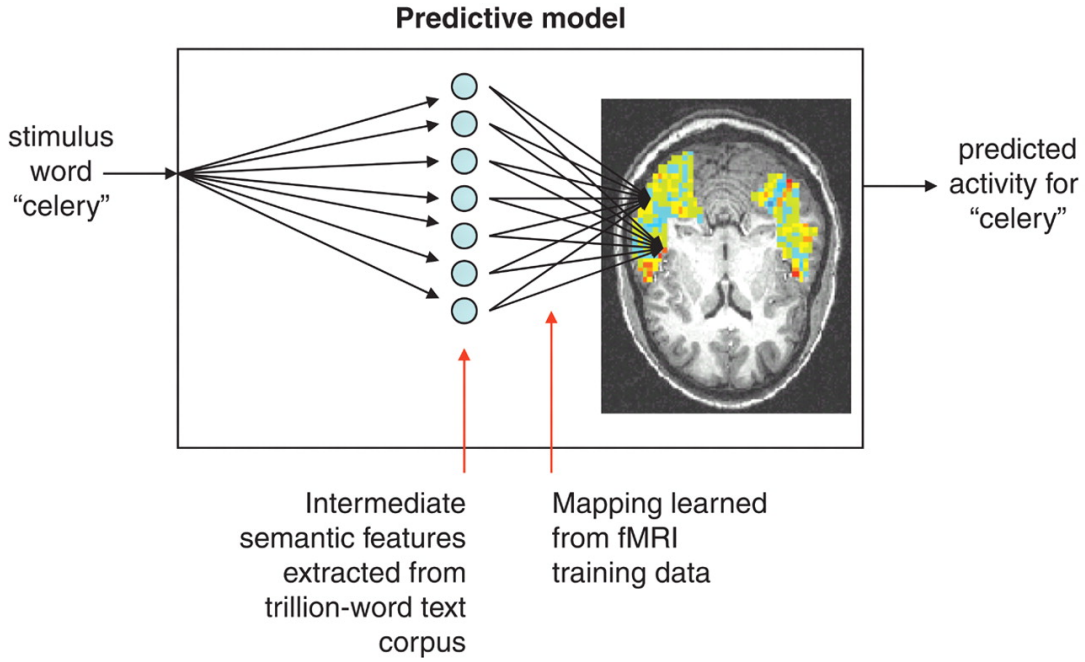


Figure 1.1: **Predicting fMRI activations for arbitrary nouns stimuli.** fMRI activation is predicted in a two-step process. The first step encodes the meaning of the input stimulus word in terms of intermediate semantic features whose values are extracted from a large corpus of texts exhibiting typical word use. The second step predicts the fMRI as a linear combination of the fMRI signatures associated with each of these intermediate semantic features. (Figure taken from Mitchell et al., 2008).

A decade later, the field of natural language processing has been revolutionized in virtually all areas thanks to neural network-based language models such as Recurrent models or Transformers. The performance of these new models far outperformed the one of co-occurrences models, making them more interesting tools to probe language processing in the brain. Schrimpf et al. (2020) ran a systematic study, evaluating several pre-trained state-of-the-art language models on the correlation of their internal representations to three human neural datasets, including fMRI and ECoG (see Fig. 1.2A). This work gave a broad overview of the ability of various language model architectures to fit brain data acquired

with several modalities, from embedding models like GloVe and Word2Vec, to RNN, LSTM and Transformer-based models.

More recently, Caucheteux and King (2022) showed that modern language algorithms partially converge towards brain-like processes. Using fMRI and MEG data, they considered the impact of architecture, training, and performance on the ability of deep language models, such as CNN or Transformers, to generate brain-like representations. They found that the activations extracted from late-middle layers of transformer-based models explained better brain data compared to the first or last layers, and that trained neural language models were better than poorly trained models at fitting brain data (see Fig. 1.2B).

Overall, recent results using neural language models to study neuro-imgaing data suggest that brain activations and neural language models partially converge to similar linguistic representations (Caucheteux and King, 2022), suggesting that NLMs are able to capture some of the key features of language processing in the human brain. This was also shown for MEG (Toneva et al., 2020), and intracranial data (Goldstein et al., 2021).

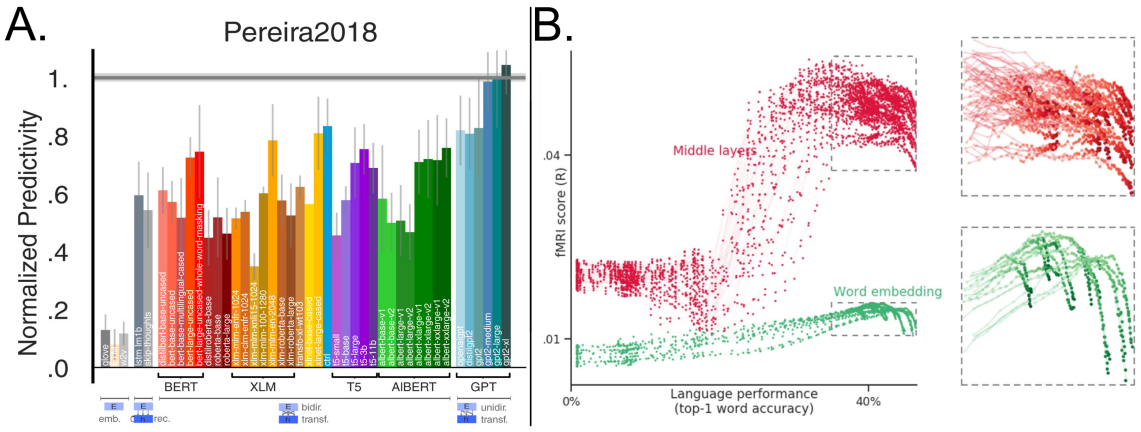


Figure 1.2: **Fitting fMRI brain data with Neural Language Models' embeddings.** A) Schrimpf et al. (2020) compared 43 computational models of language processing (ranging from fixed embeddings to recurrent and transformer models) in their ability to predict human fMRI brain voxel responses to visually presented (sentence-by-sentence) passages (Pereira et al., 2018). Model classes are grouped by color and their normalized predictivity (a.k.a. "brain score") is displayed, i.e., the fraction of explainable signal that the model can predict (Figure taken from Schrimpf et al. (2020)). B) Caucheteux and King (2022) compared models' ability to predict fMRI brain activity. They used 18 causal architectures, and displayed results separately for the input layer (word embedding, green) and the middle layers (red) (Figure taken from Caucheteux and King, 2022).

1.3.2 . Investigating Semantic and Syntactic processing using NLM

By using the approach described above, researchers have made significant discoveries. For instance, Wehbe et al. (2014a) illustrated how continuous, co-occurrences-derived, word embeddings can create rich semantic descriptions of word stimuli. In a comprehensive analysis, they presented an integrated computational model of reading that incorporates various sub-processes underlying story understanding. By fitting brain data with their descriptors of the input stimuli, they highlighted brain regions involved in the processing of syntax and semantics , but also in the processing of visual or motion features (see their Fig. 4b).

Likewise, Huth et al. (2016) used word embeddings to fit fMRI brain data, and revealed broad networks associated with semantic processing. By mapping regression weights to voxels, they show that the distribution of semantically selective areas is relatively symmetrical across the two cerebral hemispheres, and that the organization of these brain areas is consistent across individuals (see Fig. 1.3). These results diverged from older lesions studies that suggested a more localized and lateralized semantic mapping of concepts. One hypothesis put forward by the authors is that the right hemisphere responds more strongly to naturalistic stimuli compared to the words and short sentences stimuli used in most studies.

Following these tracks, recent attempts to solve the central puzzle of the neural bases of semantics and syntax tried to leverage the modeling ability of neural language models combined with naturalistic datasets.

While some studies using naturalistic stimuli have identified vast, distributed networks for syntax and semantics (Caucheteux et al., 2021; Fedorenko et al., 2020), others (Matchin et al., 2017; Pallier et al., 2011) have found more localized activations for syntax, typically

in inferior frontal and posterior temporal regions, when using constrained experimental paradigms. As a result, the extent of the independence of the representational systems as well as their neural bases still remains debated to date. So far, insights from neural language models about this central puzzle were also rather limited. This is mostly due to the complexity of the models in terms of size, training and architecture. This complexity makes it difficult to identify how and what information is encoded in their latent representations, and how to use them to study brain function.

A recent illustration of the use of NLMs to probe the neural bases of syntax and semantics can be found in Caucheteux et al. (2021). They used a neural language model, GPT-2 (Radford et al., 2019), to separate semantic and syntactic processing in the brain. Specifically, using a pre-trained GPT-2 model, they built syntactic predictors by averaging the embeddings of words from sentences that shared syntactic but no semantic properties, and used them to identify syntactic-sensitive brain regions. They defined as semantic-sensitive brain regions, the regions that were better predicted by the GPT-2's embeddings computed on the original text, compared to the syntactic predictors. They observed that syntax and semantics, defined in this way, rely on a common set of distributed brain areas.

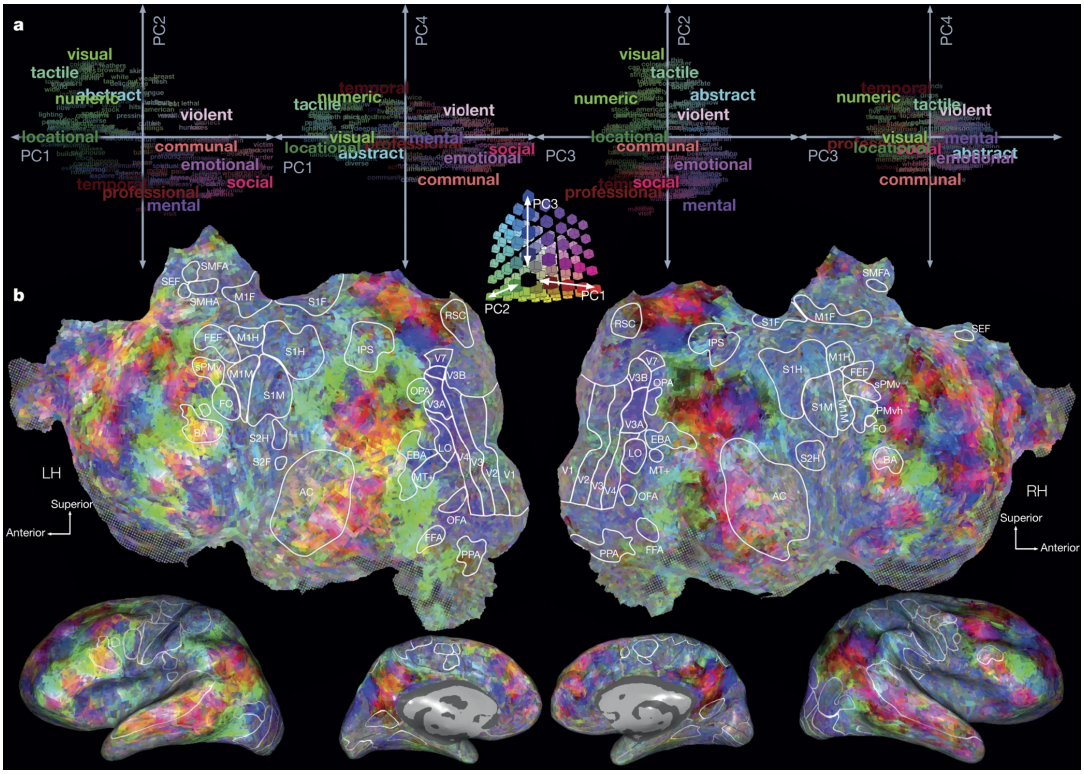


Figure 1.3: **Mapping semantic features to the human cortex.** After performing a Principal Component Analysis on word embeddings, Huth et al. (2016) mapped words into the space defined by their Principal Components (PC). Using the first three dimensions of the PCs space, they defined an RGB colormap that was used to colour both words and voxels. A) Words were projected onto two of their PCs. B) Huth et al. (2016) fitted brain data with the previous word embeddings and projected voxel-wise model weights onto the PCs space and then coloured using the same RGB colormap (Figures taken from Huth et al., 2016). Semantic information is represented in intricate patterns across much of the semantic system.

1.3.3 . Finding context-sensitive brain regions using NLM

The advent of more complex neural language models has enabled the investigation of more precise language sub-processes, such as compositionality. Following the works of Lambon Ralph et al. (2010), Bemis and Pylkkänen (2011) and Bemis and Pylkkänen (2013), a few studies have tried to leverage computational models to identify the neural bases of compositionality and quantify brain regions' sensitivity to increasing sizes of context. Some of them, using ecological paradigms, have found a hierarchy of brain regions that are sensitive to different types of contextual information and different temporal receptive fields (e.g., Jain and Huth, 2018; Toneva et al., 2022; Wehbe et al., 2014b). A notable investigation (Jain and Huth, 2018) used pre-trained LSTM (Hochreiter and Schmidhuber, 1997) models to study context integration. They varied the amount of context used to generate word embeddings, and obtained maps indicating brain regions' sensitivity to different sizes of context (see Fig. 1.4). However, because the ability of LSTMs to leverage contextual information is far from perfect, the exact hierarchy of brain regions integrating contextual information remains debated to date. Toneva et al. (2022) went deeper in the analysis of context-sensitivity by rigorously separating the neural bases of lexical and supra-lexical

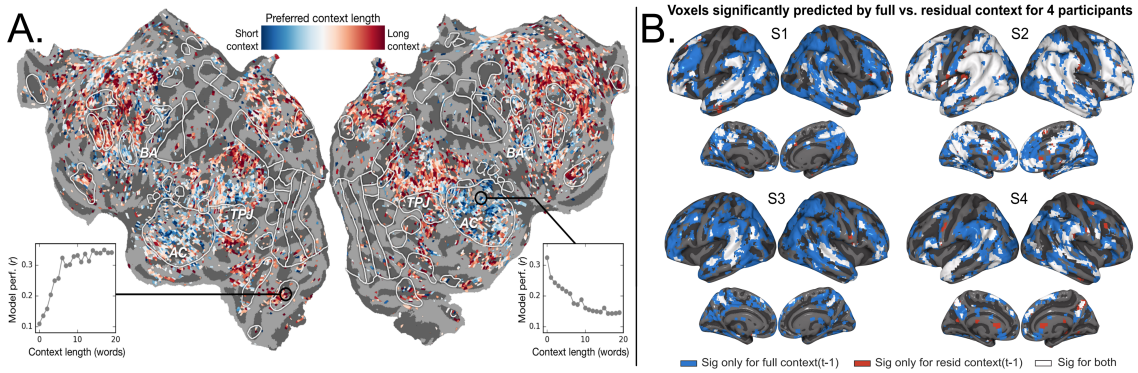


Figure 1.4: **Probing context-sensitivity using NLMs.** Jain and Huth (2018) and Toneva et al. (2022) assessed context-sensitivity in the human brain using NLMs. **A)** Context length preference across cortex. An index of context length preference is computed for each voxel in one subject and projected onto that subject's cortical surface. Voxels shown in blue are best modeled using short context, while red voxels are best modeled with long context. Non-significantly predicted voxels (mean r < 0.11) are gray. Insets show model performance with each context length for two representative voxels, one that prefers short context (right) and one long (left). Generally, voxels in low-level language areas (AC) prefer short context, while voxels in higher-level language areas prefer long (Figure taken from Jain and Huth (2018)). **B)** Voxels significantly predicted by full-context embeddings (blue), residual-context embeddings (red), or both (white), visualized in MNI space (fMRI data). Full-context embeddings are embeddings extracted from contextual-models like LSTM or GPT-2, and residual-context refers to what is uniquely explained by the supra-lexical information in full-context embeddings, that is when lexical information has been removed (Figure taken from Toneva et al. (2022)).

features. Using LSTM-based word embeddings, they assessed the brain regions uniquely predicted by supra-lexical content, by removing the shared information with the lexical processing level.

1.3.4 . Limits of the brain-ANN comparison

The use of transformer-based models and other computational models has provided valuable insights into the neural bases of language comprehension. Nevertheless, certain discrepancies between these models and the human brain have given rise to doubts about the extent to which our understanding of brain function can be advanced through their use. First, the architecture of Transformers is based on multi-head self-attention modules which does not clearly map on neural computations in biological networks (e.g., Dayan and Abbott, 2005). Does this architecture contribute to or hinder the ability of the model to predict brain activity compared to other, possibly more brain-like, architectures (e.g., recurrent neural networks)? Second, the data used to train Transformer-based models is often different from that available for children, both in type and size. Training a Transformer-based model requires massive corpora, on the order of billions of words, whereas children require orders of magnitudes less words to achieve comparable or better linguistic performance. How does the training corpus (type and size) affect the model's ability to fit brain activity? Finally, the learning and evaluation objective commonly used with these models, such as masked or next-word prediction, is at most a rough approximation of the computational problem the human brain solves during language acquisition and processing. Can

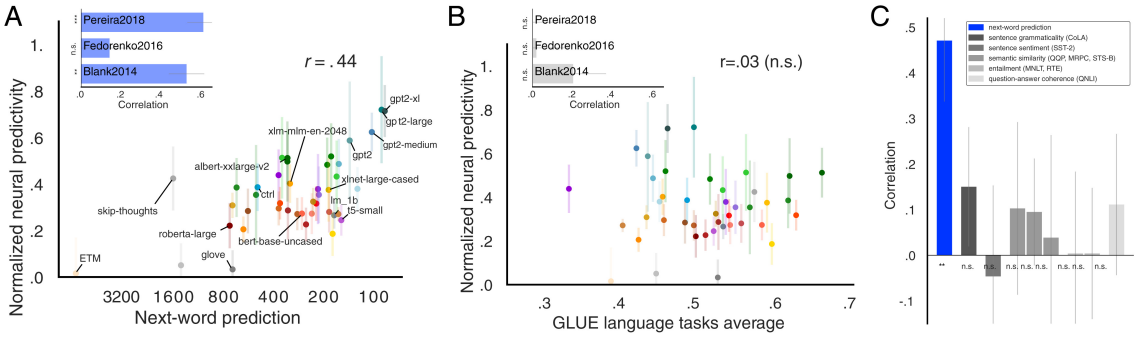


Figure 1.5: **Model performance on a next-word-prediction task selectively predicts brain scores.** (A) Next-word-prediction task performance was evaluated as the surprisal between the predicted and true next word in the WikiText-2 dataset of 720 Wikipedia articles, or perplexity (x axis, lower values are better). (B) Performance on diverse language tasks from the GLUE benchmark collection does not correlate with overall or individual-dataset brain scores. (C) Correlations of individual tasks with brain scores. Using pre-trained NLMs, Schrimpf et al. (2020) showed a partial correlation between models' accuracy in a next-word prediction task and their ability to fit fMRI brain data (Figure taken from Schrimpf et al. (2020)).

one consider that a well-trained model (according to perplexity loss) is a good model for brain activity in language tasks?

Recent investigations (Schrimpf et al., 2020) found that models' neural fits and their fits to behavioral responses are both strongly correlated with model accuracy on the next-word prediction task, and that model architecture appears to substantially contribute to neural fit (see Fig. 1.5). However, many parameters of these comparisons, like model training, are not controlled, calling for caution in the interpretation of these results.

1.4 . Thesis Outline

Despite recent advances and extensive neuroscientific and cognitive explorations, many debates remain open, such as the neural bases of semantics, syntax or the integration of contextual information among others. The following will showcase the outline of the thesis and also highlight the significant contributions made throughout its course.

All the analyses performed in this thesis use the fMRI brain data of the 51 English participants of 'The Little Prince' fMRI corpus. This dataset, henceforth referred to as *TLP*, fully described in Chapter 2, contains the fMRI brain activations of participants passively listening to 'The Little Prince' audiobook (∼90min). Chapter 2 introduces the fundamentals of functional Magnetic Resonance Imaging (fMRI), how data are pre-processed and how statistical analyses are performed. It demonstrates that the model-free approach known as the *Shared Response Model* surpasses the traditional Inter-Subject Correlation method in accurately estimating the ceiling of explainable signal.

Chapter 3 introduces Neural Language Models (NLMs), and how to extract informative representations from these models.

Chapter 4 introduces the General Linear Model and the estimation methods. Then we analyze the impact of the fMRI dataset size on the encoding model performance, and the link between this performance and the regularization hyperparameter of the encoding

model.

Chapter 5 compares various pre-trained and custom-trained Neural Language Models' ability to fit the fMRI data of TLP dataset (GloVe, LSTM, GPT-2, BERT). To achieve this, a specific training dataset named *The Integral Dataset* was created to regulate the training data and vocabulary size. A second set of analyses probes the interactions between transformer-based models' architecture (Vaswani et al., 2017) and brain regions, and quantifies information redundancy across transformers' layers.

Chapter 6 addresses the question of syntactic vs. semantic representations in the brain. A novel approach is presented which involves using word embeddings derived from *information-restricted* models to fit brain activity. These models are trained on text corpora from which specific types of information (syntactic or semantic) have been removed. The fitting performance of these models is assessed and compared to that of a neural model trained on the original, untouched dataset.

The objective of Chapter 7 is to pinpoint the specific regions of the brain that are involved in processing information beyond the lexical level, and also to determine the extent to which these regions integrate such information within a given time frame. Two distinct methods are used to investigate the role of context integration in language processing: 1) information-restricted neural language models and 2) masked-attention generation. Initially, a comparison is made between a contextual model (GPT-2) and a non-contextual model (Glove) to differentiate the brain regions responsible for lexical processing from those for supra-lexical processing. Subsequently, transformer-based GPT-2 models are used to investigate the brain regions involved in processing short (5 words), medium (15 words), and long (45 words) contexts (in number of words of past context taken into account to build an embedding). In a separate analysis, masked-attention generation is used with a pre-trained GPT-2 model to identify the window size over which past context information is integrated in each brain region. The attention mechanisms are used to regulate token interactions, in order to fit TLP data with word embeddings leveraging fixed size of context.

Finally, Chapter 8 investigates the limits of ANN-brain comparisons. In the final series of analyses, this study explores the contribution of the transformer architecture to the model's ability to predict brain activity when compared to potentially more brain-like architectures, such as recurrent neural networks. Various aspects of the models' architecture, along with the type and size of the training corpus are controlled to compare the fitting ability of trained and untrained models. Finally, we investigate the impact of training and training data on the models' ability to fit fMRI brain activity, and examine the relationship between perplexity and brain score.

2 - Investigating brain activity with fMRI

This chapter introduces functional Magnetic Resonance Imaging (fMRI). This imaging technique is used to acquire images of brain activity on which several analyses can be performed to probe the neural bases of language comprehension. This chapter first outlines the fundamentals underlying fMRI data acquisition (see Buxton (2009) for a detailed description of MRI and fMRI principles). Then, it gives a brief overview of the preprocessing required to clean and prepare fMRI data for later statistical analyses, before introducing subject-level and group-level analyses. The dataset used in all of our experiments, that is, 'The Little Prince' fMRI corpus, and the scientific motivations to put it together are then described in details. Finally, I present the different atlases used during the thesis and how to design an estimation of the maximum amount (ceiling) of fMRI signal that can be explained, using the model-free SRM method.

2.1 . MRI

Magnetic Resonance Imaging is a non-invasive imaging technique producing 3-dimensional anatomical images of body tissues. Leveraging the strong magnetic field of powerful magnets, MRI scanners causes protons to align in the static magnetic field direction (*Nuclear Magnetic Resonance*). An electromagnetic pulse is emitted, orthogonally to the static field, causing the protons to spin out of equilibrium. As the Radio Frequency field is turned off, protons return to their original alignment while releasing energy (re-emitting electromagnetic waves) that is captured by antennas. More precisely, the transient transversal moment (orthogonal to the static magnetic field) cancels with a time constant T_2, while the longitudinal moment (parallel to the static magnetic field) reaches its equilibrium with a time constant T_1. T_1 reflects the time taken to reach original magnetization, while T_2 reflects the time taken to release the energy. Both time constants depend on the chemical nature of the molecules as well as the local magnetic environment.

MRI consists in imaging in 3-dimensions the distribution of the transverse magnetization produced by the chosen pulse sequence in each brain volume. Leveraging the fact that the resonant frequency is directly proportional to the magnetic field, we can alter the magnetic field in a controlled way so that it varies linearly along any particular axis. This causes the resonant frequency to also vary linearly with position along that axis. These linearly varying fields are called gradient fields and are derived from additional coils in the scanner. There are 3 orthogonal sets of gradient coils that produce gradient fields along any axis in a MRI scanner. Thanks to these gradient fields, one can identify the signal coming from specific slices of the brain volume, as well as the contribution of each portion of this slice to the measured signal. The signal captured by the antennas is then converted into volume images, using the previously mentioned different magnetic properties.

2.2 . BOLD fMRI

Human cognition relies on the electrical activity of a vast network of neurons (about 10^{11}). Electrical activity passes from one neuron to another through ionic currents resulting from the intertwined action of membrane depolarization and neurotransmitters release.

Once the signal has been transmitted, resources need to be renewed for the signal to pass again. Resources renewal include the uptake and repackaging of the neurotransmitters as well as the restoration of ionic gradients. These processes require the consumption of Adenosin Triphosphate (ATP), which in turn, requires the glucose and oxygen brought by the Cerebral Blood Flow (CBF). Thus, the CBF increases, close to neural activations, can be used as a proxy of brain activity. This is the principle behind Positron Emission Topography (PET). However, BOLD Functional Magnetic Resonance Imaging (fMRI) does not directly measure the CBF, but the blood oxygen level dependent (BOLD) signal, i.e. it is sensitive to the proportion of oxygenated versus deoxygenated hemoglobin.

While fully oxygenated blood has a similar magnetic susceptibility compared to other brain tissues, deoxyhemoglobin, i.e. hemoglobin that is not yielding oxygen, is paramagnetic and alter blood susceptibility. BOLD fMRI measures this distortion of the magnetic field around blood vessels related to deoxygenated blood. According to the seminal work of Ogawa et al. (1990) (see also Ogawa et al. (1992)), 40% of the oxygen carried by arterial blood is extracted and metabolized. Thus, there is a substantial amount of deoxyhemoglobin in the venous vessels, which leads to a decrease in the Magnetic Resonance signal. When the neuronal activity increases in a brain region, the local blood flow increases markedly, however, the relative increase in oxygen metabolism remains small. As a consequence, the proportion of deoxygenated blood decreases, leading to a signal increase (a few percent at 1.5T, 5-15% at 4T). A more detailed analysis shows that BOLD signal arises from the interplay of blood flow, blood volume, and blood oxygenation in response to changes in neuronal activity. However, we leave the detailed description of the physical phenomena (that are still poorly understood) to studies focusing on the subject (Buxton, 2009).

The scanner captures these local signal increases in space and time, and then reconstructs the entire brain volume at each time-point (scan) acquired, resulting in 4-dimensional data, with 1 dimension for time and 3 for position in space.

An important point to keep in mind, is that the mechanisms underlying the cerebral blood volume changes are still poorly understood (Logothetis, 2002; Logothetis et al., 2001), which limits the general confidence in the fact that the BOLD signal reflects neural activity (Logothetis, 2008; Logothetis and Wandell, 2004). For example, it has been shown in Goense and Logothetis (2008); Logothetis et al. (2001); Thomsen et al. (2004) that an haemodynamic response could be observed in the absence of neuronal spiking output, or that the observed changes in BOLD signal co-localize with neural activity depending on the imaging sequence used or the magnetic field strength (Logothetis, 2008). Nonetheless, fMRI still represents one of the best imaging techniques to probe brain activity.

2.3 . Preprocessing fMRI data

Once fMRI data has been acquired thanks to the MRI scanner, it still requires several steps of preprocessing in order to be used in subsequent analyses. These preprocessing steps aim at removing artefacts:

- due to the scanner or to the subject's anatomy,
- or due to the participant motions,

as well as preparing the data for future analyses. The following will only give a brief

overview of the different steps of fMRI preprocessing. See Poldrack et al. (2011) for more details on fMRI preprocessing steps.

Data Quality Control This is the first step in making sure that the data can be used, ensuring that there is no potential outlier scans. These verifications are especially useful to find artefacts related to with-in scan effect linked to non-rigid effects of motion. It is usually a good advice to repeat this step at the end of the pre-processing pipeline.

Distortion correction Once the first high-level verification has been done, the experimenter has to correct for distortion., i.e. correct for losses of signal and geometric distortions of the fMRI signal. These artefacts are mainly due to inhomogeneities in the magnetic fields around brain regions where air and tissues interface, which causes drops in the signal or geometric distortions.

Slice timing correction fMRI volume images are built by stacking 2D MRI slices images, collected one at a time, in an order that varies depending on the acquisition method chosen. Therefore, to avoid temporal inconsistencies inside each scan, and later in the modeling of the fMRI data, slice timing has to be corrected.

Motion correction One of the main sources of artefacts in fMRI data is motion. Motion can be physiological (respiration, cardiac cycle, swallowing) or stimulus-related (when pushing a button). While part of the noise it induces can be removed by tracking these physiological or stimulus-related features and use them as confounds, there still remain two major effects: 1) a mismatch of the location of subsequent images in the time series (bulk motion), and 2) a disruption of the MRI signal (spin history effect). While (2) is impossible to correct afterwards, one can reduce the mismatch between images (1) using realignment or motion correction techniques.

Spatial normalization One might be tempted to learn about the shared bases of neural processing inside a group of subjects. In order to be compared properly, subjects have to be projected onto the same template image in a given coordinate space. More precisely, to account for anatomical variability, one need a reference frame in which individual data are projected (=*inter-subject registration* or *spatial normalization*).

Spatial smoothing Finally, a last step of data preprocessing is to spatially smooth the fMRI signal. Removing small scale changes (high frequency information) will increase the signal to noise ratio (SNR). For example, if data was acquired with small voxels to limit signal dropout near regions presenting susceptibility artefacts, one can reduce the increased noise by smoothing. Averaging information spatially also reduces the mismatch between individuals related to spatial functional variability.

2.4 . Statistical analyses

Once the data has been pre-processed, it can be properly analyzed. There are 2 levels of analysis: 1) at single subject level and 2) at the group level.

Subject-level analyses The first stage of fMRI data analysis is to model each individual participant's brain activations. More precisely, the aim is to assess whether the BOLD signal time series of an individual change in response to the stimuli or some features related to the stimuli. A common approach to estimate these relationships is to use *encoding models* (see Chapter 4).

Encoding models make predictions about representational spaces, that is, about the spaces in which live the sets of features or variables used.

When modelling fMRI brain activity, encoding models make use of the General Linear Model (GLM) (see Chapter 4) to fit voxels' timecourses using estimation methods such as the Ordinary Least Square (OLS) or penalized versions like Ridge or Lasso. More precisely, they fit a dependent variable (here the BOLD time series) with (independent) variables (aka *features*) that are assumed to partly mirror the expected BOLD responses following a given stimulus. For each voxel, an independent regression model is fitted. It takes as input time-series corresponding to the chosen features and outputs coefficients that weight the impact of each regressor on the estimation of the voxel's timecourse. In the end, a way to score models' goodness of fit is necessary to assess whether the model does better than chance. Evaluating the encoding models' performance usually entails comparing the predicted timecourses with the true timecourses, with metrics such as Pearson's correlation or the coefficient of determination.

This gives a single volume image with a scalar value per voxel indicating the performance of the encoding model at this location. In the context of this thesis, encoding models are used to test and compare brain-computational theories of language processing.

Group-level analyses After analyzing the fMRI data of each participant, it is important to evaluate the generalizability of the findings. Specifically, it is crucial to determine whether the effects observed at the individual level are consistent with those at the population level. This question is addressed with statistical tests. From the subject-level analysis, we obtained N-subjects volume maps indicating the ability of the set of predictors to fit the fMRI brain data of each subject at each voxel. Effects significance is then assessed through one-sample t-tests applied to the spatially smoothed maps, with an isotropic Gaussian kernel (e.g. having a full width at half maximum (FWHM) of 6mm). In each voxel, the test assessed whether the distribution of Pearson coefficients across participants was significantly larger than zero. These tests assume that values across participants are independent, are sampled from a normal distribution and that the dependent variable is continuous. As the number of voxels in an image and therefore the number of t-tests performed is relatively high, the proportion of False discoveries can be substantial. To control for multiple comparisons, we corrected group-level analysis maps with either a False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) correction or a Bonferroni (Bonferroni, 1936) correction.

2.5 . Introducing The Little Prince fMRI dataset

2.5.1 . Motivations behind The Little Prince fMRI Corpus

All the fMRI-based analyses of this thesis used the English subjects of *The Little Prince fMRI Corpus* (Li et al., 2022), a multilingual fMRI dataset where English (51), Chinese (35) and French (28) speakers passively listened to the same audiobook (*The Little Prince*

novella) in their native language.[1]

In naturalistic designs such as story listening, the processing of multiple levels of language comprehension (e.g., word, phrase, sentence, discourse) unfold naturally at different timescales. Such a rich contextual setting extends the range of linguistic phenomena that can be examined using the same dataset, and allows for testing assumptions on the neural mechanisms of language processing at multiple levels. For example, whether different linguistic levels coincide with different frequencies of oscillatory activity in the brain (Giraud and Poeppel, 2012) and whether these levels correspond to a hierarchically organized coding architecture (Lerner et al., 2011; Nastase et al., 2020).

Ecological experimental paradigms go beyond classical controlled experiments, where a participant performs a task while experimenting strictly controlled stimuli that are later contrasted in order to highlight a specific process of interest. Such controlled experimental paradigms include reading words on a screen or a sequence of letters for example. One of the main motivations behind the use of ecological paradigms is to avoid biases related to the task or stimulus selection. Nonetheless, deep analysis of rich and complex naturalistic datasets requires the use of machine learning tools, to leverage as much information as possible from the stimuli. However, the richness of these datasets comes with an important drawback: the entanglement of the linguistic processes inherent to language processing. Therefore, special care is needed when analysing naturalistic data and when interpreting the results.

For this thesis, the fMRI data from the 51 English subjects of *The Little Prince fMRI Corpus* were selected as the primary source of analysis. This was based on the fact that, as of 2020, it was the largest fMRI dataset available in terms of both the number of participants and the number of scans. Since then, several initiatives have created larger fMRI datasets, regarding the number of participants (Nastase et al., 2021), or regarding the number of scans per subject (e.g. LeBel et al. (2022) or Courtois-Neuromod[2])

Table 2.1 summarizes recent naturalistic datasets using a linguistic stimulus.

Dataset name	Number of subjects	Stimulus duration	Reference
Lebel	8	$\sim$ 8 hours	LeBel et al. (2022)
LPP (english)	51	$\sim$ 90 min	Li et al. (2022)
LPP (french)	28	$\sim$ 90 min	Li et al. (2022)
LPP (chineese)	35	$\sim$ 90 min	Li et al. (2022)
IBC	11	$\sim$ 90 min (LPP stimuli)	Pinho et al. (2021)
Pereira 2018	16	$\sim$ 58 min	Pereira et al. (2018)
Harry Potter	9	$\sim$ 40 min	Wehbe et al. (2014a)
Narratives	345	(unknown) $\sim$ 20 min	Nastase et al. (2021)

Table 2.1: **fMRI naturalistic datasets.**

2.5.2 . FMRI data Preprocessing

The English speakers' fMRI data were acquired using a 3T MRI scanner (Discovery MR750, GE Healthcare, Milwaukee, WI) with a 32-channel head coil. Anatomical images were collected with a high resolution T1-weighted (1 mm^3 voxel-size) with a Magnetisation-Prepared Rapid Gradient-Echo (MP-RAGE) pulse sequence. Blood Oxygen Level De-

[1] Available from https://openneuro.org/datasets/ds003643/versions/1.0.2

[2] https://www.cneuromod.ca/

pendent (BOLD) signals were collected using a T2-weighted echo planar imaging (EPI) sequence (repetition time: 2000 ms, 3x3x3 mm^3 isotropic voxels, 3 echos).

FSL's Brain Extraction Tool (Jenkinson et al., 2012) was used for skull-stripping, with a fractional intensity threshold setting of 0.5. Subsequent preprocessing steps were carried out using AFNI version 16 (Cox, 1996). Anatomical and functional images were co-registered using the in-built AFNI function `3dseg`, images were normalised to the MNI–152 template.

Multi-echo independent components analysis (ME-ICA) (Kundu et al., 2012) was used to improve the signal-to-noise ratio in these data. ME-ICA splits the $T2^*$ signal into BOLD-like and non BOLD-like components. Removing these non-BOLD components mitigates noise due to participants' head motion, physiology and scanner conditions such as thermal changes (Kundu et al., 2017). Indeed, there were no exclusions based on degree of head movement.

A global brain mask was computed to only keep voxels containing useful signal[3] across all runs for at least 50% of all participants (26,164 voxels, at 4x4x4mm resolution). Finally, linear detrending and standardization (mean removal and scaling to unit variance) were applied to each voxel's time-series.

2.5.3 . Stimuli

The stimuli consisted of nine audio segments of 10 minutes each and the nine associated text transcriptions. One functional run was acquired for each participant and audio segment.

The text comprised 15,426 words and 4,482 punctuation signs, respectively sampled among 2015 unique words and 36 unique punctuation signs. The acoustic onsets and offsets of the spoken words were marked to align the audio recording with the *The Little Prince* text.

Distributions of sentence lengths and words' part-of-speech are displayed in Fig.2.1.

2.6 . Defining a ceiling of explainable signal

One of the long-term goals of neurolinguistics is to assess the common neural bases of language processing at population level. However, there is one major factor that limits our ability to address this question, namely *variability*. Assuming that there is no variability induced by acquisition tools, variability can be split into two:

- **intra-subject variability:** variability could emerge from a change in attention, or more generally from a change in the state of the subject during an experiment (or between experiments), e.g. induced by tiredness. If acquisitions are temporally far away, changes in connectivity, grey matter volume or other anatomical factors could account for intra-subject variability.

- **inter-subject variability:** variability could emerge from differences in subjects' anatomy or subject's neural connectivity.

These sources of variability are independent of the stimuli and add a degree of complexity when modelling fMRI data. The objective of this section is to design model-free estimations

[3]Using Nilearn's *compute_epi* function

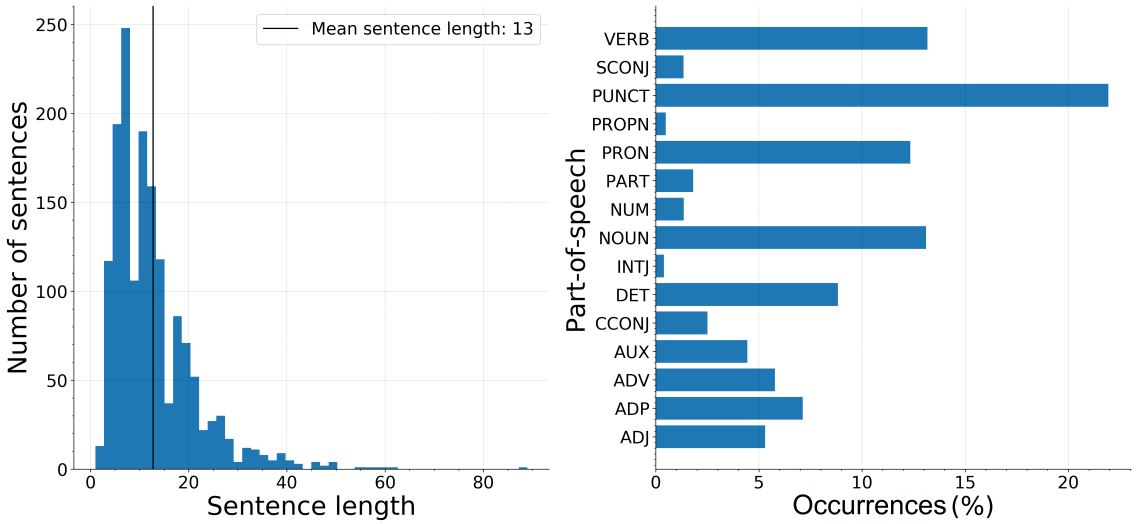


Figure 2.1: **Distribution of sentence features.** Distributions of sentence lengths, in number of words (left), and of part-of-speech occurrences, in percentage of the total number of words (right).

of the shared fMRI responses, across individuals, to a given stimuli. In the end, we want to use the model-free estimation's R scores[4] as a ceiling, that is, to consider these values as the maximal amount of fMRI signal that can be explained, and use them to normalise encoding models' performance (see Chapter 4). A model-free estimation only uses the fMRI data to make its predictions.

To assess the part of the fMRI signal that is shared across subjects for a common stimuli, Hasson et al. (2004) designed a model-free approach called *Inter-subject correlation* (ISC) analysis. ISC has the advantage of not needing a pre-defined response model, allowing the measurement of neural responses' consistency across individuals irrespective of the complexity of the stimuli. Additionally, it does not require any knowledge of the temporal composition of processes underlying the observed BOLD activity, as it only relies on subjects' haemodynamic responses. Hasson et al. (2004) used this technique to investigate common neural processes during movie watching. In ISC analyses, the signal of a voxel in a given subject is compared to the average signal in the other subjects:

$$\hat{\mathbf{X}}_i^{ISC} = \frac{1}{N} \sum_{sub=1, sub \neq i}^{N} \mathbf{X}_{sub}$$

We assessed the ISC modelling performance on the fMRI data of the English participants of 'The Little Prince' fMRI corpus. We followed the subsequent algorithm:

- we created 51 different train-test splits, using 50 participants as training data and 1 participant as testing data (the participant in the test data being different in each split)
- in each split, $\hat{\mathbf{X}}_i^{ISC}$ was estimated using the 50 training participants,

[4] A R score is a measurement of the ability of an estimation method to fit the time-course of a voxel/channel.

- we computed the Pearson coefficient for each voxel between $\hat{\mathbf{X}}_i^{ISC}$ and $\mathbf{X}_i$, obtaining 1 volume map per subject,

- we averaged across subjects the volume maps obtained at the previous step, obtaining a single volume brain map with an averaged Pearson coefficient per voxel.

From ISC's predictive R scores, we defined a ceiling of explainable signal: the maximal amount of explainable signal in each voxel is defined as the cross-validated R score obtained from the ISC method.

However, this approach is too simplistic as it completely ignores inter-subject variability. We want to design the best model-free estimation of the fMRI signal, taking into consideration inter-subject variability.

In this work, we used another model-free estimator of the explainable fMRI signal called the Shared Response Model (Chen et al., 2015)[5]. The Shared Response Model (SRM) has the advantage of taking into account inter-subject variability. Given multi-subject brain data, SRM factorizes it as a stimulus-specific response $\mathbf{S}_t$ (assumed to be sampled from a centered Gaussian) shared among all N subjects and subject-specific orthogonal transforms (bases) $\mathbf{W}_i$ (1 per subject):

$$\hat{\mathbf{X}}_{i,t}^{SRM} = \mathbf{W}_i\mathbf{S}_t + \mathbf{N}_i \text{ , } \forall \text{i=1 .. N, t=1 .. T}$$

Where $\mathbf{W}_i \in \mathbb{R}^{V\times p}$, $\mathbf{S}_t \in \mathbb{R}^{p\times T}$ and $\hat{\mathbf{X}}_{i,t}^{SRM} \in \mathbb{R}^{V\times T}$ with V the number of voxels, p the number of SRM components and T the number of time-points. $\mathbf{N}_i$ is the noise in subject i, assumed to be sampled from a centered Gaussian with covariance $\sigma_i^2 I$. Figure 2.2 details the decomposition in stimulus-specific and subject-specific components as well as the FastSRM algorithm used to perform this decomposition.

We assessed the SRM modelling performance on the English fMRI data of the 'The Little Prince' fMRI corpus, using 75 SRM components. We followed the subsequent algorithm:

- We created 51 different train-test splits. For each participant there were 9 runs of fMRI data.

- In each split, we chose 46 participants for training and 5 for testing,

- inside each split, we ran a nested cross validation on runs, where 8 runs were used for training and 1 for testing,

- we first estimated the spatial components for the 8 runs of the 46 training participants,

- we then estimated the shared responses (stimulus-specific components) for the 8 training runs and the 1 testing runs using the 46 participants for which we have the spatial components,

- knowing the shared responses, we estimated the spatial components for the 5 testing subjects,

[5]Using Richard et al. (2019) implementation

- for each test subject, we computed the voxel-wise Pearson coefficient between the predicted timecourses and the ground-truth
- we then averaged across splits (runs and participants) to obtain a single volume brain map with an averaged Pearson coefficient per voxel.

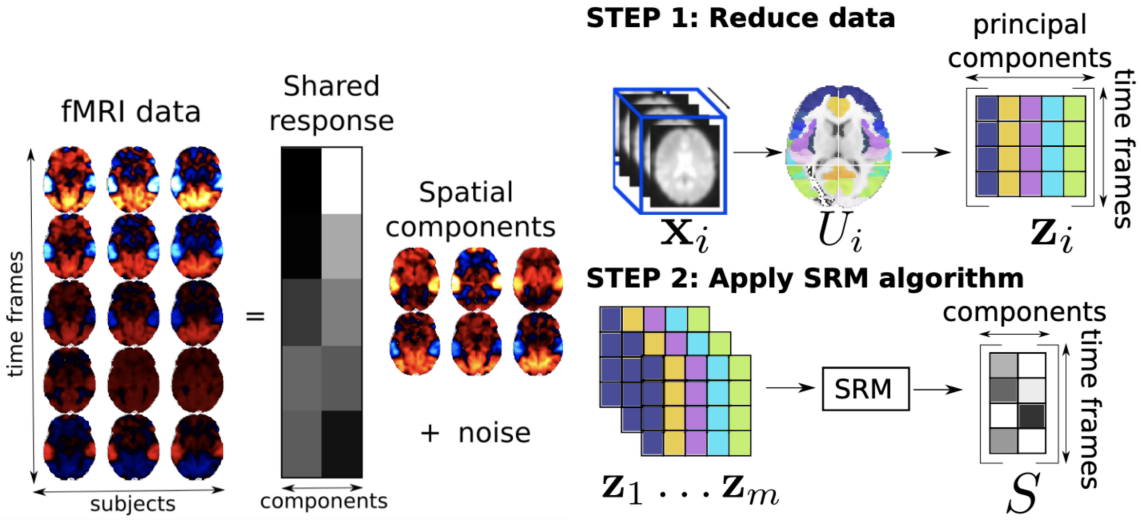


Figure 2.2: **(Left) Shared Response Model.** The raw fMRI data are modeled as a weighted combination of subject-specific spatial components with additive noise. The weights are shared between subjects and consitute the shared reponse to the stimuli. **(Right) Fast SRM Algorithm.** In step 1 (top), data $\mathbf{x}_i$ are projected onto a spatial decomposition U_i that may depend on the subject i. In step 2 (bottom), a SRM algorithm is applied on reduced data to compute the shared response. (Figures taken from Richard et al. (2019))

Figure 2.3 compares the predictive performance of the two previously described model-free estimations of the fMRI signal. It first displays the significant voxel-wise R score differences between SRM and ISC as a histogram (panel C) and on brain surface, corrected for multiple comparisons with a Bonferroni correction of $p < 0.1$ (panel A). It shows relatively high differences, with peaks (R scores around 0.15) in language related areas such as the STS[6] and frontal regions. Fig. 2.3B) and C) respectively display the ratio of R scores between ISC and SRM as a brain surface map and as a histogram, showing that, in average, SRM explains twice as much signal as ISC. However, the ratio in language related areas is higher, between 60% and 90%.

Both ISC- and SRM-derived ceiling can be artificially augmented as shown for ISC in the work of Schrimpf et al. (2020). They extrapolated the ISC modelling performance to infinitely many humans. The hypothesis behind was that the model-free estimation of the fMRI signal did not reach its full modelling potential because of the limited number of participants. By bootstrapping subsets of subjects of varying size, and computing the ISC, they fitted:

$$v = v_0 \times \left(1 - e^{-\frac{x}{\tau_0}}\right)$$

[6]Brain regions' abbreviations are listed in Appendix A.1

where v is each subsample's correlation score, x is each subsample's number of participants and v_0 and τ_0 are the fitted parameters for asymptote and slope respectively. Still, this enhancement of the ceiling does not overcome limits inherent to the modelling approach. Specifically, the enhanced ISC method still fails to consider variations among individuals.

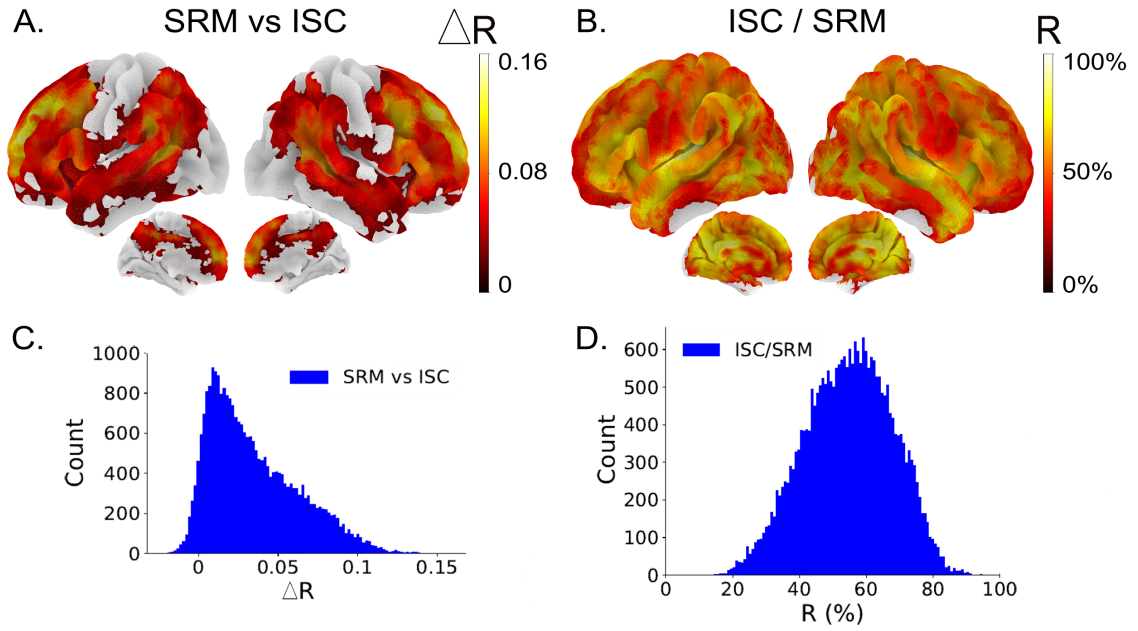


Figure 2.3: **Comparison of ISC and SRM R scores, computed on 'The Little Prince' fMRI dataset.** A) Brain regions that are significantly better predicted by SRM (in red) compared to ISC. Maps are voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with a Bonferroni approach $p < 0.1$. B) Voxels' R score for the Inter-Subject Correlation Model, in percentage of the voxels' R score for the Shared-Response Model. C) Distribution of the R scores differences between ISC and SRM. D) Distribution of the ratios between voxels' R score for the Inter-Subject Correlation Model and voxels' R score for the Shared-Response Model.

2.7 . Regions of Interest (ROIs)

Pre-processed fMRI data remains partly noisy. To gain in robustness in subsequent analyses, voxel-level information can be aggregated inside regions of interest (ROIs). A common approach to address this matter is the use of atlases. Atlases provide the location of several regions in a coordinate space. They can be probabilistic or not. In the first case, they specify the probability of each voxel to belong to a given ROI, and in the second case, they specify exactly which voxels belong to which ROI.

There exists many atlases at cortical or subcortical levels, probabilistic or not (e.g. Dadi et al., 2020; Desikan et al., 2006; Schaefer et al., 2018). In Chapter 5, we used the Harvard-Oxford atlas with 96 parcels from FSL. We used the deterministic version, called 'cort-maxprob-thr25-2mm', with images of shape (182, 218, 182).

In Chapter 7, a 'Dictionary of Functional Modes' or 'DiFuMo' (Dadi et al., 2020), is used instead. It can serve as a probabilistic atlas to extract functional signals with different dimensionalities (64, 128, 256, 512, and 1024). These models are optimized to

represent well raw BOLD timeseries, over a wide range of experimental conditions. The 'DiFuMo' atlas relies on a linear decomposition of fMRI time-series into a product of 2 matrices: one that contains spatial modes and one that contains the temporal loadings of each mode. It learns a dictionary of bases that enforces sparsity and non-negativity instead of independence on the spatial maps. In Chapter 7, we used the version containing 1024 parcels (or ROIs), with voxels edges of 3mm. Schemas of DiFuMo atlases and typical usages are displayed in Fig. 2.4.

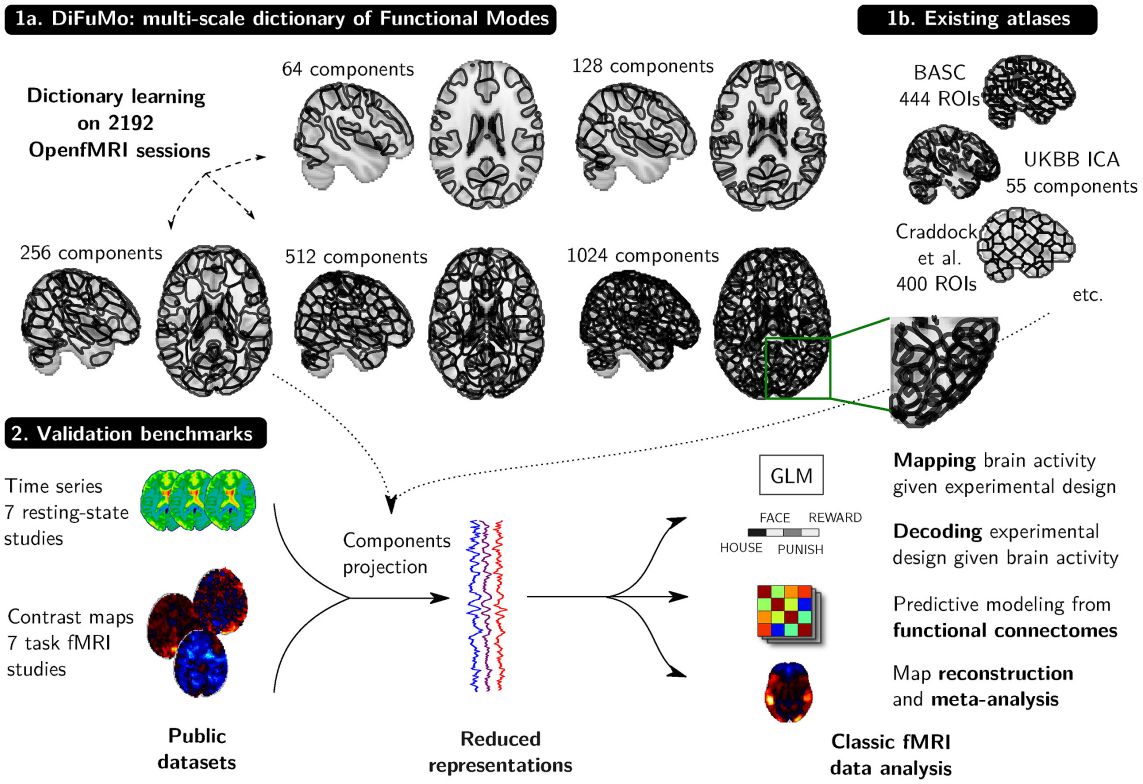


Figure 2.4: **Schema of DiFuMo atlases and their usage in typical fMRI analyses.** DiFuMo atlases are extracted from a massive concatenation of BOLD time-series across fMRI studies, using a sparsity inducing matrix factorization algorithm. The DiFuMo atlases were computed at different resolutions, up to 1024 components. Atlases were assessed in 4 benchmarks that measure suitability to classic fMRI analyses. Those are performed on reduced and non-reduced data, with different atlas sizes and a comparison between atlases. (Figure taken from Dadi et al. (2020))

This Chapter gave a brief introduction on MRI, fMRI and the required preprocessing steps before subsequent subject- or group-level analyses. We discovered 'The Little Prince' fMRI corpus, which is at the basis of all the fMRI analyses run in this thesis, and the various way the human brain could be parcellated into regions of interest. But more importantly, we learnt that the standard approach to define a ceiling of explainable fMRI signal, using Inter-Subject Correlation (ISC), has a lower sensitivity compared to the Shared Response Model (SRM). Thus, to avoid biases in the estimation of the percentages of explained signal, one should use the SRM method.

The next Chapter will introduce another fundamental block of this thesis: artificial language models, and how they can be used to build complex predictors of brain activity.

3 - Extracting linguistic features from Neural Language Models

This chapter introduces artificial neural networks, and more precisely, the variant that processes language: Neural Language Models (NLMs). NLMs have made many progress in the last decade, reaching human-level performance in several tasks. The increase in computational power and the expanding community of researchers in the field have led to a plethora of models covering a large range of architectural and training designs. The first section presents artificial neural networks as well as several instances of NLMs. The second section describes how these models encode information, and how to extract these very representations.

3.1 . Introduction to Neural Language Models (NLMs)

3.1.1 . Artificial Neural Networks

Artificial Neural Networks (ANNs), also known as Neural Networks or Neural Nets, are computing systems that are inspired by the biological neural networks found in animal brains. These systems are made up of interconnected units, called artificial neurons, which simulate the functions of neurons in a biological brain.

The connections in an ANN are called 'edges'. Similarly to the synapses in a biological brain, they transmit signal between neurons: each edge is associated with a weight that multiplies the output of the previous neuron before transmitting it to the next artificial neuron. Then, the neuron computes some non-linear function of the sum of its inputs. The weight of an edge increases or decreases the strength of the signal at a connection, and is adjusted during the learning process. In addition to these basic operations, many more specificities can be implemented. For example, neurons may possess a threshold that determines whether a signal is transmitted: the signal must exceed the threshold value in order to be transmitted.

Artificial neurons are usually grouped into layers, each performing different transformations on the inputs. The signals travel from the first layer, the input layer, to the last layer, the output layer, possibly passing through the intermediate layers multiple times. The connections between neurons can be *fully connected*, where each neuron in one layer connects to every neuron in the next layer, or *pooled*, where a group of neurons in one layer connect to a single neuron in the next layer. Some neural networks allow connections between neurons in the same or previous layers, they are referred to as recurrent networks. An example of a two-layer feed-forward artificial neural network is given in Fig. 3.1.

Neural networks learn by processing examples with known inputs and outputs, forming weighted associations between the two. The learning process tries to minimize the distance between the processed output of the network and the target output, adjusting the weighted associations accordingly until the network produces outputs that are increasingly similar to the target. The error rate is typically evaluated using a cost function, and the learning process continues as long as the cost continues to decline.

This learning process allows neural networks to perform tasks by considering examples without being explicitly programmed with task-specific rules. For example, in image recog-

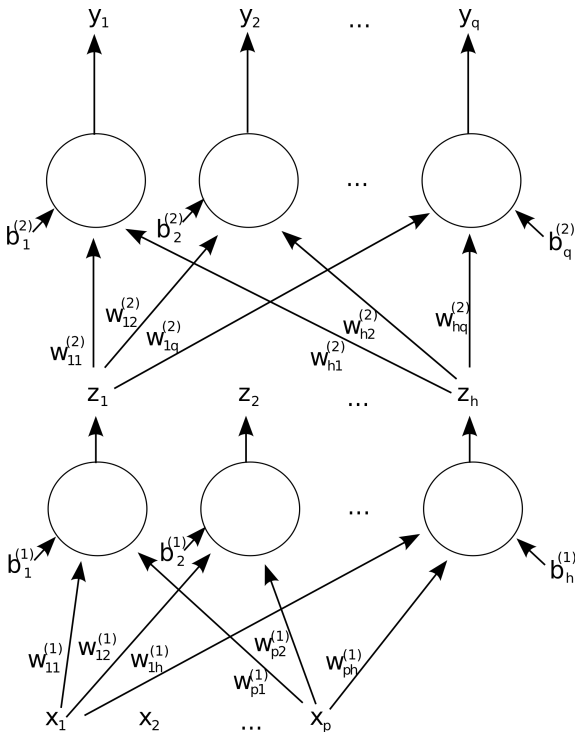


Figure 3.1: **A two-layer feed-forward artificial neural network.** w_i are the weights of each edge of the network. b_i are the biases of each neuron of the network. x_i, y_i are respectively the inputs and outputs of the network, while z_i are intermediate representations passed from the first layer to the second one. Figure taken from Wikipedia https://en.wikipedia.org/wiki/Artificial_neural_network.

nition, a neural network can learn to identify cats by processing examples of images labeled as "cat" or "not a cat". When processing new images, the network will leveraged the learnt weighted associations to identify whether the new images contain cats or not. This is done without any prior knowledge of what makes a cat, such as fur, tails, whiskers, and cat-like faces. Instead, the neural network automatically generates identifying characteristics from the examples it processes.

This work focuses on artificial neural networks applied to textual information, namely *Neural Language Models* (NLMs). These models are usually trained on tasks such as language generation, masked-language modelling, semantic/syntactic categorization or Named Entity Recognition. The following introduces different famous architectures of NLMs.

3.1.2 . Words co-occurrences models: GloVe and Word2Vec

Word2Vec The word2vec algorithm (Mikolov et al., 2013) learns to represent each unique word with a vector (aka *embedding vector*) that encodes word associations from vast amounts of text, and that is able to capture the semantic and syntactic qualities of the word.

There are two architectures used in word2vec: continuous bag-of-words (CBOW) and continuous skip-gram. Both architectures consider both individual words and the context words surrounding them as the model iterates on the corpus. Using a shallow neural network model, the CBOW model predicts the current word from the context words, while the skip-gram model uses the current word to predict the surrounding context words, with

nearby words being weighed more heavily.

Once the model has been trained, similar words have close embedding vectors in the vector space while dissimilar words have embedding vectors that are far from each other. This is because words that are semantically and syntactically similar will have similar context windows, thus learning similar embeddings.

The word2vec model can be trained using hierarchical softmax or negative sampling. The hierarchical softmax method uses a Huffman tree to calculate the conditional log-likelihood, while negative sampling minimizes the log-likelihood of sampled negative instances. While it has been noted that the quality of word embeddings increases with their dimensionality, the marginal gain decreases after a certain point. Consequently, the dimensionality of the vectors is typically set between 100 and 1,000.

GloVe GloVe (Pennington et al., 2014) is an unsupervised learning algorithm that learns to represent words with embedding vectors. It is not based on neural networks, but we still introduce it here because of its simplicity and its ability to build rich embedding vectors. GloVe embedding vectors are derived from the global co-occurrence statistics of words in a corpus. It is based on a log-bilinear model with a weighted least-squares objective and the key idea is that ratios of co-occurrence probabilities between words can reveal meaning. For instance, the ratio of co-occurrence probabilities between the target words "ice" and "steam" with other words in the vocabulary, shows that "ice" co-occurs more with "solid" and "steam" with "gas".

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Figure 3.2: **Words co-occurrences probabilities.** Figure taken from https://nlp.stanford.edu/projects/glove/.

The training objective of GloVe is to learn word embeddings such that their dot product equals the logarithm of the words' probability of co-occurrence. More precisely, it is trained on the non-zero entries of a global word-word co-occurrence matrix that specifies how frequently words co-occur with one another. Similarly to Word2Vec, GloVe's embedding vectors capture notions of semantics and syntax and perform well on tasks that evaluate word relationships, such as analogies.

Sometimes, computing words' nearest neighbors according to metrics like *cosine similarity*, reveals rare but relevant words that lie outside an average human's vocabulary. For example, the closest words to the word "frog" give: 'frogs', 'toad', 'litoria', 'leptodactylidae', 'rana', 'lizard', 'eleutherodactylus'.

3.1.3 . Recurrent Neural Networks: RNN, GRU, LSTM

A recurrent neural network (RNN, Rumelhart et al. (1988)) is a type of artificial neural network that is capable of processing sequences of inputs by maintaining an internal state or memory. This internal state allows the network to exhibit temporal dynamics, making

it suitable for tasks such as handwriting recognition and speech recognition. Unlike feed-forward neural networks, RNNs can process sequences of inputs of varying lengths.

However, traditional RNNs training can lead to the vanishing gradient problem. That is, when computing the gradient with the chain rule, the small gradient values of each layer might combine into a null-gradient update in the first layers of the model. To limit the vanishing gradient effect, RNNs have been augmented with gated states that are controlled by the network. Such gated memory is a key feature of long short-term memory (LSTM, Hochreiter and Schmidhuber (1997)) networks and gated recurrent units (GRU, Cho et al. (2014)). Since their first apparition, LSTMs have set accuracy records in multiple application domains, such as speech recognition, machine translation, and image captioning, outperforming traditional RNNs, hidden Markov models and other sequence learning methods. Fig. 3.3 illustrates the gated architecture of a LSTM cell. In this thesis, LSTM are trained on a next token prediction task. The model learns to predict the word following a sequence of input words. More precisely, it learns to predict the probability distribution over the vocabulary for the next token, given previous tokens. The objective is to minimize the error between the predicted probabilities and the actual next tokens in the training data by using a negative log likelihood loss on the predicted probabilities.

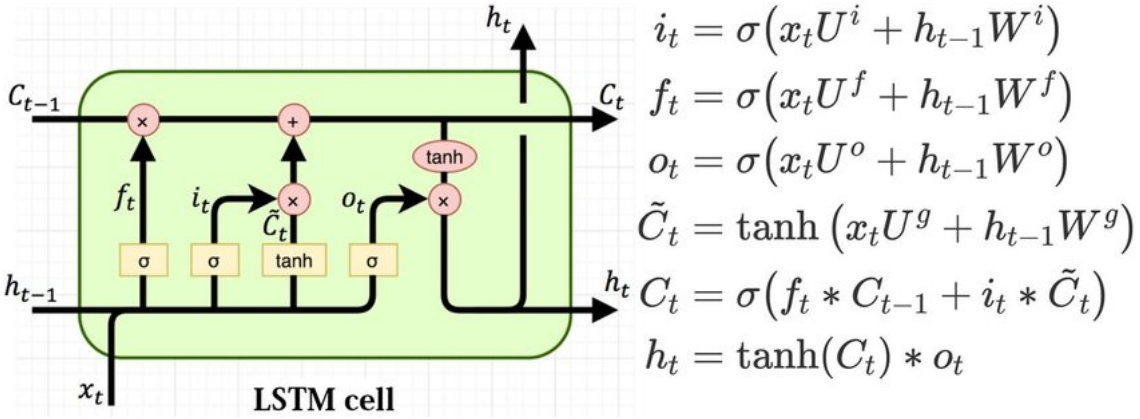


Figure 3.3: **Structure of a LSTM cell.** Each layer of a LSTM model contains several units. The structure of a LSTM unit/cell is represented here. h and C are respectively the hidden and cell states of the unit that are passed through time, while x is the output of the previous layer.

3.1.4 . Transformers: GPT-2 and BERT

Transformer Neural Language Models (Vaswani et al., 2017) are a type of deep learning models designed for processing sequential data such as text, speech, or time-series data. They make use of self-attention mechanisms, which allow them to process the entire input sequence as a whole, attending to different parts of the input sequence in parallel, rather than processing the sequence sequentially like traditional RNNs or LSTMs. More precisely, given a sequence of input tokens, transformer-based models first embed each token into a high dimensional vector. Each embedding vector can then interact with the other embedding vectors through successive projections and self-attention operations: vectors become weighted averages of themselves and their contexts, the weights being determined based on the dot-products between themselves and the other embedding vectors.

Two common and widely used transformer architectures are GPT-2 and BERT.

GPT-2 (Generative Pretrained Transformer 2, Radford et al. (2019)) is a language model developed by OpenAI that uses the Transformer architecture for natural language processing tasks. The model is trained on a massive corpus of text data through a process called unsupervised pre-training, where the model learns to generate text by predicting the next word based on the previous words in the sentence. More precisely, it learns to predict the probability distribution over the vocabulary for the next token, given previous tokens. The model is optimized using the negative log likelihood of the predicted probabilities, with the goal of minimizing the error between the predicted probabilities and the actual next tokens in the training data.

BERT (Bidirectional Encoder Representations from Transformers, Devlin et al. (2019)) is also a language model based on the Transformer architecture. Fig. 3.4 illustrates the BERT architecture. Unlike GPT-2, which is used for language generation, BERT is generally used for language understanding tasks such as named entity recognition, question answering, and sentiment analysis. It is trained using a process called masked-language modeling, where a portion of the input tokens are randomly masked, and the model tries to predict the original value of the masked tokens based on their context.

During training, BERT learns contextual representations of the input text by looking at both the left and right context of the masked tokens. This allows the model to understand the relationships between the words in the input text and to capture the context-dependent meaning of the words. After training, the pre-trained representations can be fine-tuned for specific NLP tasks by adding task-specific layers on top of the pre-trained BERT model.

In summary, both models use the Transformer architecture and are pre-trained on large amounts of text data, but GPT-2 is trained for language generation tasks, while BERT is trained for language understanding tasks.

Table 3.1: Number of trainable parameters (in millions) for several instances of LSTM, GPT-2 and BERT.

Number of layers Models	1	2	4	6	8	10	12
LSTM	81.6	86.3	95.8	105.2	114.6	124.1	133.5
GPT-2	46.3	53.4	67.5	81.7	95.9	110.1	124.2
BERT	46.4	53.6	67.7	81.9	96.1	110.3	124.4

Training such heavy architectures usually requires a specific data infrastructure. Table 3.1 gives an overview of the number of trainable parameters for several instances of LSTM, GPT-2 and BERT.

3.2 . Encoding information using latent representations

To perform the task they are trained on, NLMs manipulate continuous and high dimensional vectors on which they apply various kind of transformations. These high dimensional vectors can be seen as the representations of some properties extracted from the model's input. From these high dimensional NLMs vectors, one can build rich word representations called *embedding vectors* or *latent representations*. These embedding vectors can consist of the entire state of the model, that is all the units that compose the neural language model, or only part of it, such as the vectors from a subset of layers for example. These

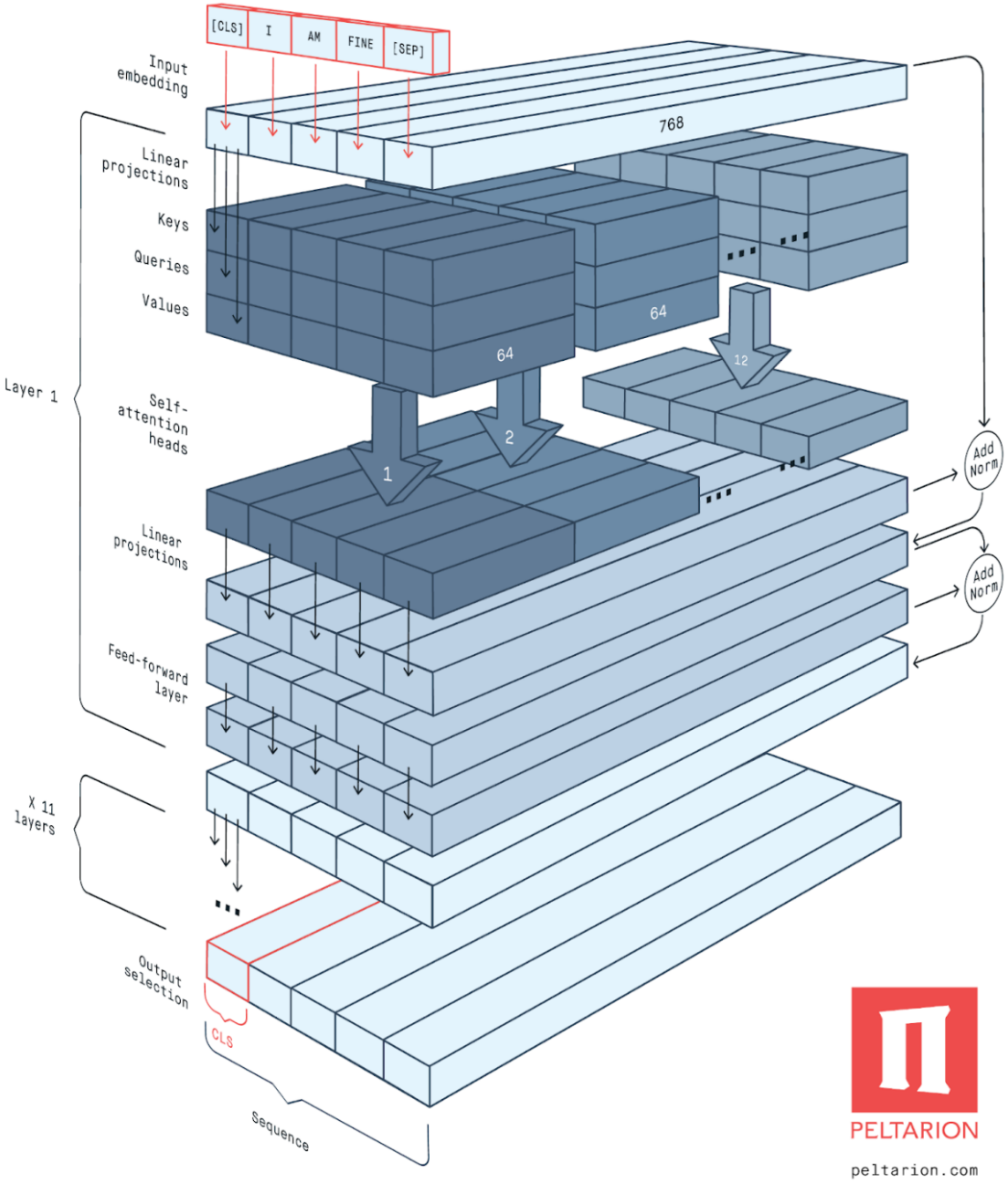


Figure 3.4: **Architecture of BERT.** 3D visualization of BERT architecture. Figure taken from https://peltarion.com/.

embedding vectors, can be fixed like with GloVe or Word2Vec, associating a single fixed embedding vector to each word of the vocabulary, or context-dependent like with LSTM, GPT-2 or BERT that modulate word embeddings depending on their context. It might be interesting to extract various embedding representations based on different subset of a network, as different subparts of the model might focus on different properties of the input. For example, in a multi-layer model, different layers might represent different level of information, eliciting a hierarchically structured network. Additionally, different neural

architectures might distribute information processing differently over their architectures.

It has been shown that the high dimensional vectors built by NLMs encode notions of semantics and syntax (e.g. Hewitt and Manning; Lakretz et al., 2019; Pennington et al., 2014), suggesting that latent representations can serve as rich descriptions of an input stimuli, going beyond hand-made characterizations. Thanks to the continuous nature of these embedded representations, it is possible to measure the syntactic or semantic similarity of words using *Euclidean distance* or *cosine similarity*.

Latent representations will be especially useful in the following chapters, when combined with linear encoding models (see Chapter 4), to probe complex and specific processes in the human brain, such as context-integration or the processing of semantic or syntactic information.

This Chapter introduced neural language models, presenting state-of-the-art models such as GloVe, LSTM, GPT-2 and BERT, and explaining how to build complex word representations (aka *word embeddings* or *latent representations*) from these models. These word embeddings can be used as rich predictors of brain activity, going beyond manually-derived features. The next chapter will explain how fMRI brain activity can be estimated using manually- or model-derived features.

4 - Mapping linguistic features to the brain: the encoding paradigm

This chapter introduces linear encoding models, and outlines their usefulness in understanding the neural bases of human cognition. The first section gives a broad overview of encoding models and motivates their use to study brain data. The second section describes the entire encoding process starting with a set of stimulus features, to the alignment of feature-derived predictors with fMRI brain data. Finally, the last section gives a more practical view, investigating model assessment, the impact of the number of time-points per subject on the encoding performance, as well as the determination of the linear model hyperparameters.

4.1 . What are encoding models and why do we use them?

Brain encoding models seek to predict brain response patterns from descriptions of the experimental conditions (Kriegeskorte and Douglas, 2019; Mitchell et al., 2008; Naselaris et al., 2011). More precisely, encoding models map the time-courses of one or several features (characteristics derived from some stimuli), to time-courses reflecting brain activity following the exposure to the same stimuli. The input features can be hand-engineered or derived from a computational model, such as a tree-parser or a neural network model for example.

How can one map a set of features to brain data when the dimensions may not have a one-to-one correspondence? An approach is to predict the raw measurements (Mitchell et al., 2008; Naselaris et al., 2011), i.e. to map the set of features to each brain channel measured - e.g. each neuron, EEG channel or fMRI voxel. Such analysis is called mass univariate analysis.

Theoretically, an encoding model is defined as a function ϕ that takes as input a matrix of features $\mathbf{X}$, and tries to best predict a matrix of channels $\mathbf{Y}$, so that

$$\|\phi(\mathbf{X}) - \mathbf{Y}\|^{\mathbf{2}}$$

is minimal. The encoding model will capture which linear combination of features best predict each brain channel. In practice, ϕ is assumed to be linear to reduce computational costs and to facilitate interpretation. Indeed, if we can linearly map the input features to a given brain channel, it means that the variations of the features are similar to the variations of the brain response channel. Following this observation, we assume that there is a high probability that the information encoded by the input features are also encoded by the well-fitted brain channels. Thus, one can learn about the information a given channel, voxel or neuron might represent if it is well fitted by a set of features that encode information you control. On the contrary, being able to map $\mathbf{X}$ onto $\mathbf{Y}$ with a complex non-linear function might increase the encoding performance, but would not tell us much about the relationship linking both variables, which is what we are interested in.

Overall, linear (brain) encoding models teach us about representational spaces and the location of the processes elicited in the brain.

4.2 . Aligning regressors with fMRI data

Upstream of the encoding model, $\mathbf{X}$ must be preprocessed to take into consideration the dynamic of the BOLD signal. The following describes how to preprocess $\mathbf{X}$ and introduces the General Linear Model as well as two estimation methods.

4.2.1 . Temporal alignment

Keep in mind that fMRI BOLD activity is a proxy of neural activity. Each voxel represents the indirect response, averaged both in time and space, of thousands of neurons. To model these average responses, that is the dynamics of the BOLD signal, we can leverage the BOLD signal linear time invariant relationship with neural responses. Time invariance means that time-shift affecting the stimulus will be similarly reflected on the BOLD response. Additionally, if neural response is scaled by a factor k, then the BOLD response is also scaled by k. Finally, linearity implies additivity: the BOLD responses of two events close in time will be the sum of the BOLD responses of the independent events. However it must be noted that BOLD linearity has its limits, especially in short range time periods when two events are very close in time. There are several methods to model the BOLD signal responses. In the following, two of them are detailed: the kernel approach and the FIR model.

The Kernel approach

The linear and additive time invariant essence of the BOLD signal naturally suggests to sum the effects of the independent events of the stimulus predictive time serie:

$$(h * f)(t) = \int h(\tau) f(t - \tau) d\tau$$

Friston et al. (1998) found that the BOLD response to a dirac stimulus was best modeled by a combination of two gamma functions: the first one modelling the shape of the initial stimulus response, while the second models the post-stimulus undershoot. This *double-gamma HRF* (h) is considered as the simplest modelling choice of the Haemodynamic response function, assuming that it is constant across brain regions and individuals.

The following gives details about the haemodynamic response function characteristics:

- *Peak height:* This is the most common feature of interest, since it is most directly related to the amount of neuronal activity in the tissue (Logothetis et al., 2001). For BOLD fMRI, the maximum observed amplitude is about 5% of the total signal, for primary sensory stimulation, whereas signals of interest in cognitive studies are often in the 0.1–0.5% range.

- *Time to peak:* The peak of the HRF generally falls within 4–6 seconds of the stimulus onset.

- *Width:* The HRF rises within 1–2 seconds and returns to baseline by 12–20 seconds after the stimulus onset.

- *Initial dip:* An initial dip in the BOLD signal occurring within the first 1–2 seconds has sometimes been reported; it is thought to reflect early oxygen consumption before changes in blood flow and volume occur. It is generally ignored in most models of fMRI data.

- *Poststimulus undershoot:* The HRF generally shows a late undershoot, which is relatively small in amplitude compared to the positive response and persists up to 20 seconds or more after the stimulus.

The FIR model

While the kernel approach presents the advantage of simplicity, its assumptions might not always be verified. In Handwerker et al. (2004), a study of the HRF shape revealed that both the time-to-peak and width of the HRF varied within subjects across different regions of the brain and across subjects, with inter-subject variability higher than intra-subject variability. An approach to better model the HRF is to use a set of HRF basis functions. Once these functions are linearly combined, they can defined a range of expected shapes for the haemodynamic response. The kernel approach previously described, is a special case with just one basis function. A more complex modelling of the HRF would add to the double-gamma basis function, its time-derivative to account for a slight temporal shift. One could furthermore learn several basis functions from the data. This is what the Finite Impulse Response (FIR) model attempts to perform. This flexible approach tries to learn the shape of the haemodynamic response function in each voxel, by stacking time-shifted versions of the stimulation signal to $\mathbf{X}$. The encoding model will then learn how a given feature contributes to the activation of a given voxel through time.

However, the flexibility of the FIR model to capture the shape of the HRF comes at the cost of an increase in the variability of the estimates (i.e., a bias–variance trade-off) as well as an increase in computation time and load that is non-negligible for models with a high number of regressors. While the bias about the shape of the HRF is reduced, the variability of our estimates increases since fewer data points are contributing to each parameter's estimate. Moreover, the variability of the estimates can also increase because of regressors' collinearity.

We only used the Kernel approach in all subsequent analyses for computational reasons.

4.2.2 . The General Linear Model (GLM)

Once the BOLD dynamic has been taken into consideration, $\mathbf{X}$ and $\mathbf{Y}$ can be aligned with the encoding model. The most common encoding model is the General Linear Model (GLM). It attempts to fit voxels' timecourses $\mathbf{Y}$ with a matrix of features $\mathbf{X}$, by learning a matrix of weights $\hat{\boldsymbol{\beta}}$, so that $\mathbf{Y} \approx \mathbf{X}\hat{\boldsymbol{\beta}}$. Where $\mathbf{Y} \in \mathbb{R}^{\mathbf{T}\times\mathbf{V}}$, $\mathbf{X} \in \mathbb{R}^{\mathbf{T}\times\mathbf{p}}$ and $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p\times V}$, with V the number of voxels, T the number of scans and p the number of predictors.

In traditional linear brain encoding analyses, one of the following two estimation methods are used.

Ordinary Least Squares (OLS)

The OLS is a type of linear least squares method for choosing the unknown parameters in a linear regression model. It minimizes the sum of the squares of the differences between the observed dependent variable (the observed voxels' time-courses) and the output of the linear function of the independent variables. For a given voxel, we have:

$$\hat{\boldsymbol{\beta}}^{v}_{OLS} = arg\min_{\boldsymbol{\beta}^{v}} \sum_{t=1}^{n} (y_t^v - \boldsymbol{\beta}^v \cdot \mathbf{x_t})^{\mathbf{2}} \tag{4.1}$$

Where $\hat{\boldsymbol{\beta}}^{v}_{OLS}$, $\boldsymbol{\beta}^{v} \in \mathbb{R}^{p}$, $\mathbf{x_t} \in \mathbb{R}^{\mathbf{p}}$ and $y^{v}_{t} \in \mathbb{R}$.

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y} \tag{4.2}$$

Where $\hat{\boldsymbol{\beta}}_{OLS} \in \mathbb{R}^{p \times V}$

OLS is generally used when the number of regressors is small.

The L2-Regularized Linear Model: Ridge

When there are many predictors or when they are highly correlated, the weights estimated with OLS are unreliable. An alternative is to use a Ridge encoding model which is a L2-regularized linear encoding model, with the L2-penalization applied on the weights of the encoding model. This regularization term is weighted by an hyperparameter α, whose determination is quite important to get an optimal fit (see next section). We obtain the following equations:

$$\hat{\boldsymbol{\beta}}^{v}_{Ridge} = arg\min_{\boldsymbol{\beta}^{v}} \sum_{t=1}^{n} (y^{v}_{t} - \boldsymbol{\beta}^{v} \cdot \mathbf{x_t})^{\mathbf{2}} + \alpha \|\boldsymbol{\beta}^{\mathbf{v}}\|^{\mathbf{2}}_{\mathbf{2}} \tag{4.3}$$

Where $\hat{\boldsymbol{\beta}}^{v}_{Ridge}$, $\boldsymbol{\beta}^{v} \in \mathbb{R}^{p}$ and $\mathbf{x_t} \in \mathbb{R}^{\mathbf{p}}$, $y^{v}_{t} \in \mathbb{R}$.

$$\hat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}^{\boldsymbol{T}}\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^{\boldsymbol{T}}\mathbf{Y} \tag{4.4}$$

Where $\hat{\boldsymbol{\beta}}_{Ridge} \in \mathbb{R}^{p \times V}$

Ridge regression is quite useful to overcome the limitations of the OLS method when using high dimensional model-derived features.

4.2.3 . Validating the alignment

Finally, once the encoding model has been trained, we have to assess the quality of the fit. To evaluate the encoding model performance, we used a cross-validation procedure: the data of each participant was split into a training and a testing set. The split was based on runs, i.e. distinct acquisitions. More precisely, for a participant having N runs of data, we used $(N-1)$ runs for training and 1 run for testing. We evaluated model performance on the test set, using the Pearson correlation coefficient (R), which is a measure of the linear correlation between models' predicted time-courses and the actual time-courses. It is defined as:

$$R(y, \hat{y})_{v,test} = \frac{\sum(\hat{y}^{v}_{t} - \bar{\hat{y}}^{v})(y^{v}_{t} - \bar{y}^{v})}{\sqrt{\sum(\hat{y}^{v}_{t} - \bar{\hat{y}}^{v})^{2} \sum(y^{v}_{t} - \bar{y}^{v})^{2}}},$$

$$\text{where } \bar{\hat{y}}^{v} = \frac{1}{T}\sum_{t=1}^{T} \hat{y}^{v}_{t} \quad , \quad \bar{y}^{v} = \frac{1}{T}\sum_{t=1}^{T} y^{v}_{t}$$

Pearson correlation coefficients were then averaged across splits for each subject, giving a cross-validated map of R values, i.e. one value for each voxel. The Coefficient of determination can also be used to evaluate the goodness of the fit. However, we chose to use Pearson correlation because we are interested in recovering the pattern of fluctuations across time and not the signal magnitude.

4.3 . Technical aspects of the encoding procedure

4.3.1 . Impact of the fMRI dataset size on the encoding performance

A known issue in neurosciences is the lack of good data. Acquiring neuroimaging data is vey expensive and complex. As a consequence, neurorimaging datasets are quite small compared to classical machine learning ones. Moreover, most of the time, data are noisy and high dimensional. For example, in a classic fMRI dataset, the number of dimensions (voxels), is usually on the order of 10^4 or 10^5, while the number of time-points for a given participant does not go beyond 10^4 for the largest datasets.

Noisy high dimensional data with few samples easily leads to overfitting when training linear encoding models. Therefore, it is important to encourage the acquisition and the use of large neuroimaging datasets.

Using word embeddings from a GloVe model (see Chapter 3) and the Shared Response Model introduced in Section 2.6, I present an analysis showing the impact of the number of scans on the encoding performance of linear encoding models. From 30 English subjects of *The Little Prince* fMRI corpus, I derived in a model-agnostic manner, using the Shared-Response Model (SRM, Chen et al. (2015)), the most "responsive" voxels. That is, the voxels whose R values were among the 50% highest ones. This set, which is referred to as "SRM50", contains 13,082 voxels.

Fig.4.1 displays the averaged R score across voxels (in the "SRM50") and subjects as a function of the number of scans used in the training dataset. It shows that using datasets with a small number of scans per subject can lead to poor encoding performance. Such dependency between the encoding model performance and the number of scans warns researchers about potential misinterpretations.

4.3.2 . Model assessment: Cross-validation

Cross-validation is a common approach to assess linear encoding model quality when having few and noisy high dimensional data. In practice, data are split into train and test sets. When working with N acquisition runs for example, one can usually keep 1 run for testing and leave the rest for training. Evaluation results are then averaged across splits, giving a better evaluation of the encoding model performance.

4.3.3 . Finding the hyperparameters: nested Cross-validation

When performing mass univariate ridge encoding analyses, the regularization parameter α has to be determined for each model fitted. To optimize the fit of the model, and to avoid overfitting, we designed a nested cross-validation scheme where the Pearson Correlations R_{test} are cross-validated in an outer loop, while the regularization parameter α is estimated in an inner cross-validation loop. More precisely, if we have N runs of fMRI data and p values for α, linearly spaced in log-scale, to test, we proceed as in the following for each voxel:

- $N-2$ runs were used to train encoding models for each α,
- 1 run was used to evaluate the encoding models, i.e. computing R_{valid},
- The previous two steps were repeated $N-1$ times by selecting a new validation run each time,

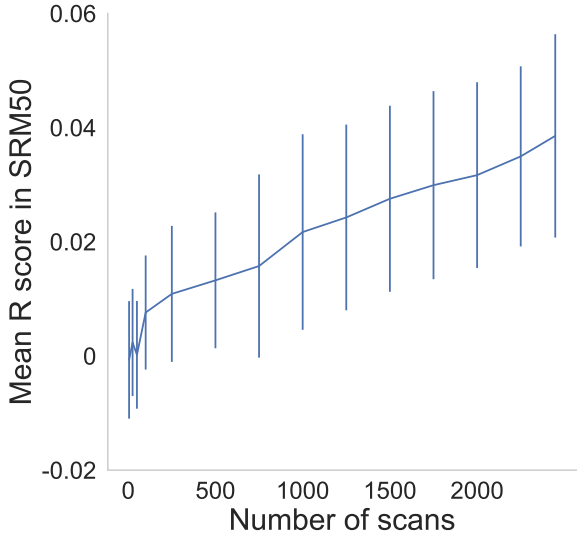


Figure 4.1: **Mean R score across subjects and voxels inside the "SRM50" voxel-set as a function of the number of scans in the training dataset.** 30 English subjects of *The Little Prince* fMRI corpus were used to fit Ridge encoding models. Each subject had 9 runs of 10 mins of data: the last run was set to be the test set while all other 8 runs were stacked and used for training. 300-dimensional word embeddings were derived from a neural language model (GloVe - see Chapter 3) and used to fit brain data. The number of scans (time-points) of the training dataset used to train the Ridge encoding model varied from 5 to 2448. Encoding models were evaluated using Pearson correlation. Finally, Pearson correlations were averaged across subjects and voxels. Error bars represent the standard deviations across subjects.

- The best α^* was selected, i.e. the one that had the highest R score in average across splits,
- An encoding model was then trained on the $N-1$ previous runs for the best alpha,
- Finally the encoding model was evaluated on the last left-out run, i.e. computing R_{test},
- The previous steps were repeated N times by selecting a new testing run each time.

The following investigates the link between the best alphas and the cross-validated R_{test}. Word embeddings derived from the 9-th layer of a BERT model were used to fit the brain data of the English participants, with α linearly spaced in log-scale between 10^{-3} and 10^4. We observed that the lower the alpha, the better the cross-validated R_{test} (Fig.4.2, left). This result is explained in the second panel (Fig.4.2, right) showing that voxels that are not involved in language processing, the voxels outside the SRM50 voxel-set, have a high α variance across runs and subjects (in red).

This Chapter introduced linear encoding models, how they can be used to probe the neural bases of language comprehension, and how a set of stimulus features should be processed and fitted to fMRI brain data. Importantly, this Chapter highlighted the importance of building and using large fMRI datasets, showing how the encoding model performance could greatly increase with the number of fMRI scans. This Chapter also gave insights on

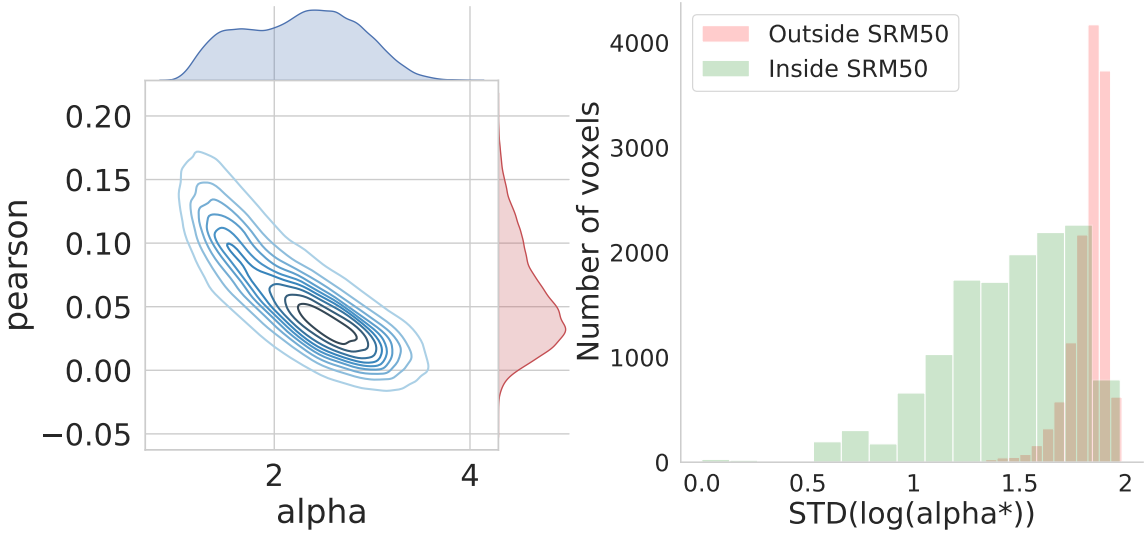


Figure 4.2: **Linking voxels' best alpha, R_{test} and involvement in language processing.** A) Density plot across voxels and subjects of the average R score as a function of the L2-penalization. 51 English subjects of *The Little Prince* fMRI corpus were used to fit a Ridge encoding models. Each subject had 9 runs of 10 mins of data: the last run was set to be the test set while all other 8 runs were stacked and used for training. We used 768-dimensional features extracted from the 9th layer of a BERT neural language model to fit brain data. Encoding models were evaluated using Pearson correlation. Finally, Pearson correlations were averaged across subjects and voxels. The R scores and alphas used in the inner-loops of the nested cross-validations are used. B) Standard deviation of the logarithm of the best alpha across subjects for all voxels. Voxels inside the SRM50 voxel-set are green, and voxels outside SRM50 are red.

the relationship linking the encoding model performance and the regularization hyperparameter, advocating to test for a large range of hyperparameters. The next Chapter will illustrate how NLM-derived features fit fMRI brain data, and how they can be used to probe the neural bases of language comprehension.

Part II

Contributions

5 - Mapping language features to The Little Prince fMRI brain data

In Chapter 4, we have seen that encoding models were able to relate a single continuous dependent, or response, variable to a set of continuous or categorical independent variables, or predictors. This chapter demonstrates how these encoding models can be used to map linguistic features to the brain. The first section briefly describes simple low-level linguistic features and the brain regions they map onto. The second section leverages neural language models' latent representations to investigate higher level language-related brain areas. In a second set of analyses, this section probes the interactions between the layered structure of transformer-based models and brain regions. Addressing questions such as: Does the layers of NLMs maps onto different brain regions? Are some brain regions better fitted by sub-modules of transformer models?

The investigation of the interactions between transformer-based models' architecture and brain regions is completed by a quantification of information redundancy across transformers' layers. That is, can fMRI brain data be better fitted by stacking the embeddings of consecutive layers compared to using only the embeddings from one layer. Finally, in a last analysis, I perform a fair comparison of NLMs' ability to fit fMRI brain data, controlling for model size, vocabulary size and the training data. Overall, this chapter is a first approach on how to use NLMs to study the brain.

5.1 . Mapping simple linguistic features to brain data

5.1.1 . Basic Features

Between the processing of the input audio signal and the construction of high-level complex representations, the brain processes various kinds of information, from phonology to lexical and supra-lexical (compositional) syntax and semantics. However, in parallel to processes of interest, the brain might activate due to some variables of non-interest.

For example, it has been shown that word's lexical frequency explains a significant amount of brain activations as rare words would create more surprise than frequent words. The intensity of the audio signal might also correlate with higher-level structures in the text such as sentence processing as the tone of the voice would decrease at the end of each sentence.

Unfortunately, there is no formal solution to the determination of all the possible variables of non-interest. As a consequence, we decided to define 3 variables of non-interest that we grouped under the name of *Basic Features* (BF):

- a) *the acoustic energy* (root mean squared, or *RMS*, of the audio signal sampled every 10ms),
- b) the *word-rate* which specifies the onset/offset of each word (one event at each word onset/offset),
- c) the *logarithm of the unigram lexical frequency* of each word, which modulates the word-rate by the logarithm of the unigram lexical frequency.

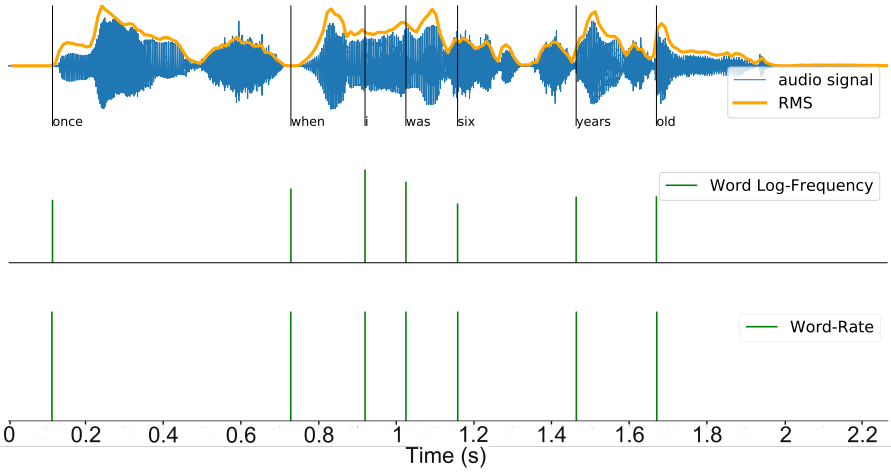


Figure 5.1: **Basic Features description.** From top to bottom are displayed the RMS, the words' LogFreq and the Word-Rate computed from a few seconds of The Little Prince stimuli. The auditory stimuli is represented in blue in the first plot.

Table 5.1: Mean correlations between the Basic Features across runs, after the convolution with the haemodynamic kernel.

Models	RMS	Word-Rate	Log-Frequency
RMS	-	0.40 (std=0.12)	0.31 (std=0.11)
Word-Rate	0.40 (std=0.12)	-	0.97 (std=0.01)
Log-Frequency	0.31 (std=0.11)	0.97 (std=0.01)	-

5.1.2 . Highlighting low-level processing brain regions

Taken together, these Basic Features gather the effects of the variables of non-interest that we want to remove when probing the comprehension of language in the brain and focusing on more specific signal of interest, such as the encoding of semantic content for example. More simply, we want to interpret the predictive power of the NLMs and how much their representational capabilities extend beyond these variables.

To understand how these variables of non-interest might affect our analyses, we address the following questions: How much fMRI signal do these variables of non-interest explain? Which brain regions do they fit?

We first determined the brain regions whose fMRI activations correlate with the time-courses predicted by these three predictors. To do so, we fitted a GLM between the fMRI brain data of each subject and each one of these Basic Features independently, before computing a map of cross-validated R for each participant and feature.

Panel A, B and C of Fig.5.2, display the group-level significant R scores for each variable of non-interest such that $p < .1$, corrected for multiple comparison with a Bonferroni correction (Bonferroni, 1936).

The R maps show quite widespread bilateral significant activations in the entire cortex, covering the classic perisylvian 'language network' (that includes the left IFG[1] and temporal lobe), with the highest correlations around Heschl's gyri. We also assessed the

[1]Brain regions' abbreviations are listed in Appendix A.1

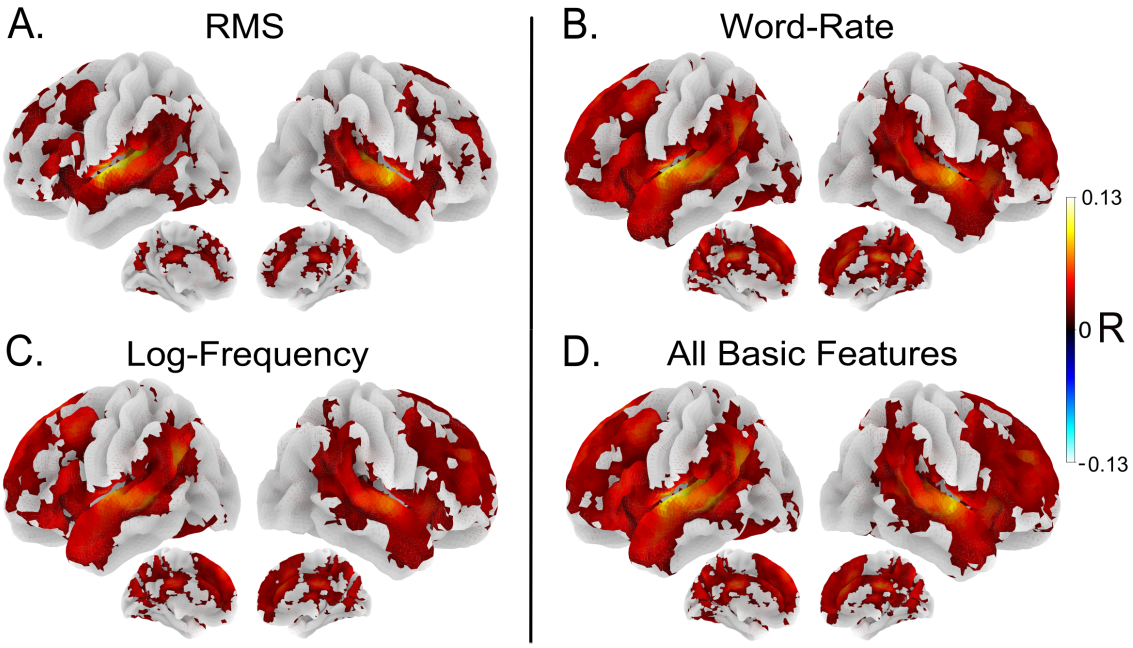


Figure 5.2: **Brain regions that are significantly fitted by the Basic Features.** Voxels showing significant activations across all participants (N=51, correction for multiple comparisons with a Bonferroni correction of 0.1), for A) the RMS of the audio signal, B) the Word-Rate, C) the Word Log-Frequency and finally when D) all the Basic Features are given to the encoding model.

cumulated effect of these three Basic Features taken together (Fig.5.2D). Fig.5.2 panel D displays the group-level significant R scores when the encoding model takes as input the stacked Basic Features, corrected for multiple comparison with a Bonferroni correction of 0.1. We observed correlation with brain data in all the language network and its surroundings, with peak activations bilaterally in Heschl's gyri with R values of 0.15. This shows that Heschl's gyri, where auditory processing is performed, are better fitted than other brain regions by our simple Basic Features.

To quantify the unique contribution of each Basic Feature to the prediction of the fMRI signal, we designed an additional contrastive experiment. We first estimated the Pearson correlation of each individual Basic Feature, then assessed the Pearson correlation obtained from the concatenation of subsets of the Basic Features. Finally, we looked at the increase in R scores relative to each individual Basic Feature when concatenating this specific Basic Feature with the other. Importantly, as the Word-Rate and word's Log-Frequency are highly correlated (see Table 5.1), we did not include them together in this contrastive analysis. Overall, we looked at:

- a) *the specific effect of RMS*, that is, the increase in R scores relative to the Word-Rate feature when concatenating RMS and Word-Rate features,
- b) *the specific effect of Word-Rate*, that is, the increase in R scores relative to the RMS feature when concatenating Word-Rate and RMS features,
- c) *the specific effect of Log-Frequency*, that is, the increase in R scores relative to the RMS feature when concatenating Log-Frequency and RMS features.

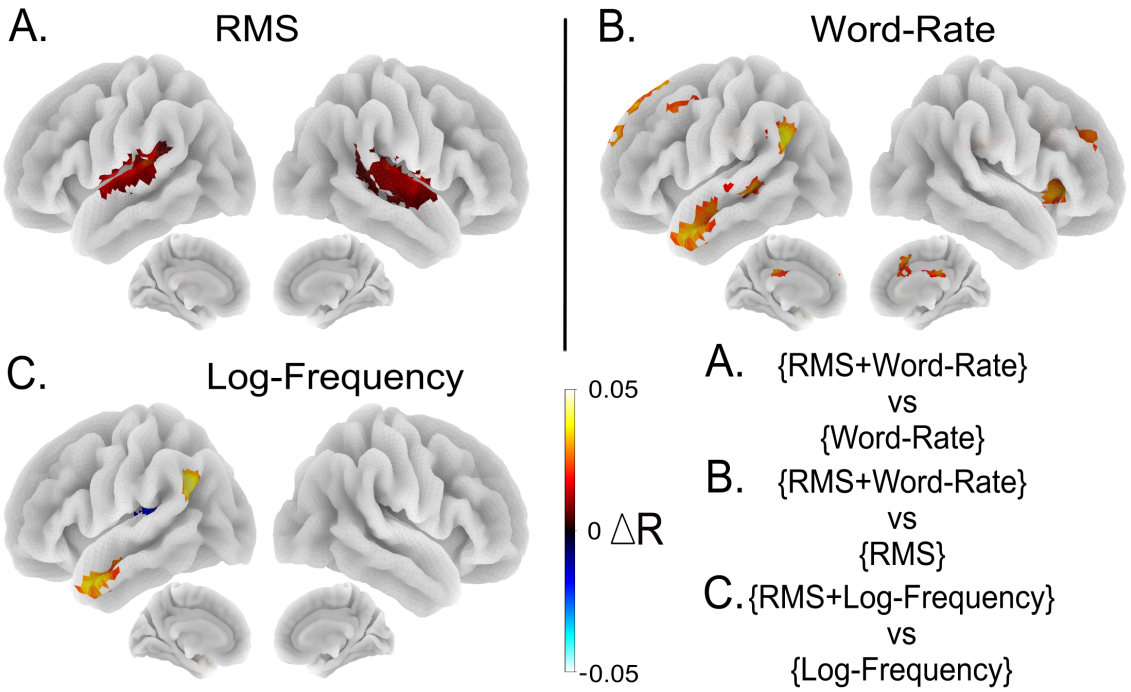


Figure 5.3: **Correlation uniquely explained by each Basic Features.** A) Increase in R scores relative to the Word-Rate feature when concatenating RMS and Word-Rate features. B) Increase in R scores relative to the RMS feature when concatenating Word-Rate and RMS features. C) Increase in R scores relative to the RMS feature when concatenating Log-Frequency and RMS features. As the Word-Rate and word's Log-Frequency were highly correlated (0.97), we only assessed their increase in R scores relatively to the RMS features.

We found specific regions for each Basic Feature considered. For the RMS predictor, the regions that are best fitted are the Heschl's gyri, where auditory processing takes place. For the Word-Rate, we found the Temporal Pole, the (anterior/middle) Superior Temporal Sulci and the Jensen sulcus. And finally, for words' Log-Frequency, the brain regions that are better fitted include the Temporal Pole and the Jensen sulcus.

In Chapter 6, we will use model-derived representations to study specific linguistic processes, namely syntax and semantics. To obtain a more accurate evaluation of the specific impact of variables of interest embedded in the model-derived representations, we will remove the contribution of the three Basic Features from all maps presented in Chapter 6. This implies that all R-maps presented there will be corrected for the contribution of these variables, that is they will display ΔR, the increase in R when adding a set of predictor features to the Basic Features versus the Basic Features by themselves.

5.2 . Mapping features derived from NLMs to the brain

5.2.1 . Highlighting language brain areas using NLMs latent representations

As described earlier, naturalistic stimuli such as story listening are not constrained by any task nor experimental design, but as a counterpart, it becomes difficult to isolate specific bits of information among high-dimensional representations. Drawing upon the developments in language models made by the machine learning community in recent decades, we opted to leverage the richness of the stimuli by using representations of the input text derived from these models. Going beyond human-made representations, neural language models (which are machine learning models specialized in the processing of textual data) build continuous representations of their input by mapping them to a dense and high-dimensional set of features (see 3.2).

As it was first shown in 2008 by the seminal work of Mitchell (Mitchell et al., 2008), and later confirm in many other studies (Huth et al., 2016; Jain and Huth, 2018; Toneva and Wehbe, 2019; Wehbe et al., 2014a), model-derived continuous representations are able to encode information that significantly explains fMRI signal in many brain regions. However, most of these studies focused on a given model such as co-occurrences models or LSTMs, giving no comprehensive insights on how the various existing NLMs fit brain data compared to one another. Thus, a few questions arise: How well can current NLMs fit fMRI brain data? Do these models better predict some brain regions than others?

To address these questions, we compared the fMRI predictive performance of GloVe, LSTM, GPT-2 and BERT (see 3.1 for more details on each model). We started by using publicly available pre-trained state-of-the-art models and used them to generate latent representations as predictors of brain activity. For GloVe, the non contextual model, we used the version shared by researchers from Stanford[2], for LSTM, the RNN, we used a version trained in Gulordava et al. (2018)[3], for GPT-2, the autoregressive transformer model, we used the 12-layer version provided by HuggingFace[4] and finally for BERT, the bidirectional transformer model, we used the uncased 12-layer version of HuggingFace[5].

While humans passively listened to the audiobook version of the novella, NLMs were provided with the exact transcription of this audiobook, enriched with punctuation signs from the written version of The Little Prince. Latent representations were then extracted from each architecture (see 3.2 for more details on the latent representations extraction procedure). We then used these latent representations to fit the fMRI brain data of each participant (see Fig. 5.4). Whole-brain, voxel-based, group analyses were performed, using one-sample t-tests applied to the individuals' R_{test} maps spatially smoothed with an isotropic Gaussian kernel with 6mm FWHM. In each voxel, the test assessed whether the distribution of R_{test} values across participants was significantly larger than zero. To control for multiple comparisons, all the maps displayed are corrected with a Bonferroni correction of $p < 0.1$. Brain maps are displayed in Fig.5.5, panels A-D.

As explained in 2.6, we derived in a model-agnostic manner from a Shared-Response Model (SRM, Chen et al. (2015)) the ceiling of explainable signal per voxel, that is the

[2]https://nlp.stanford.edu/projects/glove/
[3]https://github.com/facebookresearch/colorlessgreenRNNs
[4]https://huggingface.co/gpt2
[5]https://huggingface.co/bert-base-uncased

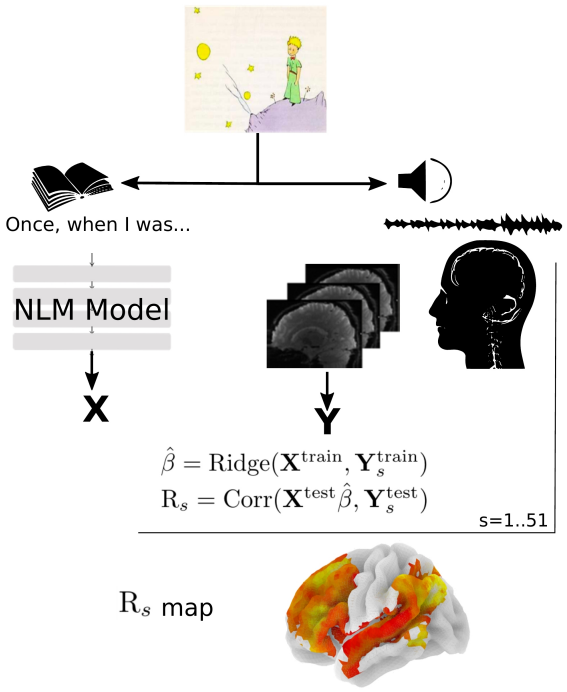


Figure 5.4: **Pipeline.** fMRI scans of human participants listening to an audio-book were obtained. The associated text transcription was input to Neural models, yielding embeddings that were convolved with an haemodyamic kernel and fitted to brain activity using a Ridge-regression. Brain maps of cross-validated correlation between encoding models' predictions and fMRI time-series were computed.

maximum R score that can be reached per voxel. We used these ceiling values to scale voxels' R scores to better visualize the quantity of signal explained by each NLM. We then displayed models' R scores distribution, as well as scaled R scores distribution in Fig.5.5, panel E.

Qualitatively, all models result into similar, widespread, activation patterns, but with intensity variations. The regions where the models best predict the signal mostly belong to the classic perisylvian 'language network'. They include the Superior and Middle Temporal Gyri, the Angular Gyrus (AG) and the Inferior and Middle Frontal Gyri. Interestingly, all NLMs capture a lot of activations in the right hemisphere, especially in the Superior Temporal Sulci. We observed that transformer-based models have higher R scores, and surprisingly, LSTM appears to performs the worst as shown with its darkened red colors (Fig.5.5, panel B) and its distribution in blue (Fig.5.5, panel E).

5.2.2 . Investigating interactions between model's architecture and brain regions

Mapping individual layers' features to brain data

In the previous analysis, we used the stacked latent representations from all layers and highlighted an ensemble of brain regions involved in the processing of language. However, this does not bring any new neuroscientific insight. One would like to be able to use these powerful machines that are NLMs to learn about the brain, about its dynamic and its organization. For example, a question that has received an increasing interest in the

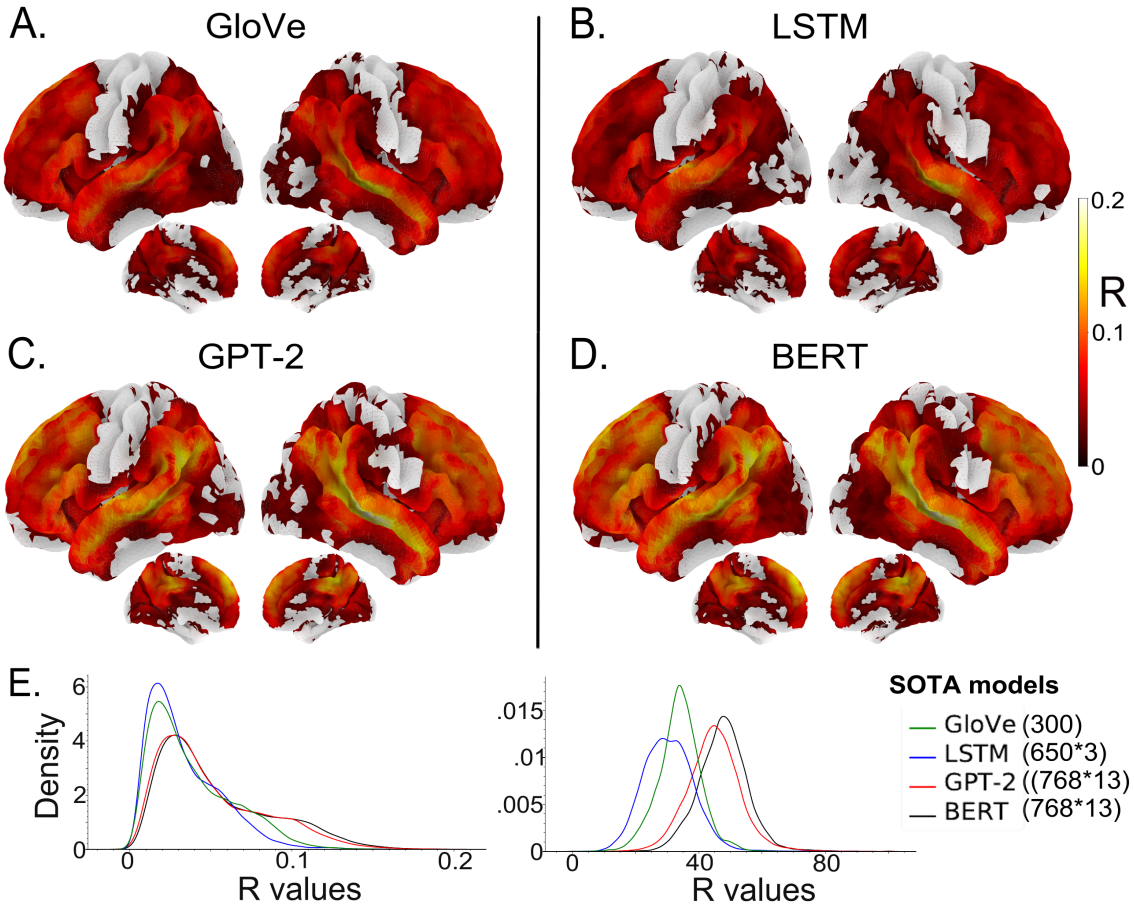


Figure 5.5: **Brain fit of pre-trained SOTA NLMs.** Voxels showing significant R scores for A) GloVe, B) LSTM, C) GPT-2 and D) BERT. E) R scores distributions for all models (left), and the same distributions scaled by the SRM-based ceiling of explainable signal (right).

last decades is whether there is a hierarchy of brain areas processing language. Several studies, such as Lerner et al. (2011) or more recently Chang et al. (2022), highlighted such hierarchical processing during language comprehension. A question that arises from the ability of NLMs to model brain data, is whether NLMs also elicit a hierarchical processing of textual information. Is there an interaction between model's architecture and brain regions? Can we find the hierarchy found by Lerner et al. (2011) using NLMs?

To address these questions, we extracted the latent representations of each layer of the 12-layer SOTA transformer-based models, and fitted an independent encoding model per layer, for each participant. We then used the Harvard-Oxford atlas (Desikan et al., 2006) (see 2.7) to order regions of interest depending on their brain score[6] and color-coded each ROI as shown in Fig.5.6, panel D.

While vision models like AlexNet elicit a hierarchical processing of their visual input similarly to the human visual system, NLMs do not elicit hierarchical processing, but a continuous transformation process. As shown in Fig.5.6 panel A and B, the brain score associated with each layer of the transformer-based models increases up to the late middle layers (layer 9-10 for a 12-layer model). Regression performance across layers is similar

[6]Brain score: median R score of a set of voxels.

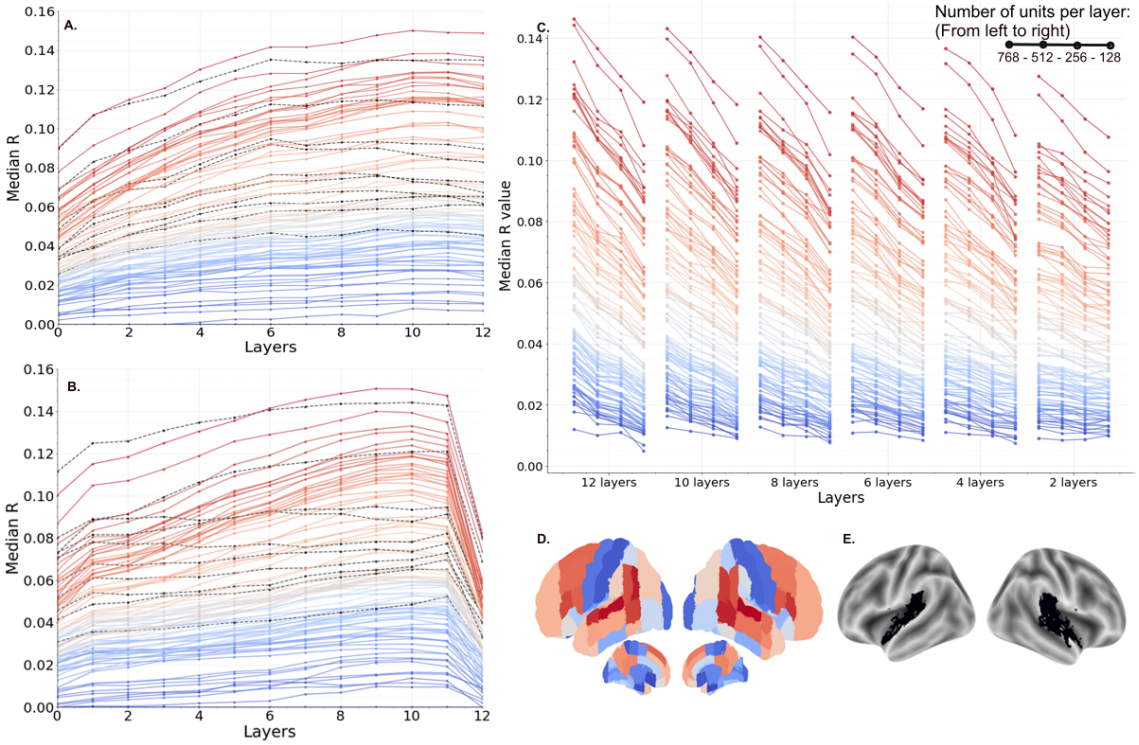


Figure 5.6: **Brain score per Region of Interest for various transformers-based encoding models.** A) Impact of layer depth on the, per-region, predictive power of BERT (resp. GPT-2 in B)). Brain scores (median R values) were computed across voxels inside brain regions defined by the Harvard-Oxford atlas; each line corresponds to a region. C) Impact of the number of layers and the number of units per layer on the predictive power of BERT. We displayed the median brain score across ROIs for different versions of BERT having a varying number of layers and a varying number of units per layers (768, 512, 256, 128). D) Color-coding scheme based on brain score (Red: highest brain scores, Blue: weakest brain scores, Light colours are in between). E) Brain regions involved in acoustic processing (Heschl's gyri and STG).

between BERT and GPT-2, increasing with layer depth and reaching a plateau (except for GPT-2 layer 12).

We tested whether it held when varying the number of layers in a BERT model, by replicating the same analysis with several versions of BERT[7] (see Fig.5.7). We corroborated the previous observations with these variants of the BERT model, showing a shared pattern of information processing and information distribution over the model architecture. We also verified that language brain regions are better fitted by the latent representations of NLMs than other brain regions, as shown in Fig.5.6 and Fig.5.7 where they appear in darker red.

Additionally, we looked at the impact of the number of layers and of the number of units per layer on the median brain score across ROIs (for the Harvard-Oxford atlas) in Fig.5.6 panel C. We used several versions of BERT, varying the number of layers (12, 10, 8, 6, 4), and the number of units per layer (768, 512, 256, 128) and observed that brain regions are ordered similarly for all models used. Finally, in Fig.5.8, we displayed the

[7]made available by GOOGLE at https://github.com/google-research/bert

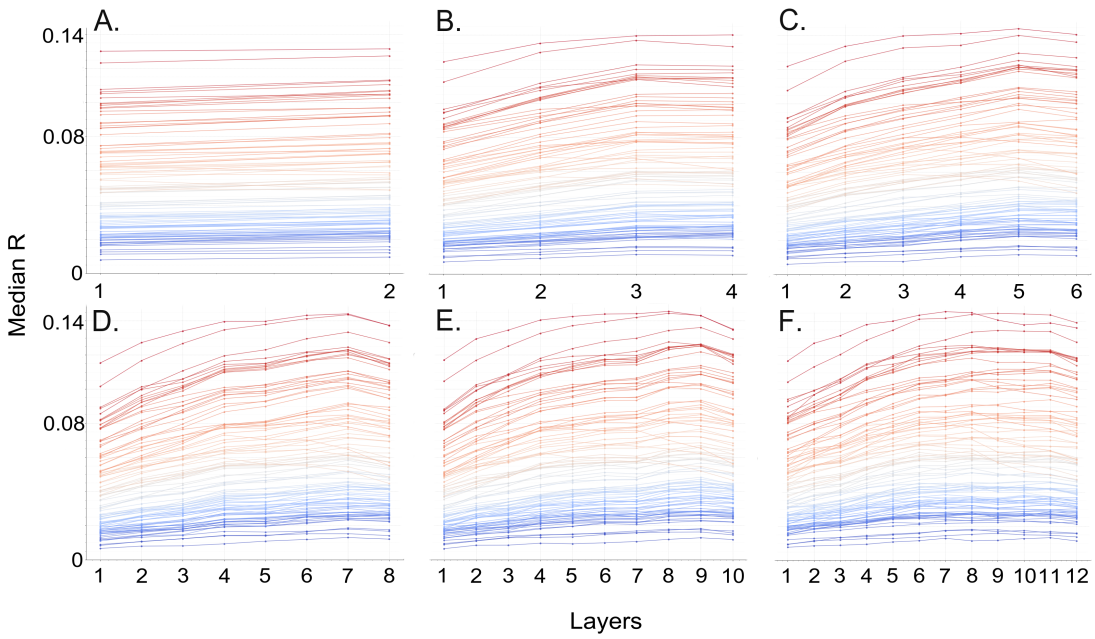


Figure 5.7: **Impact of layer depth on the, per-region, predictive power of BERT models having different total number of layers.** Impact of layer depth on the, per-region, predictive power of BERT models. A) 2-layer BERT, B) 4-layer BERT, C) 6-layer BERT, D) 8-layer BERT, E) 10-layer BERT, F) 12-layer BERT. Brain scores (median R values) were computed across voxels inside brain regions defined by the Harvard-Oxford atlas; each line corresponds to a region.

3rd quartile of voxels R score distribution as a function of the number of units per layer, showing that, for a given total number of units in a BERT model, it is better to have less layers but more units per layer.

Overall, these results suggest that there is no one to one correspondence between the layers of text transformer models and specific brain regions.

Pushing further the investigation of the interactions between model's architecture and brain regions, I conjectured that the interaction might not be visible at the layer-level but rather at the attention heads-level. To probe the existence of an interaction at the attention head level, we fitted linear encoding models on the output hidden states of each attention head (see Fig.5.9) of a 12-layer BERT (and GPT-2) model, and displayed the R score third quartile per ROI of the Harvard-Oxford atlas. We observed no specialization of attention heads to specific ROIs. In fact, if we ordered ROIs based on the fitting performance of the attention heads, they would be ordered in the same way for all attention heads as shown by the gradient of colours from pink to dark blue. Interestingly, some attention heads seem to pop up, having better average fitting performance than others. This hints at clever ways to reduce the number of features extracted from a given NLM in order to fit fMRI brain data (by selecting some attention heads). Overall, we confirmed the absence of interaction

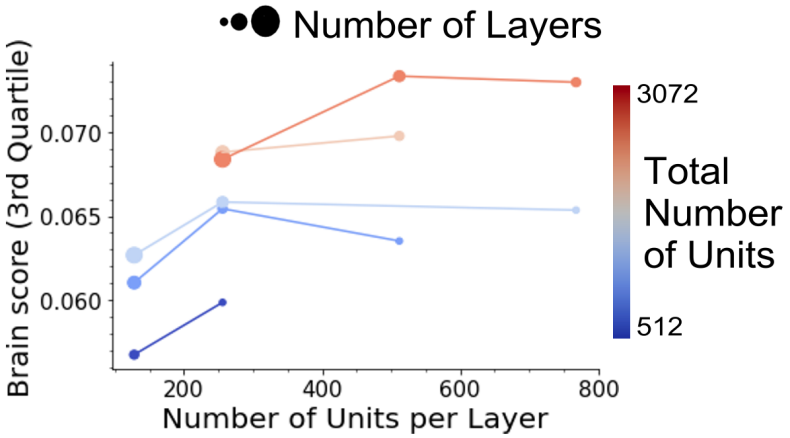


Figure 5.8: **Impact of the number of layers and the number of units per layer on the predictive performance of BERT.** Voxels' 3rd quartile as a function of the number of units per layer. Each dot is a BERT model. Each line links the models having the same total number of units. The total number of layers in each model is indicated by point size.

between model's architecture and brain regions;

Information redundancy across layers

Investigating further the absence of hierarchical processing in transformer models, we looked at information redundancy across layers. More precisely we studied how much we could gain in R score when using two consecutive layers stacked together to fit brain data compared to using just one of them.

We found that the latent representations of two consecutive layers were very similar and that using these two compared to using only the one closest to the late middle layers led to a decrease in fitting performance (see Fig. 5.10). Additionally, BERT and GPT-2 exhibit two different patterns of information redundancy across layers. This result suggests that the continuous transformation applied across layers is slow and favors information redundancy. Nonetheless, two layers that are far from each other will have less redundancy, leading to greater gains when using both of them in the regression model (see Supplementary Information - Fig.A.1).

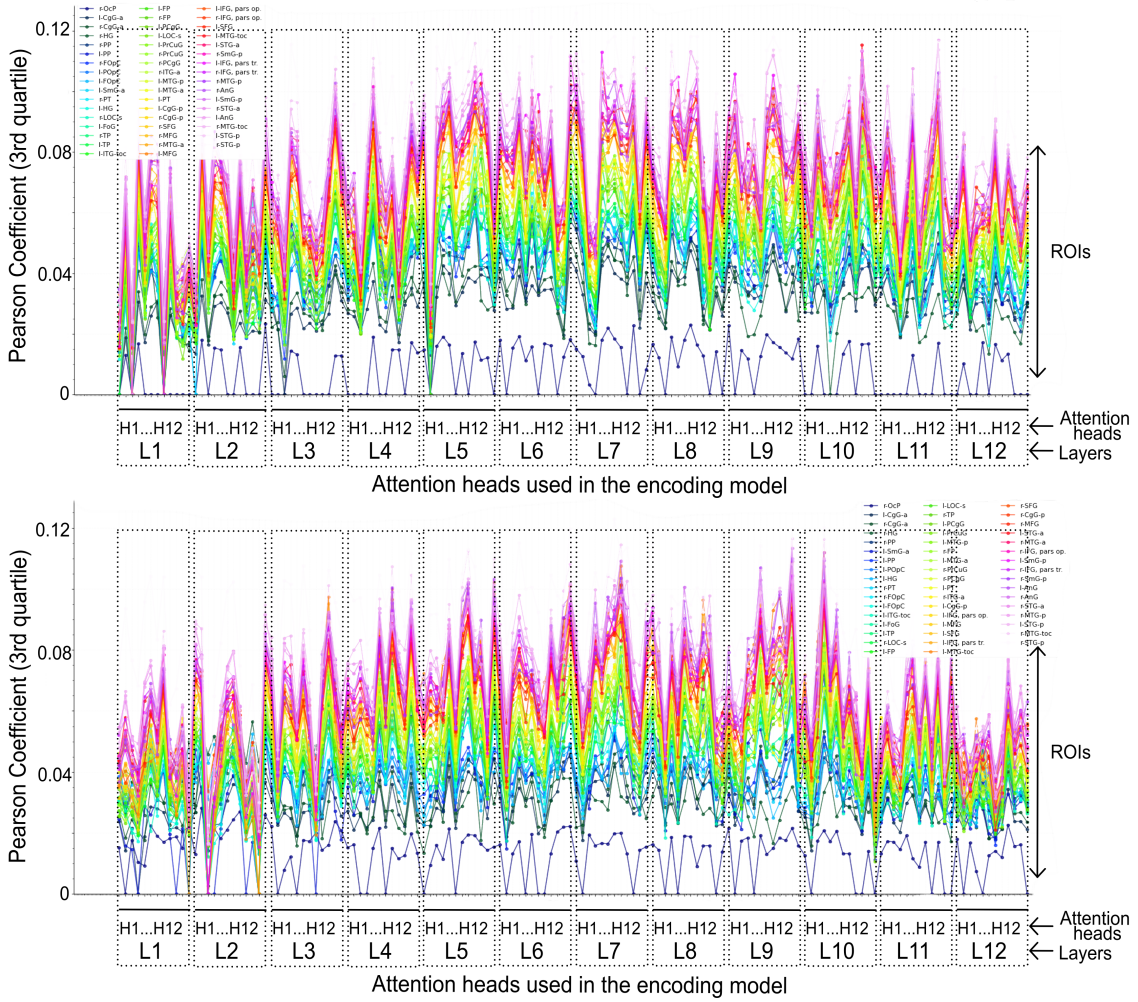


Figure 5.9: **Fitting performance of attention heads per Region of Interest.** The per-region, predictive power of BERT attention heads (top) and GPT-2 (bottom), reveals the absence of interaction between model's architecture and brain regions. The third quartile of R scores were computed across voxels inside brain regions defined by the Harvard-Oxford atlas; each line corresponds to a region.

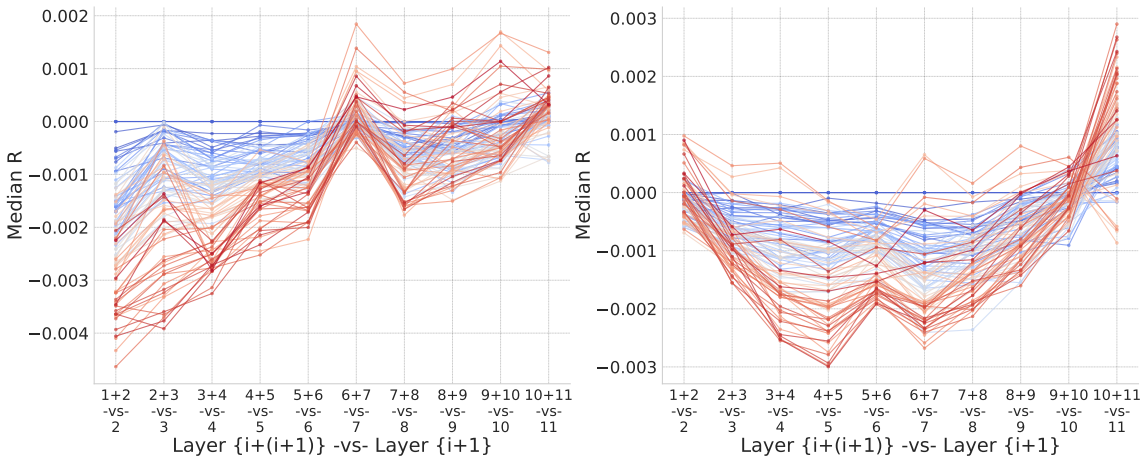


Figure 5.10: **Information redundancy across consecutive layers in transformers.** Per ROI increase in R scores relative to the embeddings of layer **i** of BERT (left) and of GPT-2 (right) when adding the embeddings of layer **i-1** in the encoding model.(the parcellation used was the Harvard-Oxford atlas) Each line corresponds to a region.

5.3 . Comparing NLMs trained on the same dataset

The former comparison of SOTA models is unfair in the sense that we were comparing state-of-the-art 12-layer transformer models trained on billions of tokens to small GloVe and LSTM models. The fact that LSTM was only trained on Wikipedia[8], while GloVe was trained on both Wikipedia and Gigaword[9] might explain the observed difference between the two.

5.3.1 . Creation of the *Integral Dataset*

We selected a collection of recent English novels (after year 1900) from Project Gutenberg (www.gutenberg.org; data retrieved on February 21, 2016) to create the *Integral Dataset*, a corpus of text on which to train neural language models. This *Integral Dataset* comprised 4.4GB of text for training purposes, 1.1GB for validation and 1.1GB for testing. We made sure that 'The Little Prince' novella was not part of the training dataset. The Integral Dataset (train, test and dev) is available at: `https://osf.io/jzcvu/`.

5.3.2 . Fitting fMRI brain data with NLMs trained on the *Integral Dataset*

We trained several versions of each architecture (details on the training procedure in 3.1):

- GloVe: 2 versions trained with embedding sizes of 768 and 1536 respectively.
- LSTM: a 1-layer model and a 4-layer model, all having a hidden layer dimension of 768.
- GPT-2: idem.
- BERT: idem.

Once we have controlled for the vocabulary size (set to 50000) and the data on which the models were trained, we need to control for the number of features extracted from each model. Indeed, we do not want to compare the predictive performance of two models from which we have extracted a different number of predictors.

In the first experiment (see Fig.5.11), we used the entire state of each model. This means that we used the activations of all the units of each model as predictive features in the encoding models. We then looked at the fitting performance of Glove (with embedding vectors of size 1536), and the 1-layer LSTM, GPT-2 and BERT. For LSTM, GPT-2 and BERT, we used the embeddings from the embedding layer (768 units) and from the unique hidden layer (768 units).

In this scenario, BERT and LSTM outperform GPT-2 and GloVe, suggesting that small size BERT and LSTM are already relatively good unlike GPT-2 whose performance suffered. This result also suggests that the poor performance of LSTM (in the SOTA experiment) was mainly due to the small dataset it was trained on. Interestingly, it is worth noticing that the R scores of our custom BERT and LSTM are close to the ones of the 12-layer SOTA transformers, which might be surprising given the models' size difference.

[8] http://dumps.wikimedia.org/enwiki/20140102/
[9] https://catalog.ldc.upenn.edu/LDC2011T07

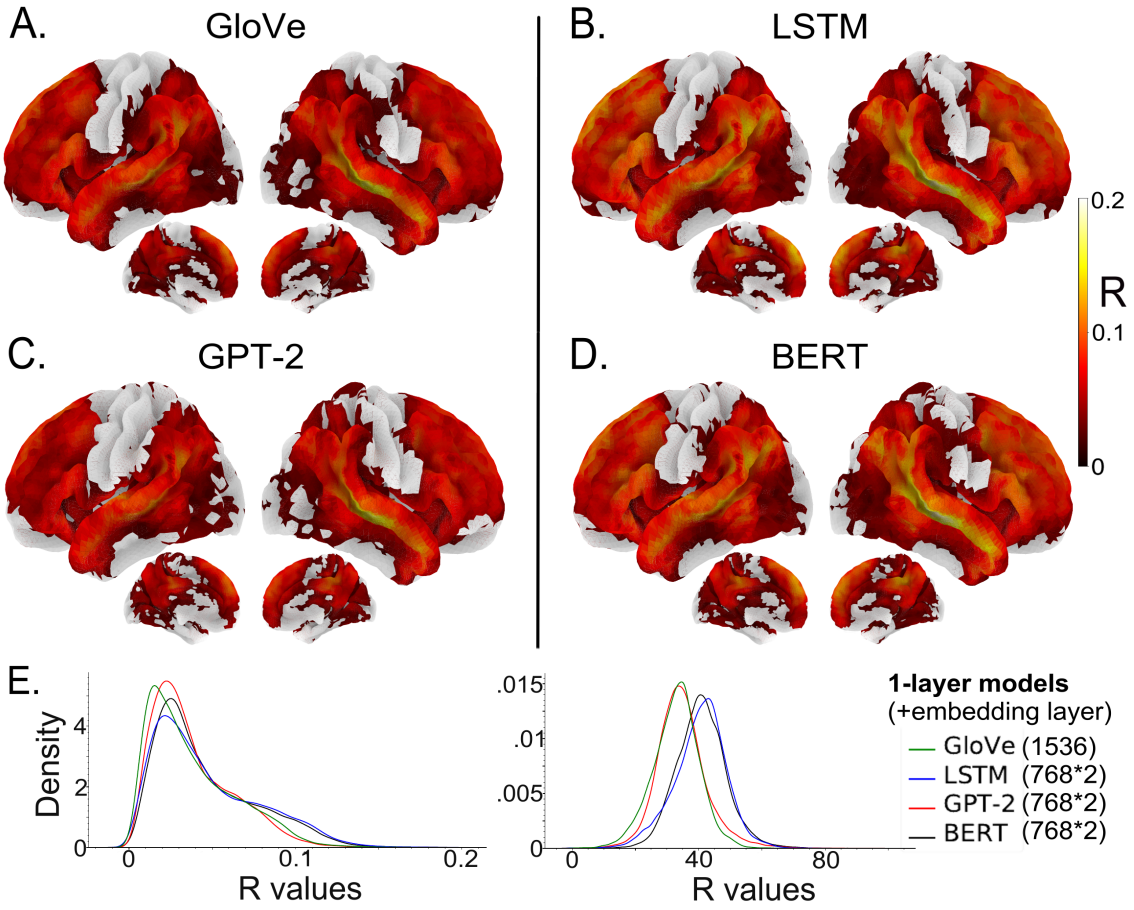


Figure 5.11: **Brain fit of 1-layer NLMs.** Voxels showing significant R scores for A) GloVe (d=1536), B) LSTM, C) GPT-2 and D) BERT. For B, C, D, the models were trained on the custom dataset derived from Project Gutenberg, with an architecture of 1 hidden layer (+ an embedding layer). For these three models, the entire state of the model was used to fit fMRI brain data. E) displays R scores distributions for all models (left), and the same distributions scaled by the SRM-based ceiling of explainable signal (right).

In the second experiment (see Fig.5.12), we used the embeddings extracted from a single layer of each model. For GloVe, we used the version with embedding vectors of size 768. For LSTM, GTP-2 and BERT, we used the third layer of the 4-layer models.

We observe that GloVe's R scores did not change much from halving the number of features. However, although we are using half the number of features compared to the previous experiment, the other three models benefited from the model size scaling, as their fitting performance improved. Especially for GPT-2 which elicited the biggest increase. This suggests that bigger models learn better representations.

In the end, the transformer-based models outperform both GloVe and LSTM, and are able to explain, in average, 50% of the explainable signal in fMRI brain data.

5.4 . Discussion

In this chapter, we have used the encoding approach to map linguistic features onto the brain.

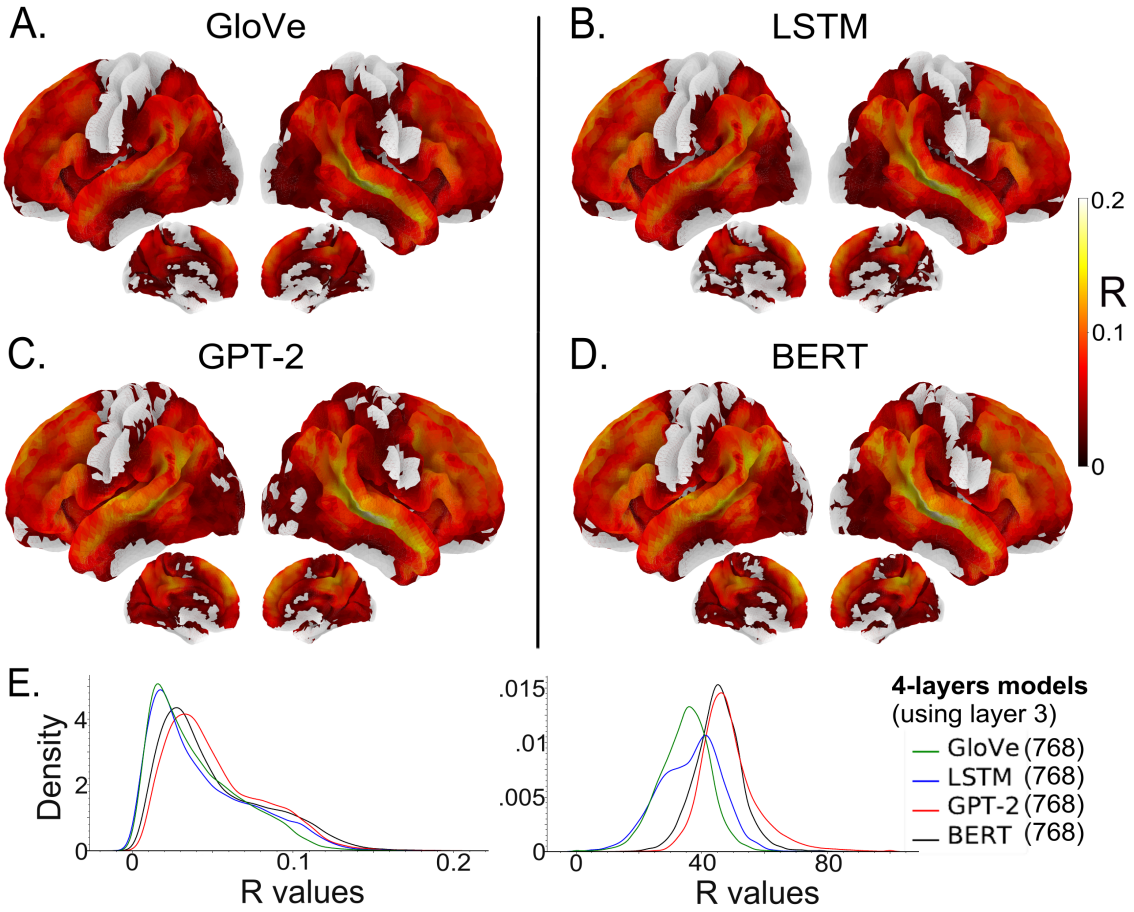


Figure 5.12: **Brain fit of 4-layer NLMs.** Voxels showing significant R scores for A) GloVe (d=768), B) LSTM, C) GPT-2 and D) BERT. For B, C, D, the models were trained on the custom dataset derived from Project Gutenberg, with an architecture of 4 hidden layers (+ an embedding layer). For these three models, only the hidden-states extracted from the third layer were used to fit fMRI brain data. E) displays R scores distributions for all models (left), and the same distributions scaled by the SRM-based ceiling of explainable signal (right).

We first observed that variables of non-interest could explain a lot of fMRI brain signal and that they should be taken into consideration when probing precise dimensions of language comprehension. One could model variables of non-interest even further, by modeling phonemes, which are the smallest units of sound that can discriminate a word from another in a particular language. However, as we investigated how NLMs' activations correlate with voxels timecourses, and as the scale of phonemes occurrences was smaller than the sampling rate of those NLMs from which we extracted word-level activations, we decided to disregard them.

The encoding paradigm has often been used to map simple manually-derived features to brain data. Here we used it to match more complex high-dimensional model-derived features. Using the latent representations extracted from NLMs (GloVe, LSTM, GPT-2 and BERT), we ran a comparative benchmark giving a relatively broad overview of NLMs ability to fit brain data, finding that transformer-based models are better compared to LSTM and GloVe to fit fMRI brain data. Importantly, NLMs were compared while con-

trolling for the training data as well as vocabulary size. NLMs' continuous representations explain even more signal than manually derived features and fit best the brain regions known to be involved in language processing. Corroborating results from (Huth et al., 2016; Wehbe et al., 2014a), we found that NLMs latent representations explain signal bilaterally in a wide network of brain regions, covering temporal, frontal and parietal regions as well as medially the Dorso-Medial Prefrontal Cortex, the Precuneus and the Cingulate and Para-cingulate gyri. An interesting neuroscientific observation is the fact that some of the highest correlations are located in the right hemisphere, that has been reported under naturalistic conditions. Additionally, comparing various instances of the different model architectures, we saw the importance of the training data and model size to perform a fair comparison. We will develop this point in more detail in Chapter.7. From all these comparisons, one point pops up: BERT has higher R values, followed by GPT-2, LSTM, and then GloVe. Schrimpf et al. (2020) did an even larger model benchmark, however, models were taken off-the-shelf, not trained on the same dataset, nor having the same vocabulary size which bias the comparison. The fact that transformers outperform LSTMs is even more surprising when knowing that LSTMs have more trainable parameters for the same vocabulary size and hidden dimension as shown in 3.1. This suggests that attention mechanisms are clearly an improvement over artificial incremental recurrent neural networks, allowing to better mix contextual information into words' embeddings. It is also intriguing to see that the non biologically plausible BERT[10] outperforms GPT-2.

A possibility is that for a given word, BERT is able to build a better word embedding by infusing information about the following words in the sentence, which might better reflect the fact that each fMRI scan is spread over 2s, mixing the activations related to an average of 10 words.

Interestingly, we observed that the R scores of the custom 4-layer BERT model are really close to the ones of the 12-layer SOTA transformers. Even the R scores achieved by specific attentions heads are fair. This result suggests that the quest for bigger and bigger NLMs will not bring us closer to understanding the brain, as opposed to focusing on smaller architectures with well-thought structure, training data and training objective. To learn about the brain using NLMs, one need to control for the information learnt by the models and remove any confound or variable of non-interest that could disturb the interpretation. The next chapter (Chapter 6) will give more details on how to control for the information learnt by NLMs. A convincing example was the fact that we could not highlight *hierarchical processing*[11] in the brain, because there was none in NLMs. A notable issue is that NLMs work in discrete space: we give them a sequence of tokens as input, while it should be a wave, or an image, like when we experience language in a naturalistic environment. Millet et al. (2022); Vaidya et al. (2022) recently tried to go beyond the discrete-input issue, by feeding the auditory input signal to a Wave2vec (2.0) (Baevski et al., 2020) model, and they show that this more biologically-plausible model elicited a hierarchical processing, with early middle layers fitting best auditory brain regions and late layers fitting best higher-level brain regions. This supports the hierarchy highlighted by Lerner et al. (2011) who scrambled the audio stimuli at different timescales corresponding

[10]BERT is considered to be non biologically plausible because it can leverage future context information.

[11]That is, we could not highlight brain regions processing linguistic structures of increasing complexity, like words, sentences, paragraphs, ...

to different linguistic structures (phonemes, words, sentences, paragraphs).

6 - Controlling NLMs feature space to probe syntactic and semantic processing in the brain

This thesis chapter originally appeared in the literature as Information-Restricted Neural Language Models Reveal Different Brain Regions' Sensitivity to Semantics, Syntax and Context
Alexandre Pasquiou, Yair Lakretz, Bertrand Thirion, Christophe Pallier, *under review*

The previous chapter used NLMs-derived latent representations to explain fMRI brain signal, and highlighted brain regions involved in language processing. However, because little was known about the information encoded in those latent representations, little could be said on the role of these regions, limiting its usefulness for the neuroscientific community. Here we would like to use these NLMs to address fundamental questions in neurolinguistics. For example, a central puzzle concerns the brain regions involved in syntactic and semantic processing during speech comprehension, both at the lexical (word processing) and supra-lexical levels (sentence and discourse processing). Using controlled GloVe- and GPT-2-derived latent representations, this chapter addresses the following question: To what extent are the brain areas involved in semantic and syntactic processing separated or intertwined?

The first section explains how to train both models in order to generate controlled latent representations, called *information-restricted representations*. The second section uses these *information-restricted representations* to probe semantic and syntactic processing in the human brain. Overall, this chapter shows that the use of information-restricted NLMs reconciles prior views on the spatial organization of syntactic and semantic processing.

6.1 . Controlling latent representations: information-restricted NLMs

6.1.1 . Crafting the feature space

As detailed in Chapter 5, we crafted a training dataset, named the *Integral Dataset*, by selecting a collection of recent English novels from Project Gutenberg (www.gutenberg.org; data retrieved on February 21, 2016). From it, we created two *information-restricted* datasets: the *semantic dataset* and the *syntactic dataset*. In the *semantic dataset*, only content words were kept, while all grammatical, function words and punctuation signs were filtered out. In the *syntactic dataset*, each token (word or punctuation sign) was replaced by an identifier encoding a (POS, Morph, NCN) triplet, where POS is the Part-of-speech computed using Spacy(Honnibal and Montani, 2017), Morph corresponds to the morphological features obtained from Spacy and NCN stands for the Number of Closing Nodes in the parse tree, at the current token, computed using the Berkeley Neural Parser(Kitaev and Klein, 2018) available with Spacy.

In this work, we refer to the content of the integral dataset as *integral features* the content of the semantic dataset as *semantic features* and the content of the syntactic dataset as *syntactic features*. Examples of integral, semantic and syntactic features are given in Table 6.1. The Integral Dataset (train, test and dev) is available at: `https://osf.io/jzcvu/`.

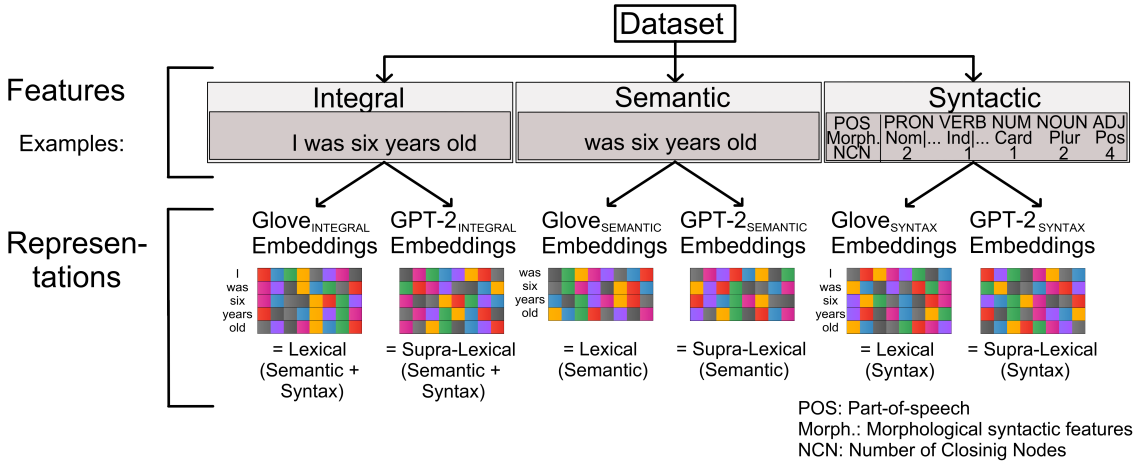


Figure 6.1: **Linguistic manipulations.** A corpus of novels was used to create a dataset from which we extracted three different sets of features: (i) *Integral features*, comprising all tokens (words+punctuation); (ii) *Semantic features*, comprising only the content words; (iii) *Syntactic features*, comprising syntactic characteristics (Part-of-speech, Morphological syntactic characteristics, Number of Closing Nodes) of all tokens. GloVe and GPT-2 models were trained on each feature space.

The semantic and syntactic datasets can be derived from the Integral Dataset using the scripts provided in `https://github.com/AlexandrePsq/Information-Restrited-NLMs`.

6.1.2 . Model training

GloVe (Global Vectors for Word Representation) relies on the co-occurrence matrix of words in a given corpus to generate fixed embedding vectors that capture the distributional properties of the words Pennington et al. (2014). Using the open-source code provided by Pennington and al. [1], we trained GloVe on the three datasets (integral, semantic and syntactic), setting the context window size to 15 words, the embedding vectors' size to 768, and the number of training epochs to 23.

GPT-2 (Generative Pretrained Transformer 2) is a deep learning transformer-based language model. We trained the open-source implementation GPT2LMHeadModel, provided by HuggingFace (**?**), on the three datasets (integral, semantic and syntactic).

The GPT2LMHeadModel architecture is trained on a next-token prediction task using a CrossEntropyLoss and the Pytorch Python package(Paszke et al., 2019). The training procedure can easily be extended to any feature type by adapting both vocabulary size and tokenizer to each vocabulary. Indeed, the inputs given to GPT2LMHeadModel are ids encoding vocabulary items. All the analyses reported in this chapter were performed with 4-layer GPT-2 models having 768 units per layer and 12 attention heads. As shown in (Pasquiou et al., 2022), these 4-layer models fit brain data nearly as well as the usual 12-layer models. We presented the models with input sequences of 512 tokens, and let the training run for 5 epochs; details about the morphological features (see A.4.1) and convergence assessments (see Fig.A.4.2) are provided in Appendix A (Chapter A).

[1] https://nlp.stanford.edu/projects/glove/

Table 6.2 summarizes the models used to probe syntax and semantics in the brain. They are organized according to the data (features) used for training and the level of information they encode, that is at the *lexical* or at the *supra-lexical* processing levels.

6.1.3 . Removing all residual syntax when training GPT-2 on semantic features

Small modifications had to be made to the model architecture of the GPT-2 version trained on the semantic features, in order to remove all residual syntax. By default, GPT-2 encodes the absolute positions of tokens in sentences. As word ordering might contain syntactic information, we had to make sure that it could not be leveraged by GPT-2 by means of its positional embeddings, yet keeping information about word proximity as it influences semantics. We achieved it by slightly modifying the architecture of GPT-2: we first removed the default positional embeddings, and added to the attention scores embeddings encoding relative positions between input tokens.

Indeed, only removing positional embeddings would have led to a bag-of-words model. Adding these embeddings encoding relative position to the attention scores induces tokens to weight the attention granted to another token depending on their distance. By doing so, information about absolute and relative positions is removed from tokens' embeddings. The following explains how this operation was performed.

Let $\mathbf{c}_W = (c_{w_1}, \ . \ . \ . \ , c_{w_m})$ be a sequence of m tokenized content words. $\mathbf{c}_W$ is fed to a n_{layers} transformer with n_{heads} of dimension d_{heads} that first build an embedding representation $\mathbf{E}_i, i = 1..m$ (of size $d = d_{heads} \times n_{heads}$) to which it appends (by default) a position embedding $\mathbf{p}_i, i = 1..m$ (of size d) for each token. To remove all syntactic content, the first step is to discard the previously mentioned positional embeddings $\mathbf{p}_i, i = 1..m$. However stopping here would only lead to a bag-of-word model where a given token might be influenced similarly by an adjacent token or one far away. As a consequence, we had to weight the attention score granted to a token depending on its relative distance.

The attention operation can be described as mapping a query (Q) and a set of key-value (K, V) pairs to an output, where the query, keys, values, and output are all vectors (generally packed into matrices) (Vaswani et al., 2017). The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. We thus modify the

Table 6.1: Examples of input sequences given to the neural language models when trained on the different feature spaces.

		Input sequence						
Integral Features		The	sixth	planet	was	ten	times	larger
Syntactic Features	Part-of-Speech	DET	ADJ	NOUN	VERB	NOUN	NOUN	ADJ
	Morphology	Definite=Def|PronType=Art	Degree=Pos	Number=Sing	Ind|Sing|Past|Person=3|Fin	Number=Card	Number=Plur	Degree=Cmp
	Number of Closing Nodes	1	1	2	1	1	2	2
Semantic Features	Content words	–	sixth	planet	–	ten	times	larger

Table 6.2: Models used to probe syntax and semantics in the brain, organized according to the data used for training and the nature of information they encode.

Training data	Model Name	Information			
		Lexical Semantic	Lexical Syntax	Compositional Semantic	Compositional Syntax
Syntactic Features	GloVe$_{Syntax}$	–	✓	–	–
	GPT-2$_{syntax}$	–	✓	–	✓
Semantic Features	GloVe$_{Semantic}$	✓	–	–	–
	GPT-2$_{Semantic}$	✓	–	✓	–
integral Dataset	GloVe	✓	✓	–	–
	GPT-2	✓	✓	✓	✓

classical attention operation:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}((\mathbf{Q}\mathbf{K}^T)/\sqrt{d_k})\mathbf{V}$$

by adding the previously described relative positional embedding $\mathbf{W}$ in the attention mechanisms:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}((\mathbf{Q}\mathbf{K}^T + \mathbf{W})/\sqrt{d_k})\mathbf{V}$$

To build $\mathbf{W}$, we first defined the matrix $\mathbf{D} = (n - 1 + j - i)_{i,j=1..m} \in \mathbb{R}^{m \times m}$ (encoding the number of tokens separating two tokens in the input sequence shifted by $n - 1$) for each input sequence $\mathbf{c}_W$, where n is the maximal input size. $\mathbf{D}$ is then embedded using a lookup table that stores an embedding of size (d_{head}) for each possible value of $\mathbf{D}$, giving $\mathbf{U}$ ($\in \mathbb{R}^{m \times m \times d_{head}}$).

Finally, the weights assigned to the value vectors are adjusted using the embedded relative distances between tokens $\mathbf{W}$ ($\in \mathbb{R}^{n_{heads} \times m \times m}$), defined as:

$$W_{i,j,k} = \sum_{d=1}^{d_{head}} K_{i,j,d} U_{j,k,d}$$

By doing so, we were able to weight words interactions depending on their relative distance in the input sequence, while removing all absolute positional information from tokens hidden-states.

6.1.4 . Validation: Decoding latent representations

Once the models were trained, we first assessed whether syntactic embeddings (the embedding vectors derived from the models trained on the syntactic features) encoded syntactic but not semantic features, and conversely, whether semantic embeddings (the embedding vectors derived from the models trained on the semantic features) encoded semantic but not syntactic features.

We designed two decoding tasks: a syntax decoding task in which we tried to predict the triplet (Part-of-speech, morphological information and number of closing nodes) of each word from its embedding vector (355 categories), and a semantic decoding task in which

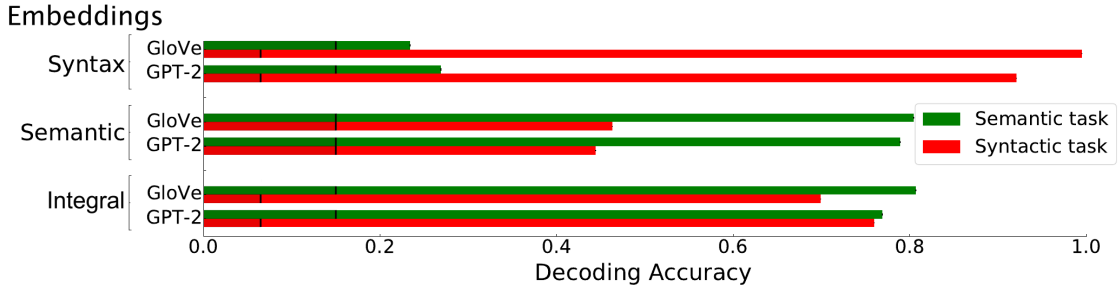


Figure 6.2: **Decoding syntactic and semantic information from words embeddings.** For each dataset and model type (Glove and GPT-2), logistic classifiers were set up to decode either the syntactic or the semantic categories of the words from the text of *The Little Prince*. We report the decoding accuracy of each model on each decoding task. Chance-level was assessed using dummy classifiers and is indicated by black vertical lines.

we tried to predict each word's semantic category (from *Wordnet*[2]) from its embedding vector (837 categories).

We used Logistic Classifiers and the text of *The Little Prince* as train and test data, which was split using a 9-fold cross-validation on runs, training on 8 runs and evaluating on the remaining one for each split. Dummy classifiers were fitted and used as estimations of chance-level for each task and model. All classifiers implementations were taken from Scikit-Learn(Pedregosa et al., 2011).

The decoding performances of the logistic classifiers are displayed in Fig.6.2. The models trained directly on the integral features, that is, the intact texts, have relatively high performance on the two tasks (75% in average for both GloVe and GPT-2). The models trained on the syntactic features performed well on the syntax decoding task (decoding accuracy >95%), but are near chance-level on the semantic decoding task (decoding accuracy around 25% with a chance-level at 16%). Similarly, the models trained on the semantic features display good performance on the semantic decoding task (decoding accuracy greater than 80%), but a relatively poorer decoding accuracy on the syntax decoding task (45%, chance level: 16%). These results validate the experimental manipulation by showing that syntactic embeddings essentially encode syntactic information and semantic embeddings essentially encode semantic information. Still, it must be acknowledged that some overlap remains. This overlap is intrinsic and derived from the very definitions of what syntax and semantics are.

6.2 . Probing Semantic and Syntactic processing in the Human Brain

6.2.1 . Correlations of fMRI data with syntactic and semantic embeddings

Our objective was to evaluate how well the embeddings computed from GloVe and GPT-2 on the syntactic and semantic features fit the fMRI signal in various parts of the brain. For each model/feature type combination, we computed the increase in R score when the resulting embeddings were appended to a baseline model that comprised low-level variables (acoustic energy (RMS), word-rate and log lexical frequency). This was

[2] https://wordnet.princeton.edu/

done separately in each voxel. The resulting maps are displayed on Fig.6.3A.

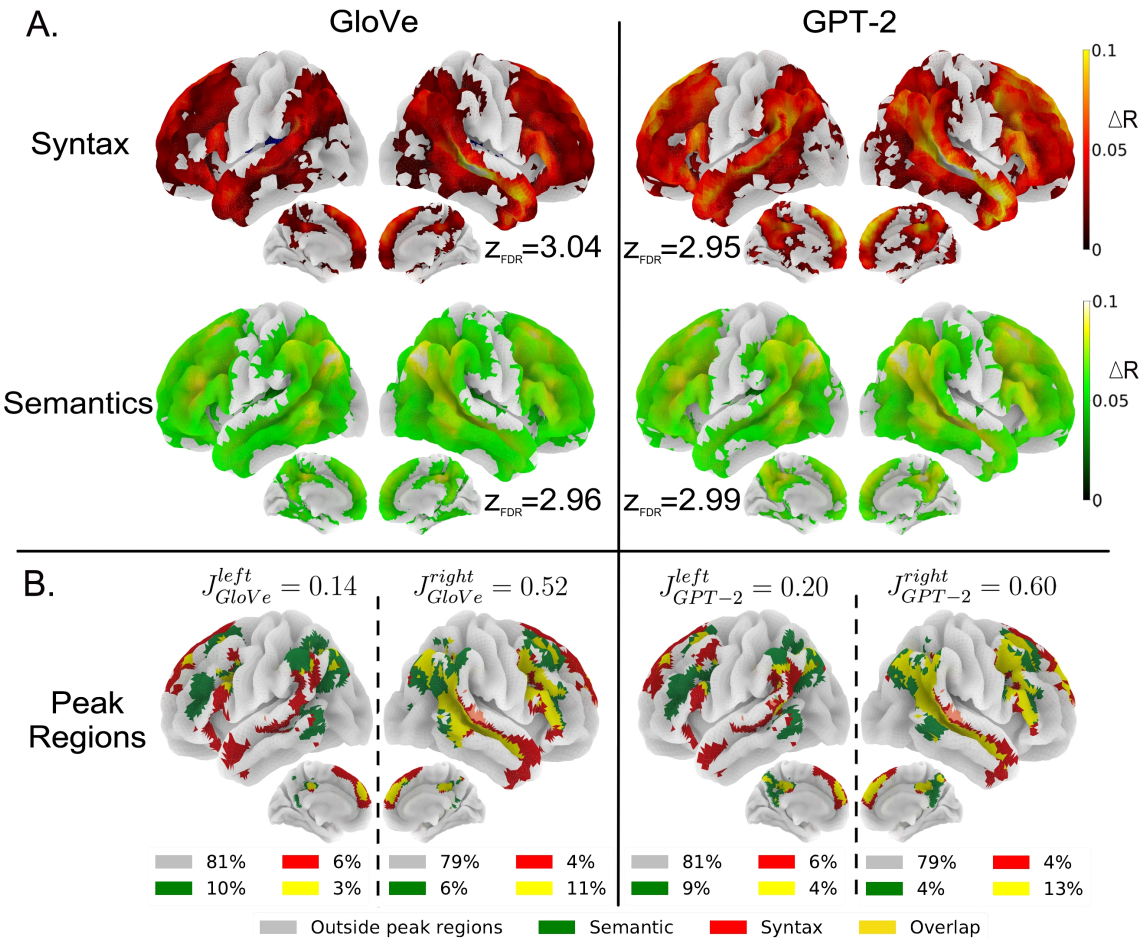


Figure 6.3: **Comparison of the ability of GloVe and GPT-2 to fit brain data when trained on either the semantic or the syntactic features. A)** Increase in R scores relative to the baseline model for GloVe (a non contextual model) and GPT-2 (a contextual model), trained either on the Syntactic features or on the Semantic features (voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with a FDR approach $p < 0.005$; for each figure z_{FDR} indicates the significance threshold on the Z-scores). **B)** Bilateral spatial organisation of syntax and semantics highest R scores. Voxels whose R score belong in the 10% highest R scores (in green for models trained on the semantic features, and in red for models trained on the syntactic features) are projected onto brain surface maps for GloVe and GPT-2 (overlap in yellow and other voxels in grey). A Jaccard score is computed for each hemisphere. It quantifies the ratio between the size of the intersection and the size of the union of semantics and syntax peak regions; the proportion of voxels of each category are displayed for each hemisphere and model.

The maps reveal that the embeddings derived from the semantic or syntactic features through GloVe or GPT-2 significantly explain signal in a set of bilateral brain regions that comprise frontal and temporal regions, as well as the Temporo-parietal junction, the Precuneus and Dorso-Medial Prefrontal Cortex (dmPFC[3]). The classical left-lateralized

[3]Brain regions' abbreviations are listed in Appendix A.1

language network, that includes the Inferior Frontal Gyrus (IFG) and the Superior Temporal Sulcus (STS), is entirely covered. Overall, a vast network of regions is modulated by both semantic and syntactic information.

Nevertheless, detailed inspection of the maps shows different R score distribution profiles (see Appendix A A.4.5). For example, syntactic embeddings yield the highest fits in the Superior Temporal Lobe, extending from the Temporal Pole (TP) to the Temporo-Parietal Junction (TPJ), as well as the Inferior Frontal Gyrus (IFG, BA-44 and 47), the Superior Frontal Gyrus (SFG), the Dorso-Medial Prefrontal Cortex (dmPFC) and the posterior Cingulate cortex (pCC). Semantic embeddings, on the other hand, show peaks in the posterior Middle Temporal Gyrus (pMTG), the Angular Gyrus (AG), the Inferior Frontal Sulcus (IFS), the dmPFC and the Precuneus/pCC.

6.2.2 . Regions best fitted by semantic or syntactic embeddings

As noticed above, despite the fact that the regions fitted by semantic and syntactic embeddings essentially overlap (Fig.6.3A), the areas where each model has the highest R scores differ. To better visualize the maxima from these maps, we selected, for each of them, the 10% of voxels having the highest R scores. Thresholding at the 90-th percentile of the distributions (threshold values displayed in Appendix A 1-Fig.A.5) produces the maps presented in Fig.6.3B.

A first observation is that the number of supra-threshold voxels is quite similar in the left (19%) and right (21%) hemispheres, whether GPT-2 or Glove is considered, showing that during the processing of natural speech, both syntactic and semantic features modulate activations in both hemispheres to a similar extent. The regions involved include, bilaterally, the TP, the STS, the IFG and IFS, the dmPFC, the pMTG, the TPJ, the Precuneus and pCC.

One noticeable difference between the two hemispheres, apparent in Fig.6.3B, concerns the *overlap* between the semantic and syntactic peak regions: it is stronger in the right than in the left hemisphere. To assess this overlap, we computed the Jaccard indices between voxels modulated by syntax and voxels modulated by semantic. The Jaccard index[4] for two sets X and Y is defined in the following manner: $J(X,Y) = |X \cap Y|/|X \cup Y|$. It behaves as a similarity coefficient: when the two sets completely overlap, J=1; when their intersection is nil, J=0. The Jaccard indices were much larger in the right hemisphere ($J^{right}_{GloVe} = 0.52$ and $J^{right}_{GPT-2} = 0.60$) than in the left ($J^{left}_{GloVe} = 0.14$ and $J^{left}_{GPT-2} = 0.20$).

The left hemisphere displayed distinct peak regions for semantics and syntax; syntax involving the STS, the pSTG, the anterior TP, the IFG (BA-44/45/47) and the MFG, while semantics involves the pMTG, AG, the TPJ and the IFS. We only observe overlap in the upper IFG (BA-44), AG and posterior STS. On medial faces, semantics and syntax share peak regions in the Precuneus, the pCC and the dmPFC. In the right hemisphere, syntax and semantics share the STS, pMTG and most frontal regions, with only syntax-specific peak regions in the TP and SFG and semantics-specific peak regions in the TPJ.

Overall, this shows that the neural correlates of syntactic and semantic features appear more separable in the left than in the right hemisphere .

6.2.3 . Gradient of sensitivity to syntax or semantics

[4]Computed using scikit-learn *jaccard_score* function from the *metrics* module.

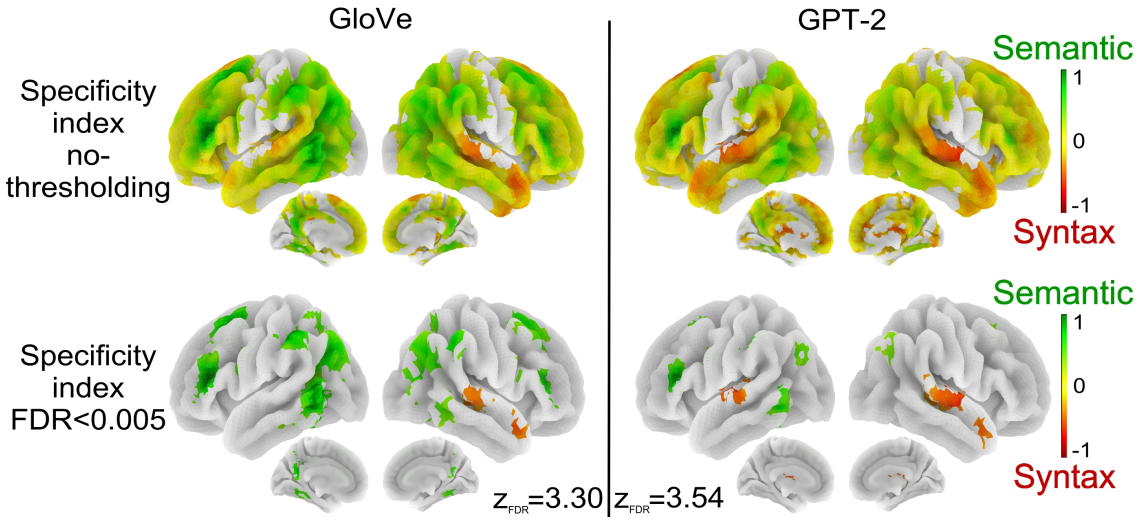


Figure 6.4: **Voxels' sensitivity to syntactic and semantic embeddings.** Voxels' specificity indexes are projected onto brain surface maps reflecting how much semantic information helps to better fit the time-courses of a voxel compared to syntactic information; the greener the more the voxel is categorized as a semantic voxel, the redder the more the voxel is categorized as a syntactic voxel. Yellow regions are brain areas where semantic and syntactic information lead to similar R score increases. The top row displays specificity indexes in voxels where there was a significant effect for semantic or syntactic embeddings in Fig.6.3A. The bottom row is the voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with $FDR < 0.005$ (for each figure z_{FDR} indicates the significance threshold on the Z-scores).

The analyses presented above revealed a large distributed network of brain regions sensitive to both syntax and semantics but with varying local sensitivity to both conditions.

We further investigated these differences by defining a *specificity index* that reflects, for each voxel, the logarithm of the ratio between the R scores derived from the semantic and the syntactic embeddings:

$$x_{specificity}(v) = log_{10}\left(\frac{r_{Semantic}(v)}{r_{Syntax}(v)}\right)$$

r_{Syntax} is the R score increase relative to the baseline model for the syntactic embeddings. $r_{Semantic}$ is the R score increase relative to the baseline model for the semantic embeddings.

Specificity indexes range from -1 to 1. A score of x indicates that the voxel is 10^x-times more sensitive to semantics compare to syntax if $x > 0$ (green), and conversely, the voxel is 10^{-x}-times more sensitive to syntax compare to semantics if $x < 0$ (red). Voxels with specificity indexes close to 0, are colored in yellow and show equal sensitivity to both conditions. Specificity indexes are plotted on surface maps in Fig.6.4. The top row shows the specificity index of voxels where there was a significant effect for syntactic or for semantic embeddings in Fig.6.3A, while the bottom row shows group specificity indexes corrected for multiple comparison using an FDR-correction of 0.005 (N=51).

We learnt from the not-thresholded Fig.6.4 top row, that voxels that are more sensitive to Syntax include, bilaterally, the anterior Temporal Lobes (aTL), the STG, the Supplementary Motor Area (SMA), the MFG and sub-parts of the IFG. Voxels more sensitive to Semantics are located in the pMTG, the TPJ/AG, the IFS, SFS and the Precuneus. Voxels sensitive to both types of features are located in the posterior STG, the STS, the dmPFC, the CC, the MFG and in the IFG.

More specifically, in Fig.6.4 bottom, one can observe significantly low ratios (in favor of the syntactic embeddings) in the STG, aTL and pre-SMA, and significantly large ratios (in favor of the semantic embeddings) in the pMTG, the AG and the IFS. Specificity index maps are consistent with the maps of R score differences between semantic and syntactic embeddings for Glove and GPT-2 (see Appendix A Fig.A.6), but provide more insights into the relative sensitivity to syntax and semantics. These maps highlight that some brain regions show stronger responses to the semantic or to the syntactic condition even when they show sensitivity to both.

Overall, the *specificity index* allows to quantify the gradient of sensitivity to syntax or to semantics, identifying brain regions more sensitive to one of the conditions.

6.2.4 . Unique contributions of syntax and semantics

The previous analyses allowed us to quantify the amounts of brain signal explained by the information encoded in various embeddings. Yet, when two embeddings explain the same amount of signal, that is, have similar R score, it remains to be clarified whether they hinge on information represented redundantly in the embeddings or information specific to each embedding. To address this issue, we analyzed the additional information brought by each embedding on top of the other one. To this end, we evaluated correlations that are uniquely explained by the semantic embeddings compared to the syntactic embeddings, and conversely.

To quantify the unique contribution of each feature space to the prediction of the fMRI signal, we first estimated the Pearson correlation explained by the embeddings learned from the individual feature space - e.g., using only syntactic embeddings or semantic embeddings.

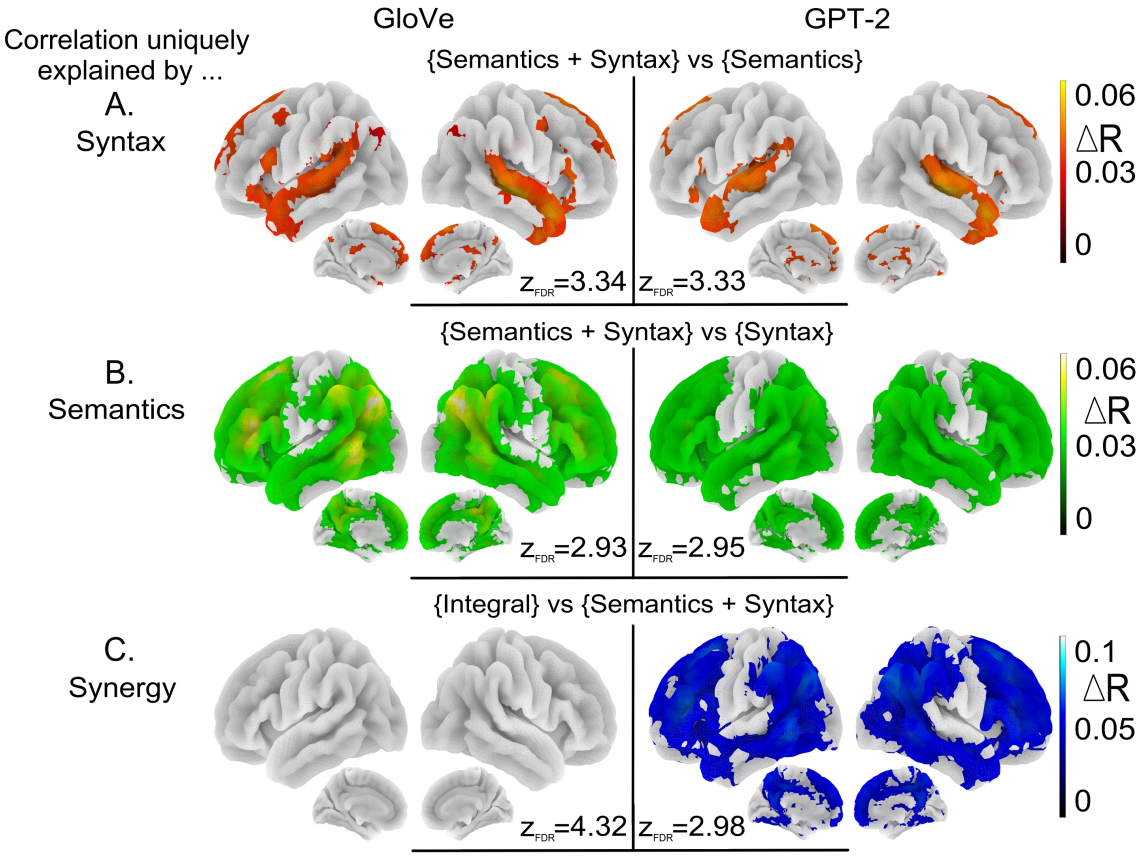


Figure 6.5: **Correlation uniquely explained by each embeddings.** A) Increase in R scores relative to the semantic embeddings when concatenating semantic and syntactic embeddings in the encoding model. B) Increase in R scores relative to the syntactic embeddings when concatenating semantic and syntactic embeddings in the encoding model. C) Increase in R scores relative to the concatenated semantic and syntactic embeddings for the integral embeddings. These maps are voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with a FDR approach $p < 0.005$; for each figure z_{FDR} indicates the significance threshold on the Z-scores.

We then assessed the correlation explained by the concatenation of embeddings derived from different feature spaces - e.g., concatenating syntactic and semantic embedding vectors (de Heer et al., 2017).

Because it can identify single voxels whose responses can be partly explained by different feature spaces, this approach provides more information than simple subtractive analyses that estimate the R score difference per voxel (see Appendix A Fig.A.6).

Syntactic embeddings (Fig.6.5A) uniquely explained brain data in localized brain regions: the STG, the TP, the pre-SMA and in the IFG, with R scores increases of about 5%.

Semantic embeddings (Fig.6.5B) uniquely explained signal bilaterally in the same wide network of brain regions as the one highlighted in Fig.6.3A, including frontal and temporo-parietal regions bilaterally as well as the Precuneus and pCC medially, with similar R scores increases around 5%.

This suggests that even if most of the brain is sensitive to both syntactic and semantic

conditions, syntax is preferentially processed in more localized regions than semantics which is widely distributed.

6.2.5 . Synergy of syntax and semantics at the compositional level

To probe regions where the joint effect of syntax and semantics is greater than the sum of the contributions of these features, we compared the R scores of the embeddings derived from the integral features with the R scores of the encoding models concatenating the semantic and syntactic embeddings (see Fig.6.5C).

For the embeddings obtained with GloVe, this analysis did not reveal any significant effect. For the embeddings obtained with GPT-2, significant effects were observed in most of the brain, but with higher effects in the semantic peak regions: pMTG, TPJ, AG and in frontal regions.

6.3 . Discussion

Language comprehension in humans is a complex process, which involves several interacting sub-components (word recognition, processing of syntactic and semantic information to construct sentence meaning, pragmatic and discourse inference, ...) (Jackendoff, 2002, e.g.). Discovering how the brain implements these processes is one of the major goals of neurolinguistics. A lot of attention has been devoted, in particular, to the syntactic and semantic components (Binder and Desai, 2011b; Friederici, 2017, for reviews) and the extent to which they are implemented in (practically) distinct or identical regions is still debated (e.g. Fedorenko et al., 2020). In Fig.6.6, we present the outcome of a meta analysis of the literature based on the search for the keywords 'syntactic' and 'semantic' in the Neurosynth database (see A.4.7). This analysis, albeit somewhat simplistic, reveals the brain regions most often associated with syntax and semantics.

It must be noted that a fair proportion of the studies included in the meta analysis relied on controlled experimental paradigms with single words or sentences, based on the manipulation of complexity or violations of expectations. To study language processing in a more natural way, several recent studies have presented naturalistic texts to partici-

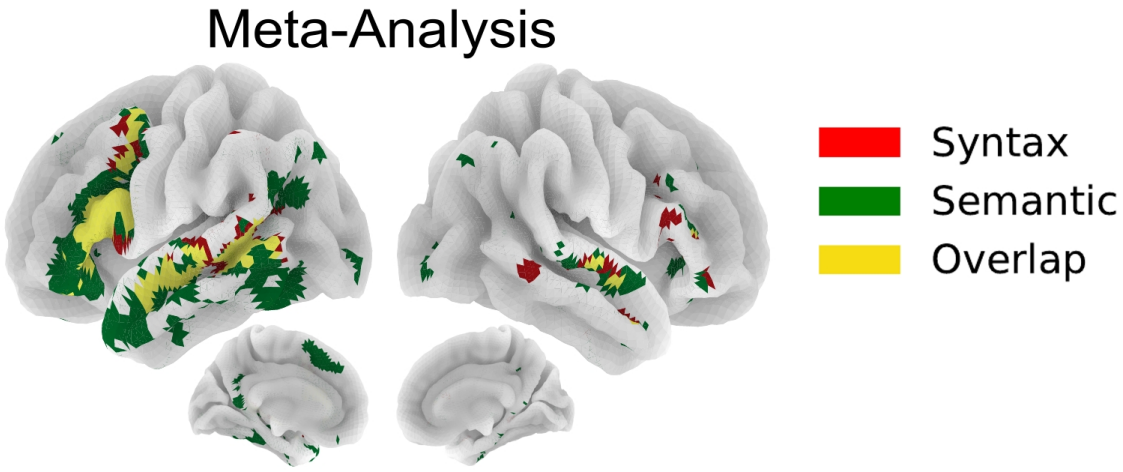


Figure 6.6: **Association maps for the terms "semantic" and "syntactic" in a meta-analysis using Neurosynth.** (`http://neurosynth.org`) The association test map for syntactic (resp. semantic) displays voxels that are reported more often in articles that include the term syntactic (resp. semantic) in their abstracts than articles that do not (FDR correction of 0.01).

pants, and have analyzed their brain activations using Artificial Neural Language Models (e.g. Huth et al., 2016; Pasquiou et al., 2022; Pereira et al., 2018; Schrimpf et al., 2020). These models are known to encode some aspects of semantics and syntax (e.g. Hewitt and Manning; Lakretz et al., 2019; Pennington et al., 2014). In the current work, to further dissect brain activations into separate linguistic processes, we trained NLP models on a corpus from which we selectively removed syntactic or semantic information and examined how well these information-restricted models could explain fMRI signal recorded from participants who had listened to an audiobook. The rationale was to highlight brain regions representing syntactic and semantic information, at the lexical and supralexical levels (comparing a lexical model GloVe, and a contextual one, GPT-2).

Whether models were trained on syntactic features or on semantic features, they fit fMRI activations in a wide bilateral network which goes beyond the classic language network comprising the IFG and temporal regions: it also includes most of the dorso lateral and medial prefrontal cortex, the inferior parietal cortex, and on the internal face, the Precuneus and posterior Cingulate cortex (see Fig.6.3). Nevertheless, the regions *best* predicted by syntactic features on the one hand, and semantic features on the other hand, are not exactly the same. While they overlap quite a lot in the right hemisphere, they are more dissociated in the left hemisphere Fig.6.3, panel B). In addition, the relative sensitivity to syntax and semantics varies from region to region, with syntax predominating in the temporal lobe (see Fig.6.4). Elimination of shared variance between syntactic and semantic features confirmed that pure syntactic effects are restricted to STG/STS, bilaterally, IFG, and pre-SMA, while pure semantic effects occur throughout the network (Fig.6.5 A-B).

Finally, the comparison between the supralexical model (GPT-2) and the lexical one (GloVe), revealed a synergy between syntax and semantics that arises only at the supralexical level (Fig.6.5C).

Models trained on semantic and syntactic features fit brain activity in a widely distributed network, but with varying relative degrees.

When trained on the integral corpus, that is on the integral features, both the lexical (GloVe) and contextual (GPT-2) models captured brain activity in a large *extended language network* (Appendix A Fig.A.4). This large extended language network goes beyond the *core* language network, that is, the left IFG and temporal regions, encompassing homologous areas in the right hemisphere, the dorsal prefrontal regions, both on the lateral and medial surfaces, as well as in the inferior parietal, Precuneus and posterior Cingulate. The result is consistent with the ones from previous studies that have looked at brain responses to naturalistic text, whether analysed with NLP models (e.g. Caucheteux et al., 2021; Huth et al., 2016; Jain and Huth, 2018; Pereira et al., 2018) or not (Chang et al., 2022; Lerner et al., 2011).

Models trained on the information-restricted semantic and syntactic features fit signal in this widely distributed network (Fig.6.3A). This is in agreement with Caucheteux et al. (2021) and Fedorenko et al. (2020) who, using very different approaches, found that syntactic predictors modulated activity throughout the language network. Caucheteux et al. (2021) first constructed new texts that matched, as well as possible, the text presented to participants in terms of their syntactic properties. The lexical items being different, the semantics of the new texts bear little relation with the original text. Then, using a pre-trained version of GPT-2, the authors obtained embeddings from these new texts and averaged them to create syntactic predictors. They found that these syntactic embeddings

fitted a network of regions (ibid. Fig5D) similar to the one we observed (Fig.6.3A). Further, defining the effect of semantics as the difference between the scores obtained from the embeddings from the original text, and the scores from the syntactic embeddings, Caucheteux et al. (2021) observed that semantics had a significant effect throughout the same network (ibid. Fig5G).

Should one conclude that syntax and semantics equally modulate the entire language network? Our results reveal a more complex picture. Figure 6.4 presents a semantics vs syntax specificity index map, showing higher sensitivity to syntax in the STG and anterior temporal lobe, whereas the parietal regions are more sensitive to semantics, consistent with Binder et al. (2009). Another point to take in consideration is that syntactic and semantic features are not perfectly orthogonal. Indeed, the logistic decoder trained on the embeddings from the semantic dataset was better than chance at recovering syntactic features (Fig.6.2), and vice versa. This might be due, for example, to the fact that some features like gender or number are present in both datasets, explicitly in the syntactic dataset and implicitly in the semantic dataset. To focus on the unique contributions of syntax and semantic, we remove the shared variance from the syntactic and semantic models using model comparisons (Fig.6.5).

"Pure" semantic but not "pure" syntactic features modulate activity in a wide set of brain regions.

The unique effect of semantics, when its shared component with syntax was removed, remains widespread (Fig.6.5B). This is consistent with the notion that semantic information is widely distributed over the cortex, an idea popularized by embodiment theories (Hauk et al., 2004; Pulvermüller, 2013), but which was already supported by the neuropsychological observations revealing domain-specific semantic deficits in patients (Damasio et al., 2004).

On the other hand the "pure" effect of syntax "shrinked" to the STG and aTL (bilaterally), the IFG (on the left) and the pre-SMA (Fig.6.5A). The left IFG and STG/STS have previously been implicated in syntactic processing (Friederici, 2011, 2017, e.g.), and this is confirmed by the new approach employed here. Note that we are not claiming that these regions are specialized for syntactic processing only. Indeed they also appear to be sensitive to the "pure" semantic component (Fig.6.5B).

The contributions of the right hemisphere.

A striking feature of our results is the strong involvement of the right hemisphere. The notion that the right hemisphere has some linguistic abilities is supported by the studies on split-brains (Sperry, 1961) and by the patterns of recovery of aphasic patients after lesions in the left hemisphere (Dronkers et al., 2017). Moreover, a number of brain imaging studies have confirmed the right hemisphere involvement in higher-level language tasks, such as comprehending metaphors or jokes, generating the best endings to sentences, mentally repairing grammatical errors, detecting story inconsistencies (see Beeman and Chiarello (2013); Jung-Beeman (2005)). All in all, this suggests that the right hemisphere is apt at recognizing distant relations between words.

The effects we observed in the right hemisphere are not simply the mirror image of the left hemisphere. Spatially, syntax and semantics dissociate more in the left than the right. (see Fig.6.3, Panel B).

Syntax drives the integration of contextual information.

The comparison between the predictions of the integral model trained on the intact texts, and the predictions of the combined syntactic and semantic embeddings from the information-restricted models (Fig.6.5C), highlights a striking contrast between GloVe and GPT2. While the former, a purely lexical model, does not benefit from being trained on the integral text, GPT-2 shows clear synergetic effects of syntactic and semantic information. GPT-2's embeddings fit brain activation better when syntactic and semantic information can contribute together. The fact that the regions that benefit most from this synergetic effect are high-level integrative regions, at the end of the temporal processing hierarchy described by Chang et al. (2022), suggests that the availability of syntactic information drives the semantic interpretation at the sentence level.

Limitations of the study

Two limitations of our study must be acknowledged.

The dissociation between syntax and semantics is not perfect. The way we created the semantic dataset by removing function words clearly impacts supra-lexical semantics. For example, removing instances of *and* and *or* prevents the NLP model from distinguishing between the meaning of "A or B" and "A and B". In other words, the logical form of sentences can be perturbed. This may partly explain the synergetic effect of syntax and semantics described above. Removing pronouns is also problematic as this removed the arguments of some verbs. Ideally, one would like to find transformations of the sentences that keep the semantic information associated to the function words like conjunctions or pronouns, but it is not clear how to do that.

A second limitation concerns potential confounding effects of prosody. One cannot exclude that the embeddings of the models captured some prosodic variables correlated with syntax (Bennett and Elfner, 2019). For example, certain categories of words (e.g. determiners or pronouns) are shorter and less accented than others. Also, although the models are purely trained on written text, they acquire the capacity to predict the end of sentences, which are more likely to be followed by pauses in the acoustic signal. We included acoustic energy and the words' offsets in the baseline models to try and diminish the impact of such factors, but such controls cannot be perfect. One way to address this issue would be to have participants *read* the text, presented at a fixed presentation rate. This would effectively remove all low-level effects of prosody.

Conclusion

State-of-the-art Natural Language Processing models, like transformers, trained with large enough corpora, can generate essentially flawless grammatical text, showing that they can acquire the grammar of the language. Using them to fit brain data has become a common endeavour, even if their architecture rules them out of plausible models of the brain. Yet, despite their low biological plausibility, their ability to build rich distributed representations can be exploited to study language processing in the brain. In this paper, we have demonstrated that restricting information provided to the model during training can be used to show which brain areas encode this information. Information-restricted models are powerful and flexible tools to probe the brain as they can be used to investigate whatever representational space chosen, such as semantics or syntax. Moreover, once they are trained, these models can be used directly on any dataset in order to generate information-restricted features for model-brain alignment. This approach is highly

beneficial, both in term of richness of the features, and scalability, compared to classical approaches that use manually crafted features or focus on specific contrasts. In future experiments, more fine grained control of both the information given to the models as well as model's representations will permit more precise characterisation of the role of the various regions involved in language comprehension.

7 - Investigating context-sensitive brain regions with NLMs

This thesis chapter originally appeared in the literature as Information-Restricted Neural Language Models Reveal Different Brain Regions' Sensitivity to Semantics, Syntax and Context
Alexandre Pasquiou, Yair Lakretz, Bertrand Thirion, Christophe Pallier, *under review*

This Chapter addresses another hot topic in Neuroscience: compositionality or the brain regions that integrate information beyond the lexical level. Previous studies, using ecological paradigms, have identified a hierarchy of brain regions that are sensitive to different types of contextual information and different temporal receptive fields (e.g., Jain and Huth, 2018; Toneva et al., 2022; Wehbe et al., 2014b). That is, brain regions that integrate information over increasing sizes: at the level of words, sentences, paragraphs, ... Using controlled GloVe- and GPT-2-derived latent representations, this Chapter addresses the following question: Which brain regions integrate information beyond the lexical level, and what is the size of the window of integration of such regions? This Chapter presents two approaches to study context sensitivity in the brain: *information-restricted NLMs* and *masked-attention generation* of latent representations. The first section uses information-restricted NLMs to identify the supra-lexical processing systems and the loci that integrate information over different context sizes. The second section uses masked-attention generation to probe the context size that modulate brain activity in each context-sensitive brain region. Overall, these two different approaches coherently identify context-sensitive brain regions as well as the context sizes they integrate.

7.1 . Probing context-sensitive brain regions with information-restricted NLMs

7.1.1 . The supra-lexical processing systems

We first dissociated the brain regions involved in lexical processing from the ones involved in supra-lexical processing by comparing GPT-2, the supra-lexical model which takes context into account, to GloVe, a purely lexical model. Using the versions of GloVe and GPT-2 trained in the last chapter on the syntactic, semantic and integral features, we derived latent representations from each model and used them to fit brain data (all details were presented in Chapter 6).

The differences in R scores between the two models, trained on each of the three datasets are presented in Fig.7.1, which displays voxel-wise thresholded group analyses on $N = 51$ subjects, corrected for multiple comparisons with $p < 0.005$ after FDR-corrected. For each panel z_{FDR} indicates the significance threshold on the Z-scores).

GPT-2 embeddings elicit stronger R scores than GloVe. The difference spreads over wider regions when the models were trained on syntax compared to semantics (respectively Fig.7.1 top left and right). The comparison for syntax led to significant differences bilat-

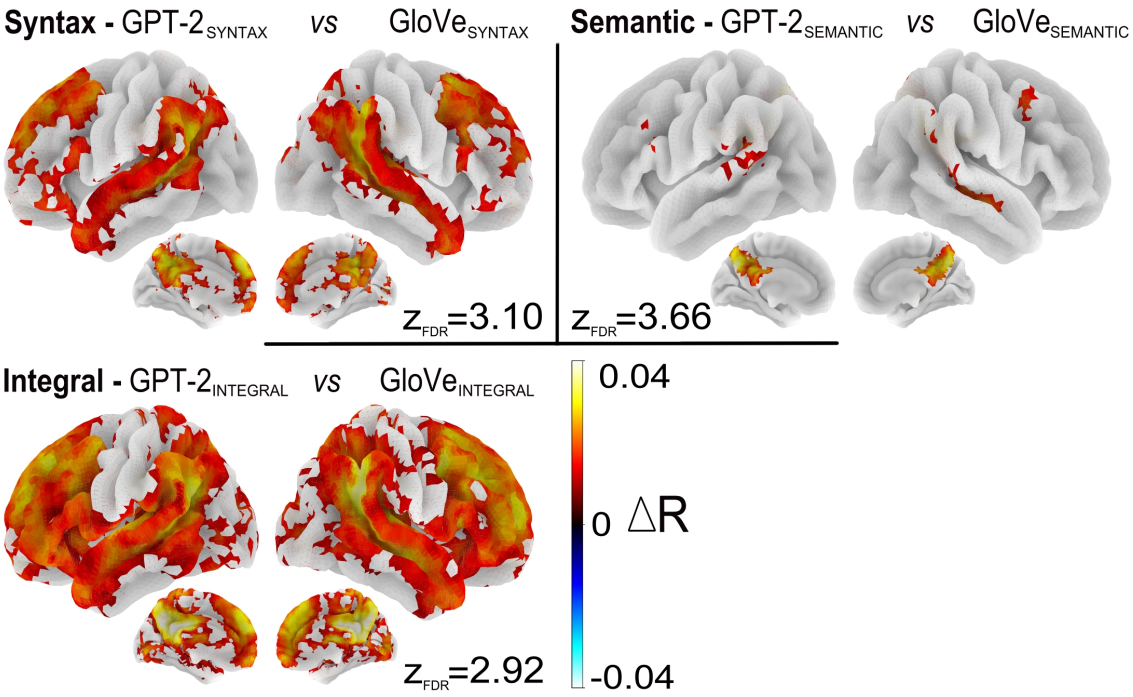


Figure 7.1: **Comparison of lexical and supra-lexical processing levels.** Brain regions that are significantly better predicted by GPT-2 (in red) compared to Glove, when trained on syntactic features (top left), semantic features (top right) and integral features (bottom left). Maps are voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with a FDR approach $p < 0.005$; for each figure z_{FDR} indicates the significance threshold on the Z-scores.

erally in the STS/STG [1], from the Temporal Pole to the TPJ, in superior, middle and inferior frontal regions, and medially in the pCC and dmPFC. For semantics, the comparison only led to significant differences in the Precuneus, the right STS and posterior STG. Fig.7.1 (bottom left) shows the comparison between GPT-2 and GloVe when trained on the Integral features. Given that both semantic and syntactic contextual information were available to GPT-2, these maps reflect the regions that benefit from context during story listening.

[1]Brain regions' abbreviations are listed in Appendix A-A.1

7.1.2 . Modelling Context-limited Features with GPT-2 by restraining information at training and inference

To show that context has an effect is one thing, but different brain regions are likely to have different integration window's size. To address this question, we developed a fixed-context window training protocol to control for the amount of contextual information used by GPT-2 (see Fig.7.2). Using the *Integral Features* defined in the last chapter, we trained three additional GPT-2 models to probe context integration. More precisely, we restricted the size of previous context k (k=5;15;45 tokens) given to the GPT-2 models during training on the *Integral Features*. [2]

When training GPT-2 with a limited amount of contextual information, each input sequence contained $k + 5$ tokens: a special token at the beginning, k context tokens, the current token for which we retrieve the activations in order to fit fMRI brain data, the token that is predicted by the current token and the 2 special tokens at the end (the last special end-of-sentence token is always preceded by a token encoding a blank space, we omit it in the following Fig.7.2).

The training procedure (model size, objective and packages) is similar to the one described in 6.1.2.

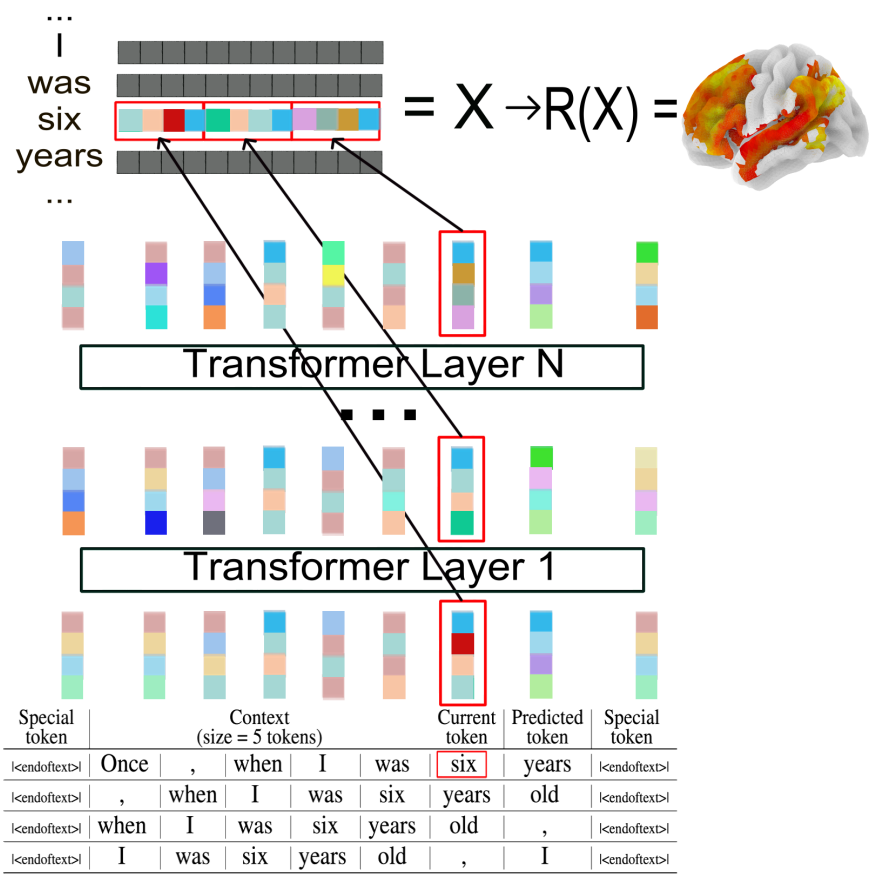
7.1.3 . Context-integration at various scales

By training models with short (5 tokens), medium (15 tokens) and long (45 tokens) range window sizes, we made sure that GPT-2 was not sampling out of the learnt distribution at inference, and not using more context than what was available in the context window. For the 0-context baseline (the non-contextualized model), we used GloVe trained on the integral features.

Comparing GPT-2 with 5 tokens to GloVe (0-size context) highlighted a large network of frontal and temporo-parietal regions. Medially, it included the Precuneus, the pCC and the dmPFC (Fig.7.3, short). Short context-sensitivity showed peak effects in the Supramarginal gyri, the pMTG and medially in the Precuneus and pCC. Counting the number of voxels showing significant short-context effects highlighted an asymmetry between the left and right hemisphere with 1.6 times more significant voxels in the left hemisphere compared to the right. Contrasting a GPT-2 model using 15 tokens of context (the average size of a sentence in *The Little Prince*) versus a GPT-2 model using only 5 tokens, yielded localized significant differences in the SFG/SFS, the TP, MFG and STG near Heschl's gyri and medially in the Precuneus and pCC (Fig.7.3, Medium). The biggest medium context effects included the left MFG, the right SFG and dmPFC and bilaterally the Precuneus and pCC. Finally, contrasting models using respectively 45 and 15 tokens of context revealed 2.8 times as many significant differences in the right hemisphere as in the left. Significant effects were the highest bilaterally and medially in the pCC, followed, in the right hemisphere, by the Precuneus, the dmPFC, MFG, SFG, STS and TP (7.3, Long).

Taken together, our results show 1) that syntax dominantly determines the integration of contextual information, 2) that a bilateral network of frontal and temporo-parietal regions is modulated by short context, 3) that short-range context integration is preferentially located in the left hemisphere, 4) that the right hemisphere is involved in the

[2] We make the approximation that the number of tokens is equal to the number of words, as 95% of words are not split into sub-words.



Special token	Context (size = 5 tokens)					Current token	Predicted token	Special token
|<endoftext>|	Once	,	when	I	was	six	years	|<endoftext>|
|<endoftext>|	,	when	I	was	six	years	old	|<endoftext>|
|<endoftext>|	when	I	was	six	years	old	,	|<endoftext>|
|<endoftext>|	I	was	six	years	old	,	I	|<endoftext>|

Figure 7.2: **Controlling for contextual information in model's activations.** To study sensitivity to context, a GPT-2 model was trained and tested on input sequences of bounded context length (5, 15 and 45). The resulting representations were then used to predict fMRI activity.

processing of longer context sizes, and finally 5) that medial regions (Precuneus and pCC) are core regions of context integration, showing context effects at all scales.

7.2 . Probing context-sensitive brain regions with masked-attention generation

In this section, we present a new method to probe the integration of contextual information using the attention mechanism of transformer-based models. We first describe how we can control for the interactions between tokens using an attention mask. Then, we apply this method to assess brain regions' context-sensitivity and the size of the window on which each context-sensitive brain region integrate contextual information.

7.2.1 . Modelling Context-limited Features with GPT-2 using attention masks

Sensitivity to contextual information was tested using the 12-layer GPT-2 SOTA introduced in Chapter 5. Here, contextual information was not controlled by constraining

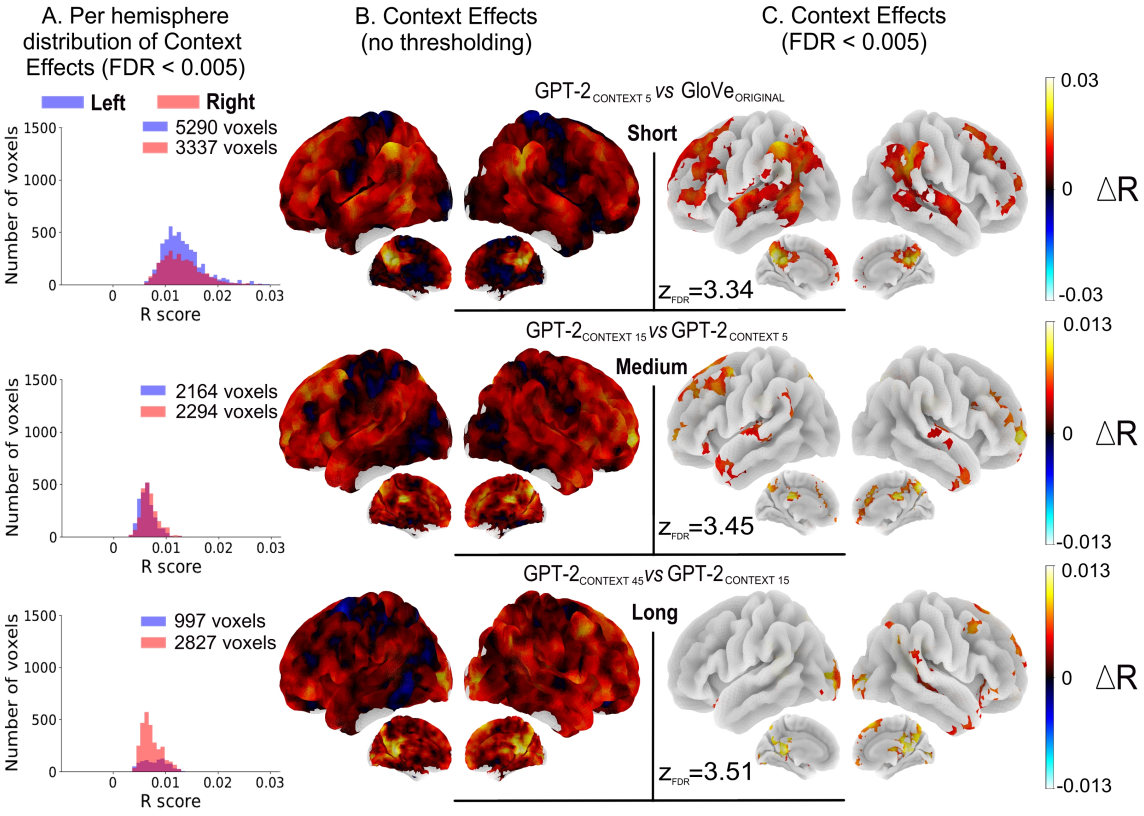


Figure 7.3: **Integration of context at different levels of language processing. A)** Per hemisphere histograms of significant context effects after group analyses (N=51 subjects); thresholded at p<0.005 voxel-wise, corrected for multiple comparisons with the FDR approach. **B)** Uncorrected group averaged surface brain maps representing R scores increases when fitting brain data with models leveraging increasing sizes of contextual information. **C)** Corrected group averaged surface brain maps representing R scores increases when fitting brain data with models leveraging increasing sizes of contextual information; thresholded at p<0.005 voxel-wise, corrected for multiple comparisons with the FDR approach (for each figure z_{FDR} indicates the significance threshold on the Z-scores). **(top row)** Comparison of the model trained with 5 tokens of context (GPT-2$_{Context-5}$) with the non-contextualized GloVe. **(middle row)** Comparison of the models respectively trained with 15 (GPT-2$_{Context-15}$) and 5 (GPT-2$_{Context-5}$) tokens of context. **(bottom row)** Comparison of the models respectively trained with 45 (GPT-2$_{Context-45}$) and 15 (GPT-2$_{Context-15}$) tokens of context.

the input sequence but rather by playing with the internal mechanisms of the GPT-2 transformer, namely: its attention mechanisms.

We constrained the model in a way that each token can only access tokens in a context window. This was done by providing an attention mask in addition to the input sequence (see Fig. 7.4). The attention mask 'truncated' the input by removing interactions with words that were outside the context window (outside the attention mask). Controlling tokens' interactions with the attention mask preserves the positional encoding of the words in the sentence, as well as input sequences that follow the training inputs statistics with complete sentences and the right use of the special tokens. More precisely, given an input sequence containing a target token for which we want to retrieve the latent representation,

and a context window of size n, we gave the GPT-2 model the tokenized sequence converted to ids, as well as a binary vector that defined the tokens that could interact together. This binary vector contains 0 everywhere, except for the target token and the $n-1$ tokens before it, where it equals 1. Thus, only the tokens belonging to the *context window* (where there are 1 in the attention mask) can 'see' each other. Tokens outside the context window have no interaction with any other token. Note that GPT-2 is an incremental model and a given token cannot integrate information over following tokens. Thus, every token in the sentence can only interact with the tokens appearing before them in the attention mask. This operation is performed implicitly during the forward pass. An important point is, that the attention mask is the same for all the tokens in the input sequence, modulo the incrementality. If we were to give a different context-window to each token, we would propagate information outside of the context window because of model's depth. For example, let's consider that we want to retrieve the latent representation of a target token with a context-window size n. If each token can look at the n tokens before itself, then at the first layer, the latent representation of each token includes information on the past n tokens. As a consequence, the target token would see its context-window size double implicitly at each new layer.

To retrieve the latent representation of all the words of an input text, we gave as many $(inputsequence, attentionmask)$ pairs to the GPT-2 model as there are words in the text. We then retrieved the target token's latent representation for each pair. An example is given in Fig. 7.4 for a context-window size of 4.

The motivation behind this approach is the following: if a word needs short-range information to build its latent representation, then the latent representation won't be affected when using a small context-size compared to a long one. On the contrary, if a word needs long-range information to build its latent representation, then the latent representation will be damaged when using a small context-size. The more we increase the context-size, the more we release the constraint on the latent representations. This variable quality of latent representations will be reflected in the fitting performance of the model. Indeed, the brain regions that are better fitted by features aggregated over long context-sizes will benefit more from increasing the context-size compared to regions that focus on lexical processing.

To summarize, we aim at using these context-damaged representations to fit fMRI brain data, and look at the brain regions that benefit from longer context-sizes.

7.2.2 . Quantifying brain regions sensitivity to context

We computed context-limited latent representations for each word in the *The Little Prince* novella, for context-window sizes in [1, 2, 3, 4, 5, 7, 10, 12, 15, 17, 20, 22, 25, 27, 30, 32, 35, 37, 40, 42, 45]. The latent representation from layer 9 of the model (dim=768) was retrieved and fitted to fMRI brain data for each context-window size and subject (N=51). Then, we examined the impact of the context-window size on the models' predictive performance (R_{test} scores).

To remove noise, voxel-level information was aggregated by computing the median R score across voxels in each parcel of the Difumo atlas (Dadi et al., 2020) with 1024 regions of interest (ROIs), and voxels with 3mm edges. The Difumo atlas is a probabilistic atlas, that is, for each parcel, the value at a given voxel indicates how strongly it is related to this parcel. For each parcel of the probabilistic Difumo atlas, we only kept voxels constituting 90% of the non-zero loadings. Then, we resampled all ROI masks to voxels with 4mm

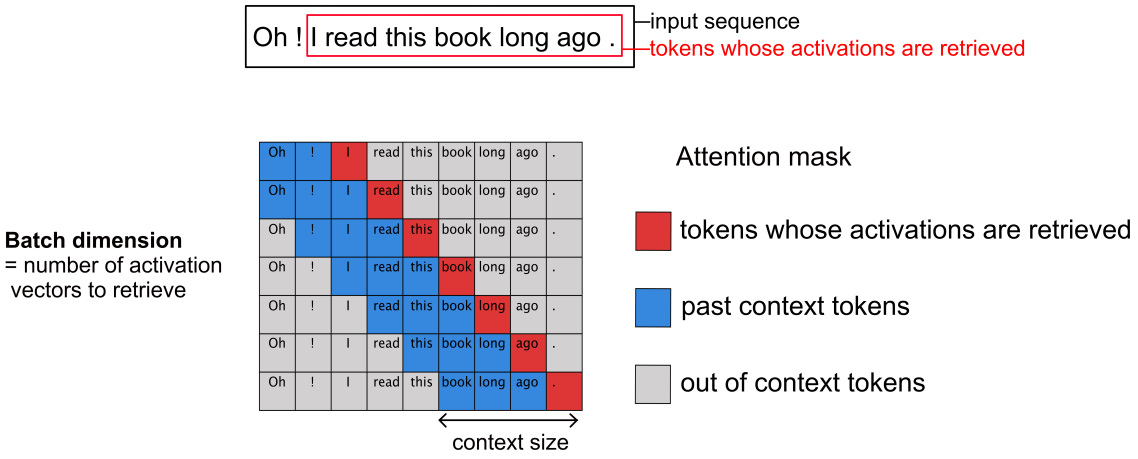


Figure 7.4: **Controlling for tokens' interaction using attention masks.** Examples of (input sequence, attention mask) pairs to retrieve the latent representation of each word of the target sentence (framed in red above). An input sequence is represented by a row, the target token is colored in red, tokens in the attention mask are blue or red (context size = 4), and out-of-context tokens are grey.

edges to match the voxel size of 'The Little Prince' fMRI dataset, giving images of size (37, 46, 38) (Fig. 7.5A).

Let's call ROI-score the median R score across voxels in a given ROI (parcel). For each participant and ROI, we fitted a Linear Regression on the (context_size, ROI-score) points to get the slope of increase of the ROI-score as a function of context-size (Fig. 7.5B). Brain regions' context-sensitivity was estimated with a t-test on the slopes of increase across subjects, with a FDR correction of 0.01 to account for multiple comparisons (Fig. 7.5C).

For each context-sensitive parcel of the atlas, we estimated its *maximal context-size*, i.e. the last context-window size over which the ROI-score is less than one standard deviation away from its maximal value (Fig. 7.6A). We reported the maximal context-sizes found in Fig. 7.6B.

Fig. 7.6 corroborates the results from Fig. 7.1. First, most of the language related brain regions are context-sensitive: the Temporal lobe from the TP to the inferior Parietal Lobule, all frontal regions except for the Frontal Pole, and medially, the dmPFC, Precuneus and posterior Cingulate gyri. This network of context-sensitive brain regions is bilateral and mostly symmetrical. Notes that low-level regions such as the auditory, motor and visual cortices are not context-sensitive.

We observed parcels sensitive to longer context sizes in the right hemisphere compared to the left hemisphere (Fig. 7.7). The brain regions integrating longer-context include the SFG, upper MFG, STS anterior and posterior, AG (+Jensen sulcus), and medially the DPMC and posterior cingulate gyri. While the TP, the anterior and middle MTG as well as the IFG and anterior MFG integrate smaller context-sizes.

Because of the aggregation of voxel R scores at the ROI level in the masked-attention analysis, we observed lower context-sizes in Fig. 7.6 compared to Fig. 7.1.

Taken together, our results show that the entire language network is context-sensitive, with the right hemisphere being involved in the processing of longer context sizes.

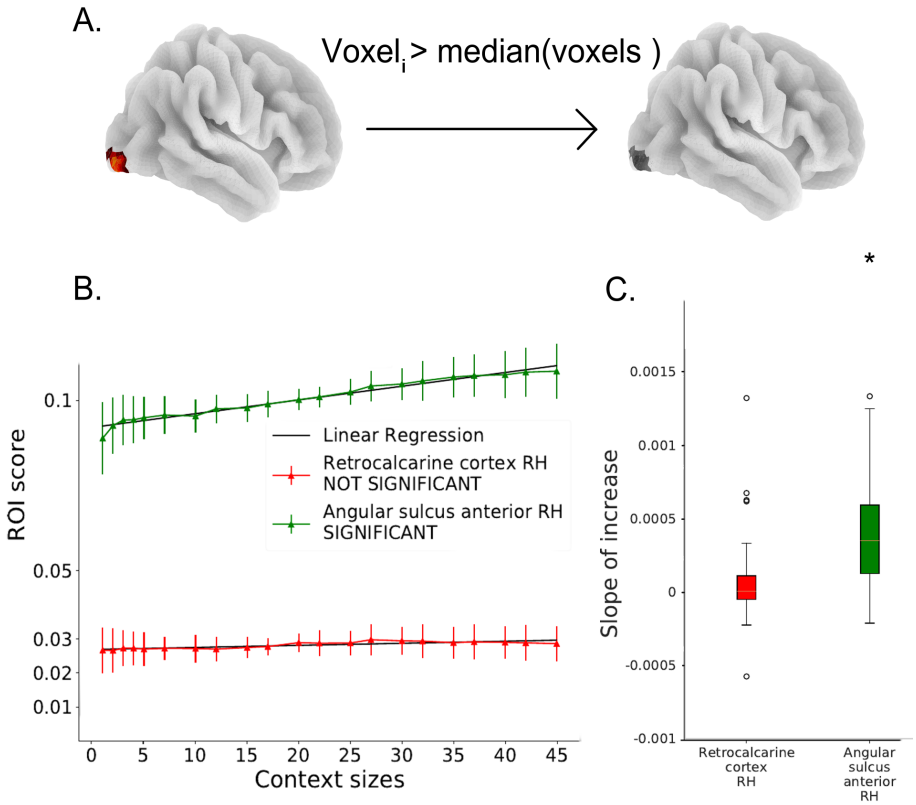


Figure 7.5: **Assessing Brain regions' context-sensitivity.** A) For each parcel of the 1024 parcels Difumo atlas, we selected voxels whose value was above the median of all non-zero values. We obtained a binary mask for each parcel, indicating which voxels belong to it. The 'retrocalcarine cortex RH' parcel is represented here. B) For each subject, we computed the median R-score (ROI-score) inside each parcel as a function of context-size, and displayed the averaged ROI-scores and standard deviations across subjects, as well as a linear regression fitted on the (context_size, ROI-score) points of each parcel. C) Parcels' context-sensitivity was assessed through a t-test on the slope of increase of the ROI-score as a function of context-size across subjects.

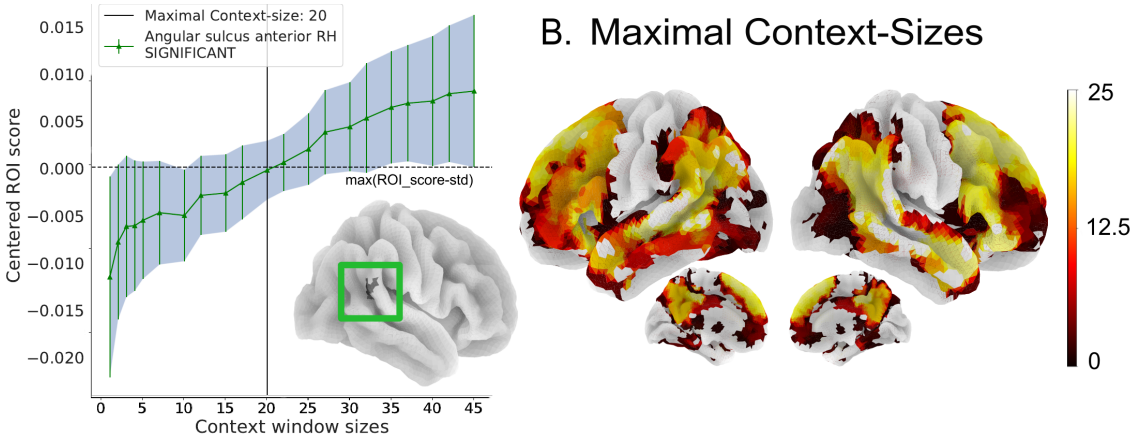


Figure 7.6: **Assessing the maximal context window size over which information is integrated.** A) Determination of the maximal context-size for each parcel of the Difumo atlas. The maximal context-size is defined as the last context-size inferior to the maximal averaged centered ROI-score minus its standard deviation. B) Surface projection of Difumo's parcels maximal context-size in context-sensitive brain regions (obtained with a 12-layer GPT-2).

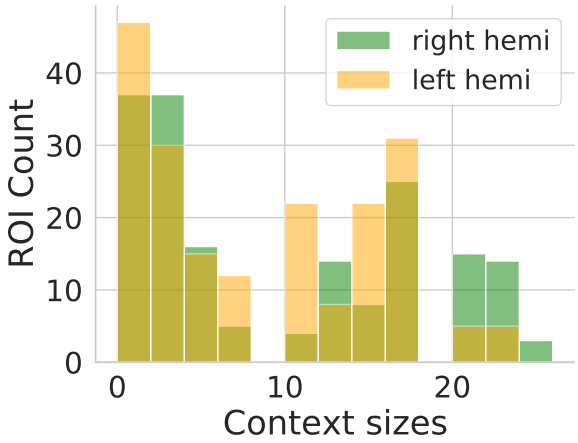


Figure 7.7: **Comparison of the maximal context sizes per ROIs in the left and right hemispheres.** Histograms representing the maximal context sizes distribution across context-sensitive ROIs, in the left hemisphere (orange), and in the right hemisphere (green).

7.3 . Discussion

In this chapter, we investigated the integration of contextual information in the human brain. Using two different approaches, namely *information restricted NLMs* and *masked-attention generation*, we found context-sensitive networks of brain regions that integrate information at different scales.

Most of the cortex is context-sensitive.

The comparison between the supralexical model (GPT-2) and the lexical one (GloVe), as well as the analyses of context-sensitivity using masked-attention generation and Information-restricted NLMs, revealed brain regions involved in compositionality (Fig.7.1, Fig.7.3 and Fig.7.6). Similarly to Jain and Huth (2018) who varied the amount of context fed to LSTM models, from 0 to 19 words, we found a bilateral and mostly symmetrical ensemble of brain regions, involving most of the cortex except for: sensory areas, motor areas, the Frontal pole and the ITG. However, the lack of signal in the ITG prevents us from excluding properly context-sensitivity in the ITG. Like in Jain and Huth (2018), we found short context sensitivity around the middle part of the temporal lobe and the IFG, as well as longer context-sensitivity in the right hemisphere. The shorter context-sizes found in Jain and Huth (2018) are probably due to the limited ability of LSTM models to integrate long range information, compared to transformers.

Our results are also consistent, to some extent, with the ones from Lerner et al. (2011) who probed hierarchical processing in the brain by scrambling the audio stimuli at different timescales corresponding to different linguistic structures (phonemes, words, sentences, paragraphs). They discovered a hierarchy departing from the superior temporal lobe, corresponding to word level information, and going up to the Angular gyri and the superior and middle frontal gyri, corresponding to sentence or paragraph level information. First, we both find medial regions, namely the medial Prefrontal Cortex, the Precuneus and the pCC, to be core components of long context integration. Secondly, we identified a similar widely distributed network of brain regions involved in supra-lexical contextual integration (ibid green and blue regions in their Fig.3). Nonetheless, some differences between our works can be observed. Lerner et al. (2011) identified brain regions that are involved at different levels of information processing: brain regions integrating word-level information, sentence-level information and paragraph-level information. While our work identified brain regions whose activity is modulated by semantically richer representations, whether by controlling the linguistic structures that enrich word representations (the words before, the words in the sentence, or the words in the paragraph) or by directly controlling the number of words of preceding context.

Syntax drives the integration of contextual information.

Comparing the contextual GPT-2 with the non-contextual GloVe, when trained on either the semantic, syntactic or integral features, confirmed that the integration of contextual information benefits more from the presence of syntactic information compared to semantic information (Fig.7.1). Indeed, including syntactic information at the supra-lexical level induced a higher gain compared to semantic information. Among a large set of regions, the STS up to the AG, and medial regions showed the highest increases. This observation is partly supported by Siegelman et al. (2019), who found the left posterior part of the STS going up to the Jensen sulcus to be involved in the processing of sentence-level

syntactic information.

The right hemisphere preferentially integrates long context, while the left hemisphere preferentially integrates short context.

Importantly, a dissociation between the left and right hemispheres arises from our analyses. In Fig.7.3, we observed that short-range context integration is mainly located in the left hemisphere, while the right hemisphere is involved in the processing of longer context sizes. Corroborating results from Chapter 6, these findings are coherent with other brain imaging studies that have supported the role of the right hemisphere in higher-level language tasks (see Beeman and Chiarello (2013); Jung-Beeman (2005)).

Medial regions are core components of the integration of contextual information.

Finally, the Precuneus and the posterior Cingulate gyri (pCC), appear as core regions of context integration, showing the highest context effects at all scales. The Precuneus/pCC, inferior parietal and dorsomedial prefrontal cortex are part of the Default Mode Network (DMN) (Raichle, 2015). The same areas are actually also relevant in language and high-level cognition. For example, early studies examining the role of coherence during text comprehension had pointed out the same regions (Ferstl and von Cramon, 2001; Xu et al., 2005): coherent discourses elicit stronger activations than incoherent ones. Recent work by (Chang et al., 2022) has revealed that the DMN is the last stage in a temporal hierarchy of processing naturalistic text, integrating information on the scale of paragraphs and narrative events, see also (Baldassano et al., 2017; Simony et al., 2016). These regions are not language-specific though, as they have been shown to be activated during various theory of mind tasks, relying on language or not, and have thus also been dubbed the "Mentalizing network" (Baetens et al., 2014; Mar, 2011).

Limitations of the study

A limitation of these analyses must be acknowledged.

While our work highlighted context-sensitive brain regions, the detailed analyses of the window sizes of integration did not show brain regions processing strictly defined linguistic structures, but rather brain regions whose activity was modulated by semantically richer representations.

8 - The limits of NLM-brain comparisons

This thesis chapter originally appeared in the literature as 'Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps'
Alexandre Pasquiou, Yair Lakretz, John Hale, Bertrand Thirion, Christophe Pallier, *ICML 2022*

Capitalizing on the several advances made by NLMs in the last decade, the neuroscience community has used NLMs to study neural activity in the human brain during language processing, shedding light on the hidden mechanisms underlying the processing of semantics or syntax or even the integration of contextual information. These findings have suggested a partial convergence between the representations of NLMs and the ones of the brain. This chapter examines the similarity of brains and artificial neural language models, showing that even if NLMs can be fruitfully used to probe language processing in the brain, there are pitfalls and limits to which attention should be paid. The first section investigates the limited encoding performance of NLMs when fitting fMRI brain data, while the second section highlights the divergences between brains and ANNs.

Overall, we show the limits and pitfalls of the comparisons between brains and ANNs, and support the need to grant specific care to potential confounds, as well as carefully design experiments in which ANN's design and training are controlled.

8.1 . A Limited Encoding Performance

8.1.1 . Removing confounds and variables of non-interest

Variables of non-interest, nuisance factors as well as confounds pose a major problem to the interpretation of the encoding models. As seen in Chapter 5, variables of non-interest already predict a significant amount of BOLD signal. However they are most of the time orthogonal to the dimensions that are being probed in the brain (like in Chapter 6 for example).

Thus, controlling for potential nuisance factors, variables of non-interest and confounds is essential when designing encoding experiments. Nuisance factors can regroup movement artefacts, heart beats or respiration, while confounds can regroup low-level features such as the Basic Features (see Chapter 5); but they are not limited to these model-independent examples. For example, when using latent representations extracted from a given NLMs, it is possible that the very design of the model architecture already bends the feature space, creating some patterns in the latent representations without even training the model. Does model architecture contribute to or hinder the ability of the model to predict brain activity?

We investigate this question by comparing four types of trained and untrained language models (GloVe, LSTM, GPT-2 and BERT) in their ability to fit functional Magnetic Resonance Imaging (fMRI) timecourses of participants listening to *The Little Prince* audiobook. Importantly, we conduct the model comparison while controlling for various aspects of the architecture of the models as well as the type and size of the corpus on which they are trained. More precisely, we used as training data the *Integral dataset* crafted from the Project Gutenberg (see Chapter 5 and 6 for details).

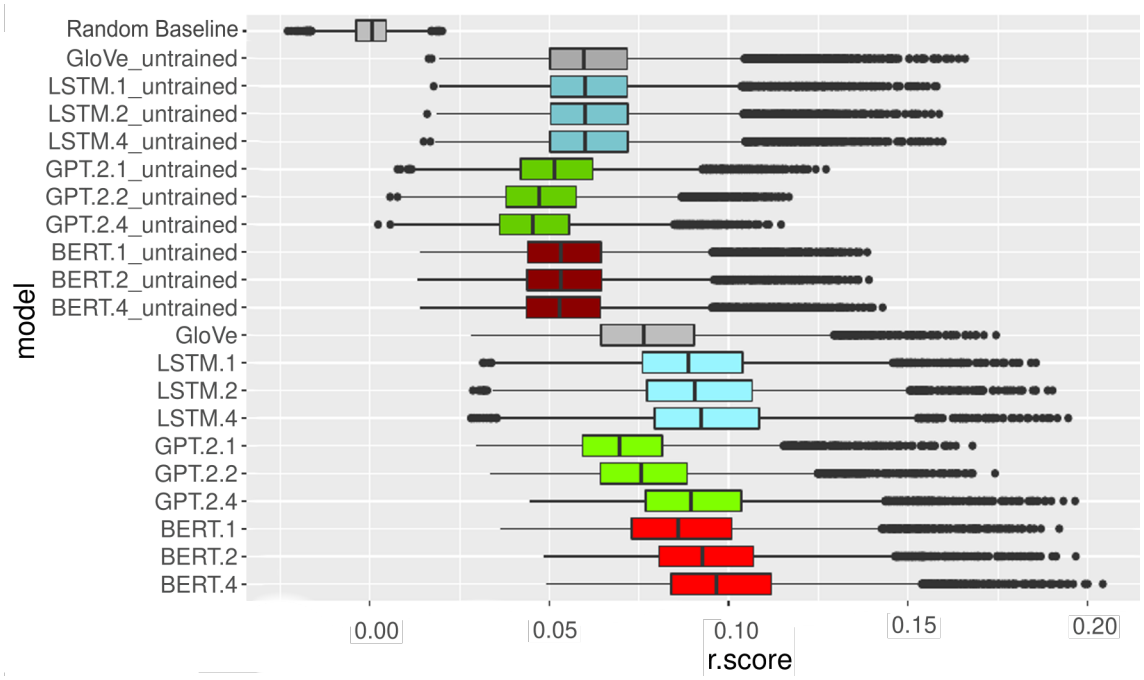


Figure 8.1: **Distributions of R_{test}-values across voxels in the 25% most reliable voxels across subjects (SRM25)** for untrained and trained versions of GloVe (static word embedding), LSTM, GPT-2 and BERT models, having 1, 2 or 4 layers as well as for a random baseline. The random baseline consisted in generating a random vector following a normal distribution with mean 0 and a standard deviation of 1 for each token. For example, two different random vectors were generated for two different occurrences of the same word.

In our first analysis, we assessed whether the model class and number of layers bias its ability to fit the fMRI brain data of the English participants of 'The Little Prince' fMRI corpus. We instantiated several untrained versions of each model class, varying the number of layers, and generated latent representations from these models before fitting them to brain data. For each model, the latent representations were built using all the hidden-states of all layers, including the embedding layer. We also defined a Baseline model whose latent representations are obtained by associating a fixed embedding vector of size 768 (size of each model's layer) to each word of the text. It is equivalent to an untrained GloVe model (and will be referred to as such). For each, we obtained 3D brain maps displaying the average R_{test} values in each voxel. Then, we derived in a model-agnostic manner from a Shared-Response Model (SRM, Chen et al. (2015)) (see 2.6) the most "responsive" voxels. That is, the voxels whose R values were among the 25% highest ones. This set of 6,541 voxels, which we will refer to as "SRM25" is displayed on a brain surface in Fig.8.2A. Finally we displayed, in Fig.8.1, boxplots of the R_{test} values distributions in the SRM-defined voxel selection named *SRM25*.

Remarkably, all untrained models, regardless of their architecture, explain signal better than chance (significantly better than 0). Untrained LSTM and untrained GloVe (that is, Fixed Random Embeddings) perform equally well with an average score around 6.3% (SE=0.02%), and significantly better than Transformers as attested by direct comparisons between untrained 4-layer models: LSTM.4−GPT-2.4 (1.6% SE=0.02%); LSTM.4−BERT.4

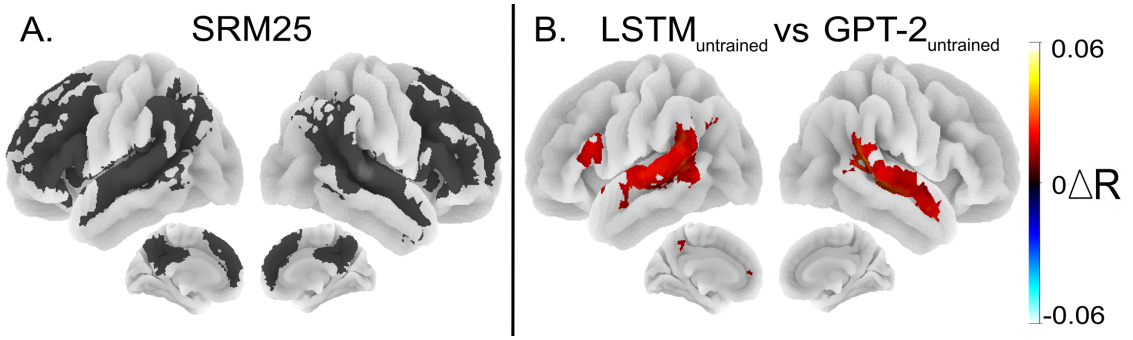


Figure 8.2: **A) SRM25 and B) LSTM.2 vs GPT-2.2 architecture.** A) Voxels whose R values were among the 25% largest ones. B) Brain regions in which an untrained LSTM.2 outperforms an untrained GPT-2.2 model.

(0.7% SE=0.004%). [1] Overall, untrained GPT-2.4 had the worst performance (BERT.4−GPT-2.4 (0.9% SE=0.01%)).

The brain regions where $\text{LSTM}_{\text{untrained}}$ performs significantly better than $\text{GPT-2}_{\text{untrained}}$ are displayed on Fig.8.2B. They are located within the left hemispheric language network and its right counterpart (Superior Temporal Gyrus/Superior Temporal Sulcus and Inferior Frontal Gyrus pars opercularis).

Looking at the effect of the number of layers for LSTM, GPT-2 and BERT models, we observed on Fig. 8.1, a change in performance either flat (for LSTM and BERT) or negative (for GPT-2) for untrained models. Comparing 4-layer models to 1-layer models yields the following: LSTM (-0.02% SE=0.002%); GPT-2 (-0.6% SE=0.004%), BERT (-0.02% SE=0.003%).

For trained models, performance improves with the number of layers. The increase in performance (4-layer model's performance - 1-layer model's performance) is more marked for Transformers — GPT-2 (2% SE=0.006%) and BERT (1% SE=0.006%) — than for LSTM (0.4% SE=0.005%).

Overall, untrained models already explain an important part of the signal explained by trained models. It's worth noting that with untrained GloVe, each word in the corpus is given a fixed random vector. On the other hand, untrained Transformers map each word to a variable vector that depends on the surrounding context. This means that untrained Transformers generate different random embeddings for the same word in different contexts, leading to reduced brain scores compared to an untrained GloVe which is better at predicting how the brain responds to frequently occurring words in both the training and test data (e.g., function words). The validation of this discovery is supported by the fitting performance of a random baseline. Instead of assigning a fixed random embedding per word we assigned a random vector to each token, resulting in two different random vectors being generated for two different occurrences of the same word. This process destroys the consistency of the random vector associated with each word, ultimately leading to a decrease in the fitting performance to 0.

This result highlights the importance of controlling for all confounds even the ones that are model-related as they can lead to misinterpretations.

8.1.2 . The best NLMs only explain up to 60% of SRM's R scores

[1] LSTM.X, GPT-2.X or BERT.X mean an X-layer version of the model.

Trained and untrained NLMs' latent representations significantly predict fMRI brain data. When comparing with the ceiling of explainable signal in Fig.5.5, Fig.5.11 and Fig.5.12 panels E, they explain from 15% to 60% of the signal across voxels.

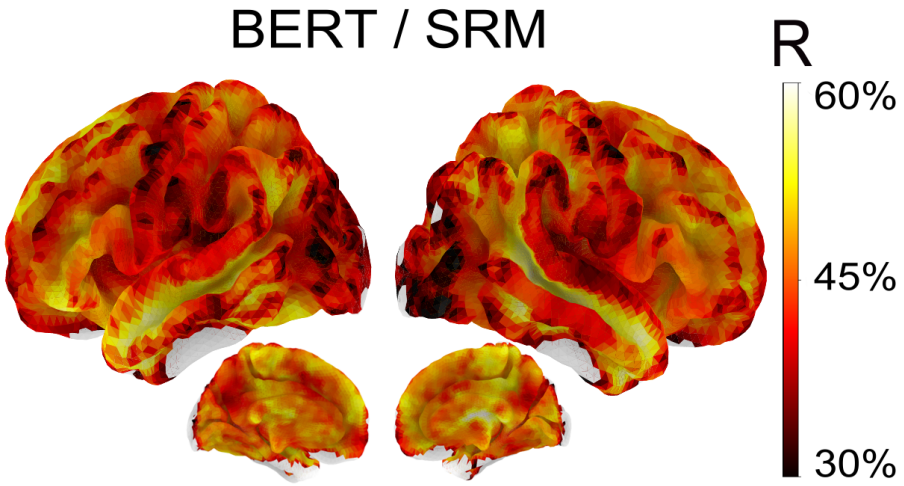


Figure 8.3: **Percentage of explainable fMRI signal fitted by BERT SOTA.** Voxels' R score for the 12-layer BERT SOTA, in percentage of the explainable signal as defined by the SRM model.

In this section we selected the model that best fitted brain data: the 12-layer BERT SOTA, and looked at the percentage of explained signal across voxels (Fig.8.3). Percentages range from 30% in badly fitted brain regions such as the occipital and motor cortex, up to 55-60% in the regions that are best fitted by the model, namely the STS[2] from the TP up to the AG, the IFS, the SFG and the dmPFC, all of them bilaterally. It is reassuring to see that language related brain regions are best fitted by the language model compared to other brain areas. This suggests a greater similarity between the activation patterns of these brain regions and the model's latent representations. This greater similarity is consistent across subjects, supporting a relative convergence between model's representations and brain representations.

8.2 . Divergence between brains and models

8.2.1 . A limited convergence between brains and language models

A point that stands out clearly among previous results, is that trained models better fit brain data than models initialized with random weights. To quantify this improvement for each model type, we computed, in each voxel, the difference in R_{test} between the trained model and the untrained model. Fig.8.4A shows the distributions of these training effects, while Fig.8.4B shows the locations of voxels where the R_{test} increases are significant.

First, all differences were statistically significant: GloVe (1.5% SE=0.02%); LSTM (3.1% SE=0.02%) ; GPT-2 (4.5% SE=0.02%); BERT (4.4% SE=0.02%); in Student T-tests, all $p < 10^{-16}$). Additionally, the effect of training is spatially consistent across models, that is, displays similar topographies across models; and the R-score improvements are comparable in high-order language networks across models.

[2]Brain regions' abbreviations are listed in Appendix A.1

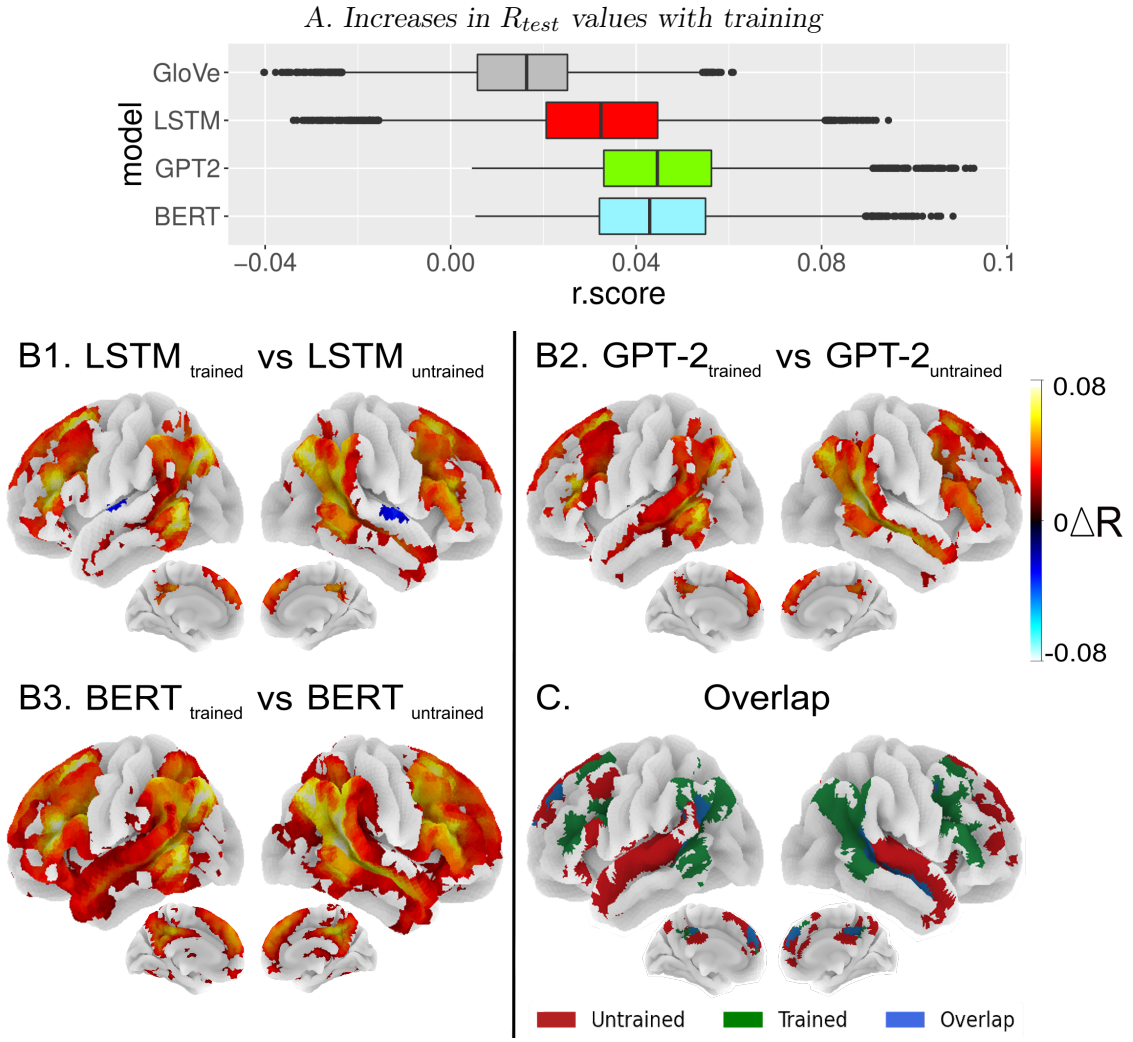


Figure 8.4: **Effect of model training.** A) Distributions of R_{test} increases for the 2-layer versions of LSTM, GPT-2 and BERT. B) Brain areas showing significant increases: LSTM (B1), GPT-2 (B2) and BERT (B3). C) Regions showing the strongest gains in R scores with training across the three models (intersection of the three previous maps thresholded at the 10% upper percentile), in green. Regions showing the strongest R scores across the three 2-layer untrained models: LSTM, GPT-2, BERT (intersection of the three maps thresholded at the 10% upper percentile), in red. There is a 18% overlap between these two highlighted networks (in blue).

To assess the similarity between the hotspots on these maps, we thresholded them, keeping the 10% of voxels (2,617 voxels) showing the highest gains with training. We then computed the percentage of overlap across the resulting binarized maps. Results are presented in Table 8.1. There is a 75% overlap between the maps of all 3 models. For simplicity, this overlap is called *Training Gain Overlap* in the following.

We then thresholded the untrained models brain maps, keeping again the 10% (2,617 voxels) of voxels showing the highest R score, and computed the percentage of overlap across the resulting binarized maps. Results are presented in Table 8.2. There is a 79% overlap between the maps of all 3 models. For simplicity, this overlap is called *Untrained*

Overlap in the following.

Finally, we studied the intersection between the *Training Gain Overlap* and the *Untrained Overlap* of the three models, and synthesized differences and similarities in Fig. 8.4C. The *Training Gain Overlap* and the *Untrained Overlap* are presented in Fig. 8.5.

The similar topographies across models confirm previous results showing a convergence of the representations learnt by the NLMs during training with brains' representations. In addition, this convergence is more localized in the brain regions processing language and is similar across NLMs' architectures.

However, as previously stated, this convergence is partial. Fig. 8.6 (panels C-D-E) displayed the relation between model's perplexity and brain score[3] for several training checkpoints of GPT-2 and LSTM, starting from the end of the first epoch up to epoch 5 for GPT-2, and up to epoch 14 for LSTM. For panel C and E, perplexity was computed on the test set derived from Project Gutenberg (see Chapter 6). For panel D, we used the text of *The Little Prince* novella to compute GPT-2's perplexity. The untrained checkpoint is not represented because its perplexity was too high. Unlike what would have been expected from other studies (Schrimpf et al., 2020), that state that model's perplexity correlates with model's brain score, we found no clear relation. For example, in panel C, the 1- and 2-layer versions of GPT-2 have decreasing perplexity after the first epoch, and decreasing brain score. On the contrary, the 4-layer model obtains higher brain score while its perplexity decreases. For LSTM, while the perplexity keeps decreasing, there is no monotonous variation of the brain score which seems to oscillate. Finally, when computing the perplexity and brain score on the same dataset (The Little Prince), the brain score decreases with training after the first epoch.

Overall, the first epoch of training brings brain's and model's representations closer, however, the following epochs show no clear pattern. At a certain point, it seems that the training objective lead the latent representations further away from brain like representations.

Model	GPT-2	BERT
LSTM	79%	86%
GPT-2	.	85%

Table 8.1: **Overlap between training effect brain maps.** The percentage of common voxels when the maps were thresholded at their 10% upper percentile. There is a 75% overlap between the maps of all 3 models.

Model	LSTM	GPT-2	BERT
LSTM	.	81%	92%
GPT-2	.	.	86%

Table 8.2: **Overlap between untrained brain maps.** The percentage of common voxels when the maps were thresholded at their 10% upper percentile. The overlap between the three maps is 79% across all 3 models.

[3]mean R score across the SRM25 voxelset

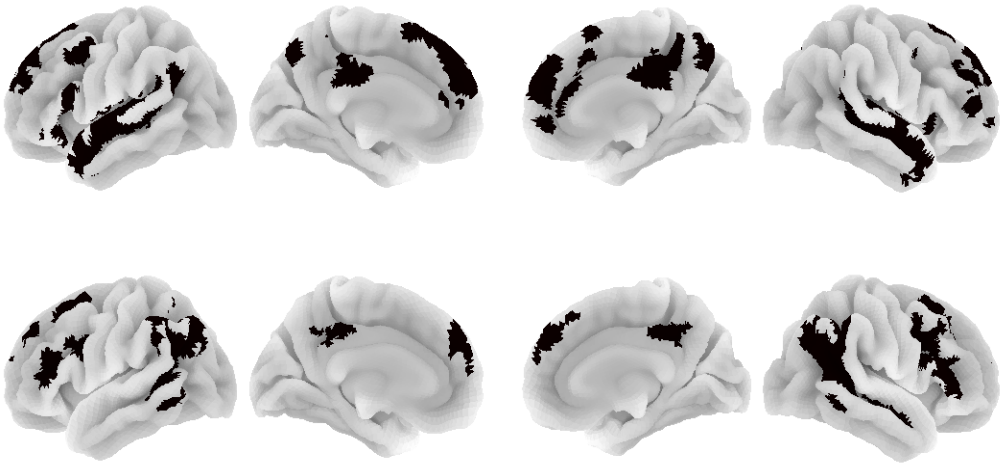


Figure 8.5: **The Training Gain Overlap and the Untrained Overlap.** (Top) Overlap between untrained brain maps. (Bottom) Overlap between training gain brain maps.

8.2.2 . The relation between Perplexity and Brain score is not monotonous

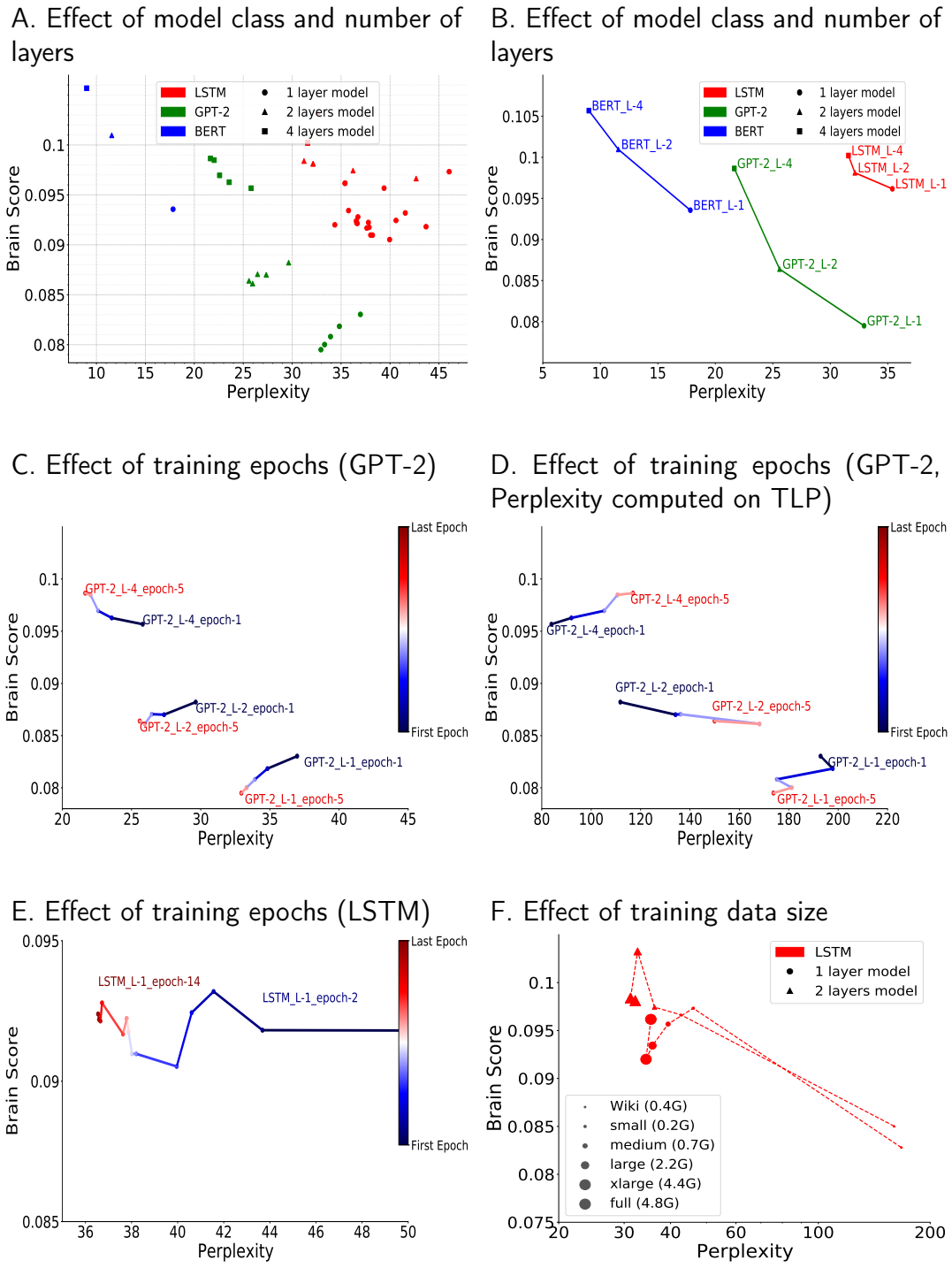


Figure 8.6: **Detailed analyses of the relation between brain score and perplexity as a function of model class (B), number of layers (B), training epochs (A-E) and training datasets (F).**

Finally, we investigated the general relation between perplexity and brain score. Using the set of trained LSTM, GPT-2 and BERT models, we evaluated them using the standard loss, that is, the average logarithm of model perplexity, computed on the test set (from Project Gutenberg). For each model, we also computed the brain score, defined as the average R-value within the SRM25 voxelset.

Fig.8.6A shows the relationship between perplexity (model loss) and brain scores derived from various models, architectures, training sets and training stages. Unlike previous reports (Schrimpf et al., 2020), we did not observe a clear monotonic relationship between the two variables (see Fig. 1.5). For example, the average LSTMs perplexity is worse than that of GPT-2, but the average brain score is higher. We investigated in more details the effects of model class, number of layers, training epochs and training dataset size on the relationship between brain score and perplexity. The results are presented in Fig.8.6. In Fig.8.6 panel B, we observed that within each model class, increasing the number of layers improves perplexity and brain score. However, as previously described, within a given model class, there is not always a monotonic relationship between brain score and perplexity as shown by the effect of training epochs in panels C and D for GPT-2 and panel E for LSTM.

8.2.3 . The training set has a significant effect on the prediction of fMRI brain data

Finally, we investigated the importance of the training data on the model ability to fit the brain data of the English participants of the 'The Little Prince' fMRI corpus. We first explore the relation between perplexity and brain score for LSTMs trained on different datasets. We used Wikipedia[4] (0.4G), the training dataset derived from Project Gutenberg (described in Chapter 6), refered to as the *xlarge* dataset (4.4G), and we also defined subsets of the *xlarge* dataset, namely the small (0.2G), medium (0.7G) and large (2.2G) datasets. Where small $\subset$ medium $\subset$ large $\subset$ xlarge $\subset$ Full, and Full = Wikipedia + xlarge. Then, we compared models trained on Wikipedia and on the Full dataset (for LSTM and GloVe).

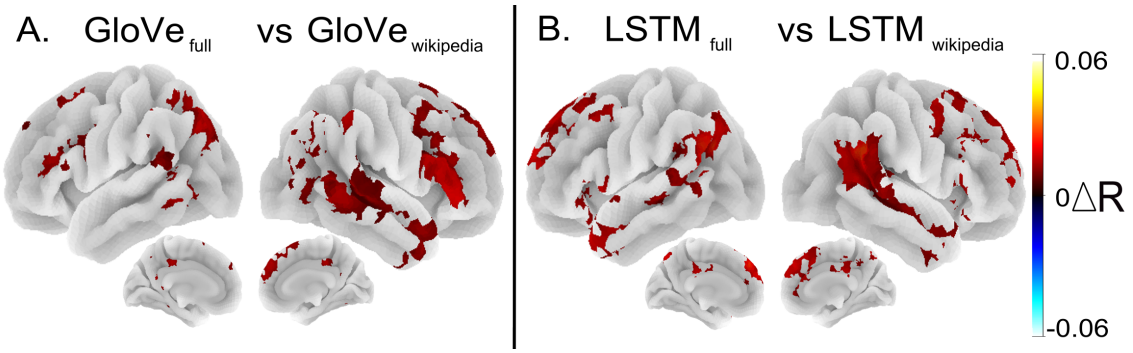


Figure 8.7: **Influence of Training dataset on R_{test}.** LSTM and GloVe better fit brain data when learning on more training data. This shows the dependence of models contrast on training data. Our full dataset comprised Gutenberg + Wikipedia (4.8 GB) while Wikipedia represented 425 MB.

[4]http://dumps.wikimedia.org/enwiki/20140102/

Manipulating training dataset size with LSTM shows no obvious relationship between brain score and perplexity (Fig.8.6F). Overall, the data used for training have a strong influence on the outcome. Fig.8.7 presents contrasts maps obtained when training LSTM or GloVe with our custom Full dataset versus Wikipedia, showing significant differences in favor of the versions trained on the Full dataset. This shows that off-the-shelf models trained on Wikipedia likely lack statistical power to detect brain activations.

8.3 . Discussion

Previous work has shown that brain activity during visual or language processing can be significantly predicted from artificial neural-network activations (see Section 1.3 for a review). In the present chapter, we examined the limits of the connection between brains and ANNs.

We studied the impact of models' architecture (assessing GloVe and LSTM, GPT-2 and BERT models with varying number of layers), models' perplexity and training corpus, on their ability to predict functional Magnetic Resonance Imaging timecourses of participants listening to an audiobook. We made several observations: (1) there is an important amount of signal that is not explained by NLMs, (2) untrained versions of each model already explain significant amount of signal in the brain, with the untrained LSTM and GloVe outperforming the others; (3) training NLP models improves brain scores in the same set of brain regions irrespective of model's architecture; (4) Perplexity is not a good predictor of brain score; (5) training data have a strong influence on the ability to fit brain data.

One discovery is that all architectures are not equal, but training them consistently increases brain scores in the same set of brain areas. From a neuroanatomical point of view, they are located on the border of the core language regions (IFG and STS) and partly overlap with regions assigned to the Default mode network (Angular Gyri, Dorso mesial prefrontal cortex). In the two preceding chapters, we found that these very brain regions were involved in semantic processing and context integration. This is consistent with the fact that NLMs learn during training to build a semantic representation for each word that depends on the surrounding context. These results are coherent with previous work investigating the processing of contextual information by either using LSTM models (Jain and Huth, 2018) or by scrambling the stimuli at different levels (Lerner et al., 2011), confirming that this network (in green/blue in Fig.8.4) is at the center of combinatorial language processing in the human brain. However, this convergence is partial. Investigating the relation between model loss and model brain score, we found that optimizing one does not mean optimizing the other one.

We observed that even if transformers start with a disadvantage regarding the ability to fit fMRI brain data, they benefit more from training than LSTMs, and they are able to take advantage of stacks of layers to improve their fitting performance. The comparison of untrained LSTM and untrained GloVe (i.e., random embeddings) showed no significant differences, whereas the comparison of untrained GloVe and untrained transformers showed significant R-score differences in some regions. The difference between untrained LSTMs and untrained Transformers might be due to their different architectures. However, there is an alternative explanation. Note that for untrained GloVe (random embeddings), each word in the corpus is assigned a fixed vector, whereas for untrained Transformers, each word is mapped to a variable vector, depending on the context that surrounds the word.

Therefore, untrained Glove better predicts brain responses to words that occur frequently both in the training and in the test data (e.g., function words). In contrast, untrained transformers generates different embeddings for the same word (e.g., 'the' in the train and test sets), due to their context sensitivity. This variability reduces the brain score of untrained Transformers compared to that of untrained GloVe. Thus, the ability of untrained models to fit brain data is not related to their architecture (unlike Schrimpf et al. (2020) stated), but rather to their ability to consistently identify the same words in the train and test sets. This result was confirmed by the use of a random baseline which associates a fixed random embedding to each token instead of each word. Indeed, the fitting performance dropped to zero when fitting brain data with these randomly generated embeddings. Finally, our results suggest that untrained LSTM are more similar to untrained Glove, having less context sensitivity compared to Transformers. The weaker LSTM sensitivity to context compared to transformers justifies the approach adopted in Chapter 7 on context-sensitivity. Taken together, this suggests that most of what the untrained baselines capture is similarity in brain responses to words that appear in both the train and test sets. Thus, the analysis of the untrained latent representations' predictive power highlights potential pitfalls of using complex NLMs to study brain data and interpreting without taking the time to understand all potential confounds.

We also observed an important amount of signal that is not explained by NLMs, reaching only 60% in the very best fitted voxels. What is the unexplained signal? Why doesn't the best model capture it? How could we capture it? Future work should investigate the 40% remaining signal that are captured by the model-free Shared-Response Model but not by the NLMs. Understanding this difference might shed light on limits inherent to the modelling approach or on the limits of current NLMs to fit brain data. If it is indeed a limitation of the NLMs, it could be related to known limitations of NLMs such as compositional generalisation (Dankers et al., 2022)[5], the ability to build a world-model (Ruis et al., 2020), learn rules and generalize to unknown environment (Bastings et al., 2018; Dessì and Baroni, 2019; Loula et al., 2018). Many studies have investigated compositional generalisation in language models and found that neural networks struggle to interpret compositions unseen in training. They can make successful zero-shot generalizations when the differences between training and test are small, but fail dramatically when generalization requires systematic compositional rules (Lake and Baroni, 2018; Lake et al., 2019; Ruis et al., 2020). Compositionality in neural language models is also constrained by their limited ability to integrate context which makes them lose contextual information over time. However, the steady increase in model sizes and training data sizes lead to better and better language models, like GPT-3 (Brown et al., 2020) for example.

Building better model that go beyond the previously listed limitations is a promising direction to understand the gap of the missing 40%. Several assumptions related to the modelling approach could also explain such a gap. First, one of the main assumption in the analysis of fMRI BOLD data is the linearity assumption. The seminal work of Boynton et al. (1996); Vazquez and Noll (1998) identified that the bold response behaved as a linear time invariant system for time durations separating events greater than 2 seconds. Such hypothesis simplifies drastically the subsequent analyses by allowing the use of the General Linear Model. However, during story listening, words occur at a sampling frequency of about 200ms, which might violate the linearity assumption of the elicited BOLD signal.

[5]that is the ability to combine concepts together

Therefore, under the linearity hypothesis, nonlinear interactions create a prediction error that may reduce sensitivity (Wager et al., 2005).

Finally, the discrepancy between brain score and perplexity indicates that training is not a guarantee of convergence towards brain-like representations (see also Hale et al. (2019)). Relatedly, other research also indicates that perplexity minimization is not a royal road to cognitive models (see, e.g., Clark (2000)). A last methodological word of caution stem from our results: data used for training have a strong influence on the outcome, showing that off-the-shelf models trained on small datasets, like Wikipedia, lack statistical power to capture brain activations and should be avoided to probe brain representations.

Overall, the convergence between brains and ANNs is partial and subject to many factors. These results call for caution when comparing model's and brain's representations.

9 - General Discussion

Through the use of functional Magnetic Resonance Imaging and latent linguistic representations, the studies presented in this document provide a thorough analysis of several neural mechanisms involved in language comprehension, shedding new light on representation similarities between brains and ANNs as well as on brain functions topography.

The assumption underlying our approach is that the information processed by a NLM and a brain region are probably similar if the latter is well fitted by the latent representations extracted from the NLM. This paradigm enables us to address questions such as: How does the structure and function of artificial neural networks compare to that of the human brain when processing language? To what extent are the brain areas involved in semantic and syntactic processing separated or intertwined? Which brain regions integrate information beyond the lexical level, and what is the size of the window of integration of such regions?

The investigation of the neural bases of language comprehension using artificial language models represents a complementary approach to traditional language neuroscience research. One advantage of this approach is its ability to investigate multiple aspects of language processing using a single naturalistic dataset. However, achieving complete control over all variables is challenging. Part of this work seeks to highlight the limitations and biases of ANN-powered naturalistic language processing.

9.1 . Summary of significant findings and contributions

In our view, the main contribution of this work to the field of neurolinguistics is the introduction of information-restricted NLMs to probe the neural bases of semantics, syntax and context-integration (Chapters 6 and 7). Information-restricted NLMs are customized NLM architectures trained on feature spaces containing a specific type of information. This work has shown that restricting the information provided to a model during training can reveal which brain regions encode that information.

The application of information-restricted NLMs can extend to a wide range of representational spaces, including syntax and semantics. From large amounts of text, several representational spaces can be crafted, each representing a dimension of language processing. By training information-restricted NLMs on each feature space, one can investigate the associated linguistic dimensions. The interest lies in the fact that a single naturalistic stimulus can later be used to probe a plethora of processes. There is no need to design a complex experiment per investigated process, which lessens the experimental burden and saves time. Using naturalistic stimuli also reduces the risk of biases induced by the task or the constraint put on the stimuli. It also allows to acquire larger neuroimaging datasets with easily available stimuli. Once trained, these models can generate information-restricted embeddings that can be used to fit brain activity in any dataset. This approach offers several advantages over classical approaches that rely on manually crafted features or focus on specific contrasts, including feature richness and scalability. In future experiments, fine-grained control over the information given to the models and their representations will enable a more precise characterization of the various regions involved in language comprehension.

The specific application of information-restricted NLMs to the analysis of syntactic and semantic processing resolved previous disagreements regarding their spatial organization. Information-restricted models were able to explain both the widespread, bilateral, network of semantic- and syntactic- sensitive brain regions that extends beyond the traditional language network[1] (Caucheteux et al., 2021; Fedorenko et al., 2020; Huth et al., 2016; Pereira et al., 2018), as well as the more localized semantic- and syntactic-sensitive brain regions found by traditional experiments using constrained stimuli (Friederici, 2011, 2017; Pallier et al., 2011). To assess the validity of both views, our approach was to quantify the varying relative degrees of sensitivity to syntax and semantics by defining a *specificity index* (see Chapter 6), but also by removing the shared component between syntactic and semantic embeddings. These manipulations revealed a higher sensitivity to syntax in the left IFG, STG/STS and anterior temporal lobe (bilaterally) (e.g., Friederici, 2011, 2017), and a higher sensitivity to semantics in parietal regions, consistent with Binder et al. (2009). In summary, sensitivity to semantics and syntax is widespread over both hemispheres, but with different patterns of sensitivity. Specifically, the brain regions that are most attuned to syntax tend to be more localized.

Secondly, the analysis of brain regions sensitivity to syntax and semantics revealed different patterns of sensitivity across hemispheres. The regions that are best predicted by syntactic and semantic features overlap substantially in the right hemisphere, while they are almost dissociated in the left hemisphere.

The second contribution of this work is the demonstration that 'surgical' operations can be performed on complex neural language model architectures in order to probe precise linguistic processes. A 'surgical' operation on a neural language model refers to modifications of its architecture (e.g., unit ablation), or of its internal operations (e.g., modification of the attention mechanisms). Chapter 7 illustrated it on context-sensitivity with both masked-attention generation and information-restricted models. Both approaches yielded consistent evidence of context-sensitivity across most of the cortex (Jain and Huth, 2018) and revealed a further differentiation between the left and right hemispheres. Specifically, it was observed that the right hemisphere selectively incorporates longer contextual information, whereas the left hemisphere selectively incorporates shorter contextual information. Both methods were able to generate brain maps indicating the size of context over which each context-sensitive brain region integrates contextual information, leading to results consistent with Lerner et al. (2011) and Jain and Huth (2018), and finding that medial regions are core components for the integration of contextual information.

Two major findings of this work are: 1) the fact that perplexity is not a good predictor of brain score, and 2) the influence of the NLMs' training data on their fitting performance. The absence of clear relation between brain score and perplexity suggests that simply training a model is not a guarantee of achieving brain-like representations (see also Hale et al. (2019)). To be more specific, the training of neural language models initially leads to a convergence of brain and model representations, resulting in improvements in both brain score and perplexity. However, after a certain point, perplexity continues to improve while brain score begins to decline. Our findings, along with those of Caucheteux and

[1] that includes the IFG and temporal regions

King (2022), support this observation. The correlation between perplexity and brain score observed in Schrimpf et al. (2020), albeit not so high, seems to be mainly driven by the GPT-2 models (ibid, Figure 1-5). This aligns with previous research on the limitations of using perplexity as a measure of cognitive models (see, e.g., Clark (2000)). Our results also show that the nature and size of the data used for training greatly affects the outcome, indicating that using off-the-shelf models trained on small datasets such as Wikipedia is not effective in capturing brain activations and should be avoided when studying brain representations. Still, the capacity to fit the same set of brain regions during training is enhanced irrespective of the architecture and training data employed. These findings highlight the importance of meticulously planning experiments that regulate the design and training of ANNs.

Another contribution relates to the rigorous comparison of NLMs' ability to fit fMRI brain data. Chapter 5 performed a comprehensive comparison of NLMs taken off-the-shelf or trained under controlled conditions. Our findings replicate Wehbe et al. (2014a)'s and Huth et al. (2016)'s results, indicating that the latent representations of NLMs are capable of explaining signal bilaterally across a broad network of brain regions, including temporal, frontal, parietal, and medial areas. Notably, we observed high correlations in the right hemisphere, which is consistent with previous naturalistic conditions research. By varying NLMs' architecture, size and training data, these results gave broad insights on NLMs' ability to fit fMRI brain data. These findings complete the work of Schrimpf et al. (2020), who ran a systematic comparison of 43 models. While Schrimpf et al. (2020) conducted a larger model benchmark, it is important to note that the models used were taken off-the-shelf and not trained on the same dataset, with the same vocabulary size, which introduces bias into the comparison, as shown in Chapter 5. Despite this, we observed that transformers outperformed LSTMs, even though LSTMs have more trainable parameters for the same vocabulary size and hidden dimension, as shown in Table 3.1. Our results suggest that attention mechanisms represent a significant improvement over traditional incremental recurrent neural networks by allowing for better integration of contextual information into word embeddings. Although transformers better fit brain data, no interaction between their architecture and brain regions could be observed. In other words, no brain region is better suited to fit specific submodules of transformers: regardless of the attentional head or layer used, the predictive performance of brain regions are ordered in the same way.

Finally, the last contribution addresses the estimation of the ceiling of explainable signal and the size of the fMRI dataset that should be used. Studies investigating the fitting performance of neural language models report results as percentage of a ceiling of explainable signal (Caucheteux and King, 2022; Millet et al., 2022; Schrimpf et al., 2020). While this approach allows to better contextualize models' ability to fit brain data, it is likely biased by the modelling ability of the estimation method used to find the ceiling. Results from Chapter 2 shows that the standard estimation approach using Inter-Subject Correlation is outperformed by the Shared Response Model estimation method. The relative difference between these two estimations is not the least, as it is above 20% everywhere in the brain. As a consequence, the results reported in studies using Inter-Subject Correlation, such as Schrimpf et al. (2020), are overestimated. Regarding the assessment of the influence of

the number of fMRI scans on the encoding model performance, it was shown in Chapter 4 that encoding models' performance keeps increasing beyond thousands of scans. More importantly, this increase occurs in language-related voxels. This analysis revealed the importance of using large neuroimaging datasets with thousands of scans per subjects, and shed doubts on conclusions drawn from studies leveraging a few minutes or tenth of minutes of neuroimaging data.

9.2 . Implications of the findings

This work and recent studies found that NLMs explain a significant amount of signal in brain data (Caucheteux and King, 2022; Schrimpf et al., 2020). However, the lower sensitivity of encoding models compared to the model-free SRM-based estimation of the ceiling of explainable signal (Fig. 8.3), shows the limits inherent to the modelling approach or to NLMs. To advance in the understanding of language processing, **it is important to develop more advanced neural language models** that can acquire rules, compositional understanding, as well as generalize to novel environments and even construct a model of the world. If such findings cannot bridge the gap of the undetected brain activations, then, this might suggest an issue in the modelling approach. For example, nonlinear interactions between the brain activations, elicited from rapidly occurring words, might create prediction errors that may reduce sensitivity (Wager et al., 2005). Additionally, the influence of the number of scans per subject on the encoding model performance encourages researchers to **use large neuroimaging datasets (in the number of time-points per participant)**. Indeed, the average performance in Fig. 4.1 does not reach a plateau, even with more than 2000 scans, indicating that future neuroimaging datasets should contain hours of data per subject.

The investigation of the structural and functional similarity between brains and transformers revealed no interaction between transformers' architecture and brain regions, ruling out these architectures as direct explanatory models of the brain (see Fig. 5.9). Moreover, the state-of-the-art 12-layer BERT model's predictive performance only exceeded slightly that of our customized 4-layer BERT model, which was trained on a smaller dataset (see Fig. A.7). This finding hints that the quest for bigger and bigger language models does not bring us closer to understanding the brain. On the contrary, this provides impetus for exploring smaller and more brain-like architectures, as *developing more biologically plausible and interpretable models could aid in our comprehension of language processing as a whole*. For instance, performing 'surgical' operations on brain-like architectures could enable more precise investigations into the neural bases of cognition. Future work could design brain-like language model architectures using Neural Architecture Search (Liu et al., 2019; Luo et al., 2019; So et al., 2019) and objective functions based on fMRI encoding performance or connectivity as well as classical NLP tasks (such as masked-token prediction or next-token prediction).

Beyond the architecture of the models, the input format is likely to play an important role, as there is a mismatch between the continuous inputs received by the brain and the symbolic representations given to NLMs. A recent attempt to get more biologically-plausible inputs can be found in Vaidya et al. (2022) and Millet et al. (2022) who used

the audio stimuli as input to both NLMs and human participants. Using these models, they were able to highlight a hierarchy of brain regions processing language. In a broader sense, **multi-modal models that construct a comprehensive feature space shared across input modalities (text, audio, image) exhibit great potential for the modelling of brain data**.

Finally, in addition to the model architecture and input format, the amount of training data has a strong influence on the model ability to fit fMRI brain data (see Chapters 5 and 8). Fitting brain data with a model trained on a small dataset leads to a lack of statistical power to detect brain activations, and therefore to a questionable interpretation (see Fig. 8.7). As a consequence, studies training NLMs on Wikipedia, such as Caucheteux and King (2022), likely used underperforming models, which might have affected model comparisons or ROI-analyses. Thus, **it is necessary to use NLMs trained on large enough datasets**.

Result interpretation is also highly dependent on potential confounds and variables of non-interest. The investigations of the effects of the basic features (Chapter 5) and the effects of untrained NLMs (Chapter 8) is a warning to researchers, encouraging them to **find optimal controls in order to isolate effects of interest**.

Lastly, we found that our understanding of the neural bases of syntax, semantics and context-integration was improved thanks to the use of information-restricted NLMs. These results are an incentive to **design (better) feature spaces to probe specific brain processes**. They illustrate how the manipulation of the architecture, the parameters, and the training data and objective of these models can be used to explore the different processing strategies and mechanisms during language comprehension.

9.3 . Limitations and future directions

Several limitations of this work have to be acknowledged.

The first one is the absence of replication on different stimuli. Even if 'The Little Prince' fMRI dataset is a large corpus of 51 English participants, from which 90 minutes of data were recorded, there is still a risk that the results do not replicate on another dataset. Future work will intent to replicate these results on the french participants or other fMRI corpora.

One of the major limitations of the approach resides in the lack of biological plausibility of the NLMs. Their complex architectures, and lack of interpretability are barriers to the optimal use of computational model to study the brain. Since they do not accurately reflect the neural processes that occur in the human brain, their ability to provide insights into the neural mechanisms underlying language comprehension is limited. Despite their complexity, current Neural Language Models neglect some aspects of language processing. Natural language processing involves various cognitive and neural processes that interact and operate simultaneously. As an illustration, during language processing, humans may trigger associative responses such as memory retrieval, which can evoke emotions. Additionally, humans experience language (and stimuli in general) in rich and multi-modal environments, whereas NLMs solely receive symbolic linguistic information, learning only joint probability distributions over the vocabulary [2]. This can limit the ability of NLMs

[2]However, some recent models are multi-modal **??**.

to build rich representational spaces like the brain.

NLMs have a limited ability to acquire rules (Bastings et al., 2018; Dessì and Baroni, 2019; Loula et al., 2018), to understand compositionality[3](Dankers et al., 2022), as well as to generalize to novel environments (Ruis et al., 2020); they are not tied to world models. While they may be able to make accurate predictions in some cases where the differences between their training and testing data are minor, they tend to fail badly when trying to make sense of more complex combinations that require the use of systematic rules (Lake and Baroni, 2018; Lake et al., 2019; Ruis et al., 2020). Furthermore, because neural language models have a limited ability to integrate context, they may lose important contextual information over time.

Thus, NLMs may struggle to capture the variability and complexity of language use across different contexts and situations, which can limit their ability to accurately model the neural mechanisms involved in language comprehension. Nevertheless, as language models grow in size and the amount of data used to train them increases, they are becoming increasingly capable, as demonstrated by models such as GPT-3 (Brown et al., 2020). Overall, while neural language models have the potential to provide insights into the neural mechanisms underlying language comprehension, they still face a number of limitations that need to be addressed to advance our understanding of the neural bases of language processing.

Aside from NLMs, the semantic and syntactic feature spaces designed in Chapter 6 were not flawless. For instance, the removal of function words led to damaged supra-lexical semantics, while the omission of pronouns caused the loss of subject-related information (see 6.3). For example, "he cried: “What? You fell from the sky!”" becomes "cried fell sky". The lexical semantic building blocks are preserved, but the way they combine together to converge to the global meaning is mostly lost: we do not know if someone cried while falling from the sky, if someone was crying while another person was falling from the sky, etc... Ideally, one would like to create sentence transformations that preserve the semantic information associated with function words or pronouns. More generally, the optimality/purity of the feature space design on which neural models are trained is key to highlighting precise linguistic processes.

Interpretation can also be limited by potential confounding effects such as prosody. It cannot be ruled out that syntactic embeddings correlated with prosodic variables (Bennett and Elfner, 2019). For example, word accentuation can correlate with word length which can also correlate with word category: determiners or pronouns are shorter and less accented than others. Additionally, models learn to predict end of sentences, which are often followed by pauses in the acoustic signal. To minimize the impact of such factors, acoustic energy and the words' offsets were included in the baseline models. However, these controls cannot be entirely perfect. One possible solution to address this issue would be to have participants read the text, presented at a fixed presentation rate. This would effectively eliminate all low-level effects of prosody.

Finally, one last limitation would be individual differences. The neural bases of language processing varies across individuals due to anatomical and functional differences (Amunts et al., 1999; Prat et al., 2007). The latter can be influenced by factors such as age, education and language background (Malik-Moraleda et al., 2022). The difficulty to

[3]the ability to construct larger linguistic expressions by combining simpler parts, beyond semantic-only composition

control for these individual differences can limit the generalizability of the findings. However, these discrepancies between participants could be reduced with functional alignment methods (Thual et al., 2022).

9.4 . Concluding remarks

Artificial neural networks (ANNs) have been increasingly used to model and simulate the neural processes involved in language comprehension. These models build on the idea that the structure and function of the human brain can be mimicked by computational models composed of interconnected nodes that process information through a series of weighted connections. One of the main advantages of ANNs is their ability to provide insights into the neural mechanisms underlying language comprehension that are difficult or impossible to obtain using traditional neuroimaging techniques alone. For example, ANNs can be used to identify the specific features of linguistic stimuli that activate different brain regions, and to model the interaction between syntax, semantics, and pragmatics in language comprehension. Additionally, ANNs can be used to test different hypotheses about the neural bases of language comprehension, such as whether certain brain regions are specialized for processing specific aspects of language (e.g., syntax or semantics) or whether the brain uses distributed networks to integrate information from multiple sources. However, it is important to note that ANNs are not a realistic model of the human brain, and there are limitations to what can be inferred from these models. For example, ANNs do not accurately capture the dynamic and complex nature of neural processes, and may oversimplify or overlook important features of human language comprehension (e.g. taking as input symbolic representations). Overall, the use of ANNs in the investigation of the neural bases of language comprehension is a promising approach that can provide complementary insights to traditional neuroimaging techniques, but also requires careful consideration of the limitations and assumptions of these models.

In conclusion, this thesis has explored the neural bases of language comprehension using latent linguistics representations. The joint use of techniques from machine learning and neuroscience has provided novel insights into the spatial organization of syntactic and semantic processing as well as given a comprehensive understanding of context-sensitivity in the human brain. This work highlights the importance of 1) controlling the design and training of the computational models used, 2) using large training corpus to train the models on, as well as large neuroimaging datasets, and 3) seeking interpretable and biologically-plausible computational models. The design of neuroscience-inspired language models and their use in studying brain activations are promising avenues for future research. Ultimately, I hope that this work will contribute to a more comprehensive understanding of how the human brain processes language, and pave the way for new approaches to study the brain with artificial neural networks.

A - Supplementary Information

A.1 . Abbreviations

Brain Regions

- STG: superior Temporal Gyrus
- STS: superior Temporal Sulcus
- TP: Temporal Pole
- IFG: inferior Frontal Gyrus
- IFS: inferior Frontal Sulcus
- dmPFC: Dorso-Medial Prefrontal Cortex
- pMTG: posterior Middel Temporal Gyrus
- TPJ: temporo-parietal junction
- pCC: posterior Cingulate Cortex
- AG: Angular Gyrus
- SMA: Supplementary Motor Area

Other

- NLP: Natural Language Processing
- NLM: Neural Language Model
- LM: Language Model
- ANN: Artificial Neural Network

A.2 . Analyses Reproducibility

All analyses, as well as model training, features extraction and the fitting of encoding models were performed using Python 3.7.6 and can be replicated using the code provided in the same Github repository (`https://github.com/AlexandrePsq/Information-Restrited-NLMs`). The required packages are listed there. A non-exhaustive list includes Numpy (Harris et al., 2020), Scipy (Virtanen et al., 2020), Scikit-learn (Pedregosa et al., 2011), Matplotlib (Hunter, 2007), Pandas (McKinney et al., 2010) and Nilearn (`https://nilearn.github.io/stable/index.html`).

A.3 . Chapter 5

A.3.1 . Information Redundancy

Comparison of the fitting performance of consecutive layers for BERT and GPT-2. We also represent the gain in fitting performance when stacking two layers compared to using only one.

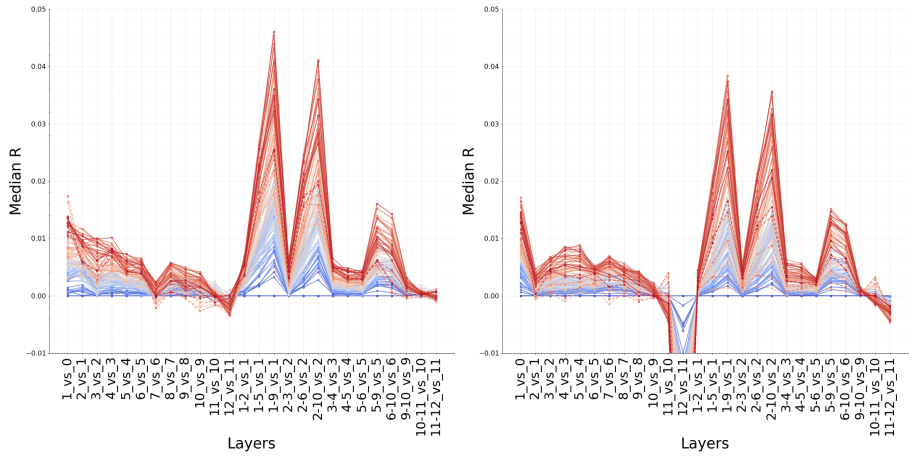


Figure A.1: **Comparison of the fitting performance of different layers from BERT and GPT-2.** Per ROI difference in R score between encoding models using the features derived from different layers (left: BERT, right: GPT-2).

A.4 . Chapter 6

A.4.1 . Models training

We trained GloVe and GPT-2 on syntactic or semantic features by adapting both vocabulary size and the associated tokenizer. Table A.1 provides examples of the features extracted from a short passage. Once features have been extracted from a text corpus, a vocabulary listing all possible feature instances is created for each feature type. A unique id is then associated to each element of the vocabulary. The tokenizer converts each feature to its unique id. Finally, the model is fed sequences of ids and learns to perform its task.

		Input sequence						
Integral Features		The	sixth	planet	was	ten	times	larger
Syntactic Features	Part-of-Speech	DET	ADJ	NOUN	VERB	NOUN	NOUN	ADJ
	Morphology	Definite=Def|PronType=Art	Degree=Pos	Number=Sing	Ind|Sing|Past|Person=3|Fin	Number=Card	Number=Plur	Degree=Cmp
	Number of Closing Nodes	1	1	2	1	1	2	2
Semantic Features	Content words	–	sixth	planet	–	ten	times	larger

Table A.1: Examples of input sequences given to the neural language models when trained on the different feature spaces.

The Morphology field contains a list of morphological features, with vertical bar (|) as list separator and with underscore to represent the empty list. All features represent attribute-value pairs, with an equals sign (=) separating the attribute from the value. In addition, features are selected from the universal feature inventory (https://universaldependencies.org/u/feat/index.html) and are sorted alphabetically by attribute names. It is possible that a feature has two or more values for a given word: Case=Acc,Dat. In this case, the values are sorted alphabetically.

Note: for display purposes, the morphology attribute values were removed for 'was', it was originally equal to

'Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin'.

A.4.2 . Convergence of the language models during training

In Chapters 5, 6, 7 and 8, GPT-2 models were trained on the *Integral Features*, the *Semantic Features* or the *Syntactic Features*. Fig. A.2 shows the convergence of these GPT-2 models during training.

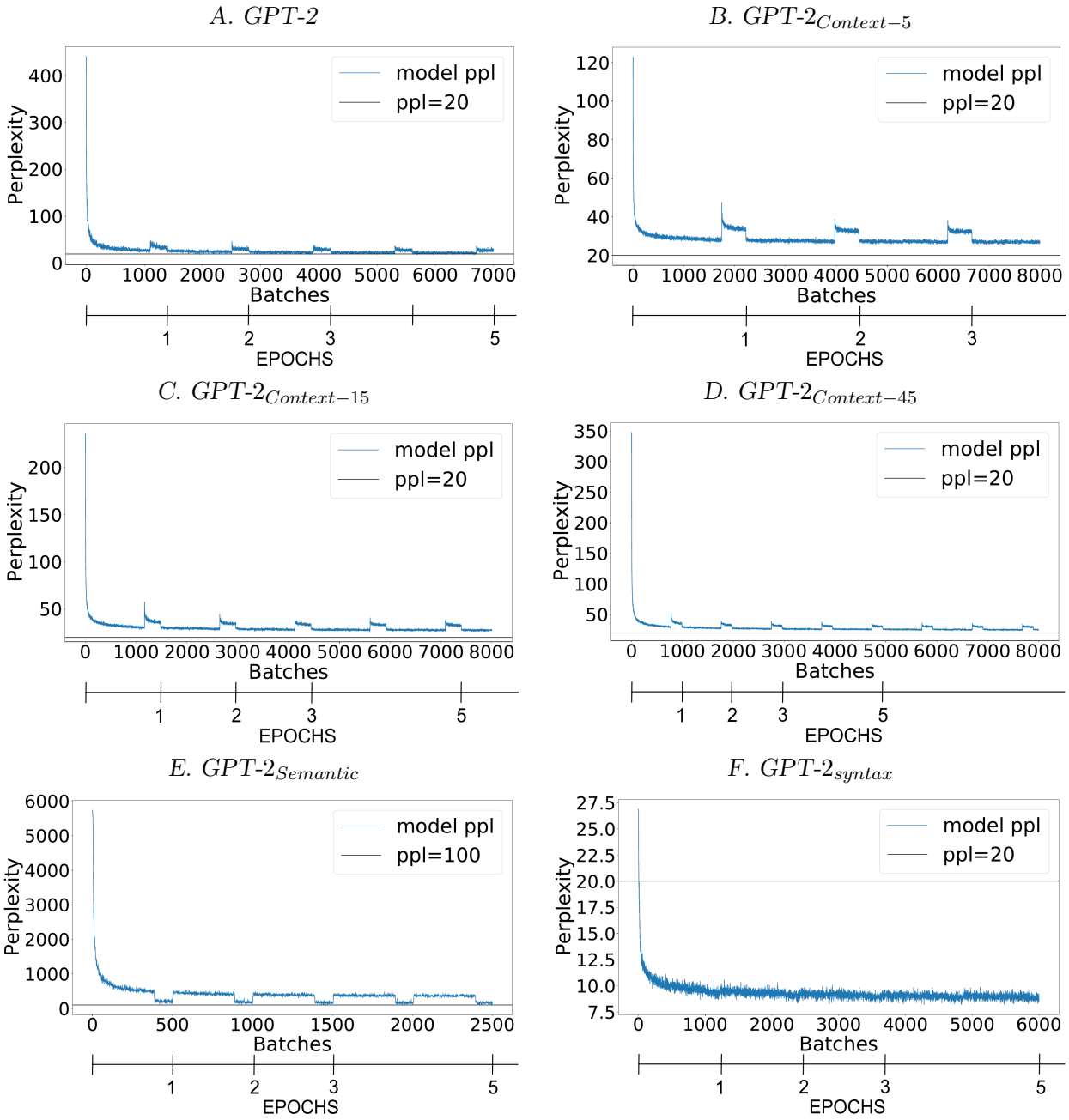


Figure A.2: **Model convergence during training.** The models represented in panels A to D were trained on the integral features. Models in panels E and F were respectively trained on the semantic and syntactic features.

A.4.3 . The Basic Features baseline model

To assess the specific impact of NLMs' embeddings, the maps shown in Fig.6.3 display *increases in R values* relative to a *baseline model* which comprised three variables of non-interest:

- acoustic energy (root mean squared of the audio signal sampled every 10ms)
- word offsets (one event at each word offset)
- log of the lexical frequency of each word (modulator of the words events).

More generally, as we looked at increases in R scores between models, the baseline model was appended to all other models studied in order to cancel out the effects of the 3 features of non-interest. Appendix - Fig.A.3 below displays the cross-validated correlations obtained from this baseline model.

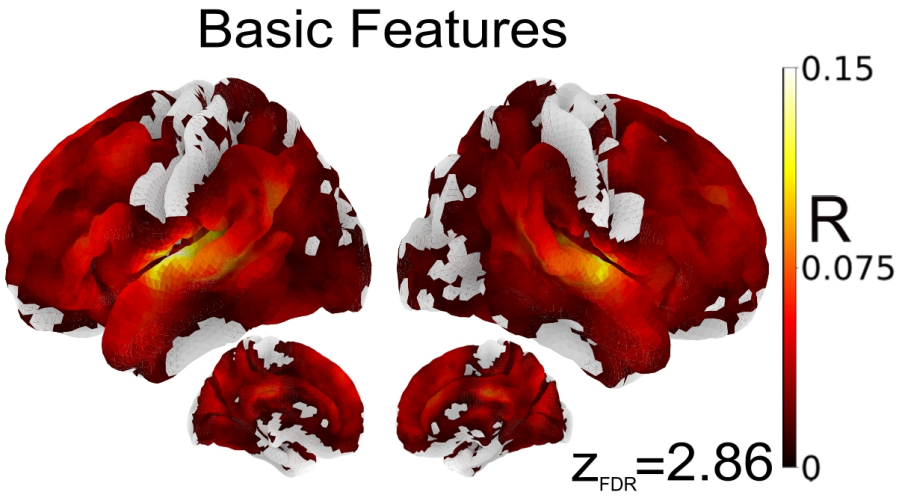


Figure A.3: **Brain regions showing significant activations for the Basic Features baseline model.** Using the Basic Features (BF) baseline model to fit fMRI brain data, we displayed voxels where there was a significant correlation (voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with a FDR approach $p < 0.005$; z_{FDR} is the FDR threshold on the z-scores). The effects from the Basic Features baseline model were discarded from all the analyses in the paper.

A.4.4 . Brain fit of GloVe and GPT-2 when trained on the Integral Features

Appendix - Fig.A.4 shows the increase in R relative to the baseline model, when using the embeddings of GloVe and GPT-2 trained on the Integral Features, that is, the intact text.

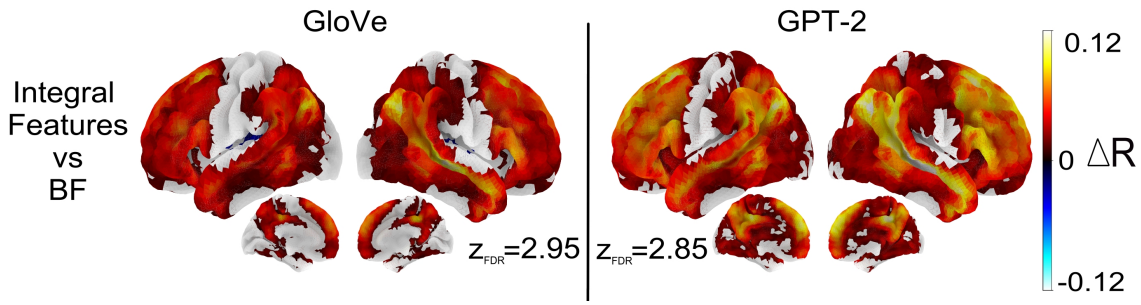


Figure A.4: **Brain regions showing significant R score increases compared to the Baseline Model for GloVe and GPT-2 when trained on the Integral Features.** Increases in R scores relative to the baseline model for GloVe (a non contextual model) and GPT-2 (a contextual model), trained on the Integral features (voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with a FDR approach $p < 0.005$; z_{FDR} is the FDR threshold on the z-scores).

A.4.5 . R Scores Distribution for GloVe and GPT-2 Trained on Semantic or Syntactic Features

Appendix - Fig.A.5 shows the distributions of the R scores increases across voxels (averaged across participants), obtained from GloVe and GPT-2 trained on semantic or syntactic features, relatively to the baseline model.

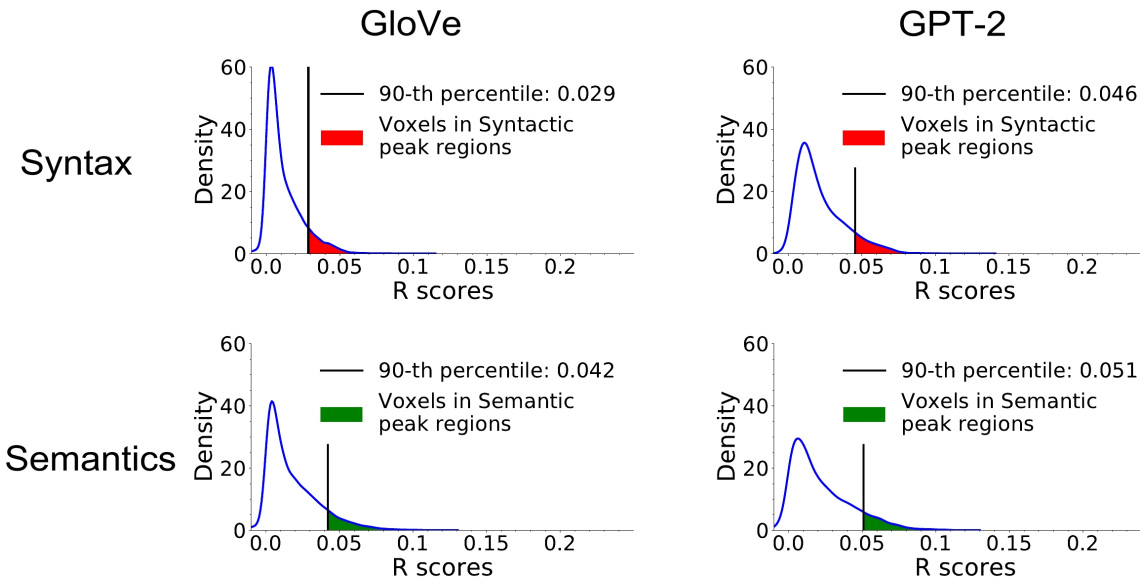


Figure A.5: **Distribution of R scores derived from GloVe and GPT-2 semantic and syntactic embeddings.** The 90th-percentile of the R scores distribution is highlighted with a vertical black line and used to select voxels for the peak regions analyses.

A.4.6 . Comparison of the models trained on Semantic features with the models trained on Syntactic features

Appendix - Fig.A.6 shows the differences in R scores between the semantic and syntactic models, for Glove and GPT-2. Correcting for multiple comparisons (N=51; $p < 0.005$ after FDR correction), we observed significant differences in favor of the syntactic embeddings in the STG, and significant differences in favor of the semantic embeddings in the pMTG, the AG and the IFS and SFS.

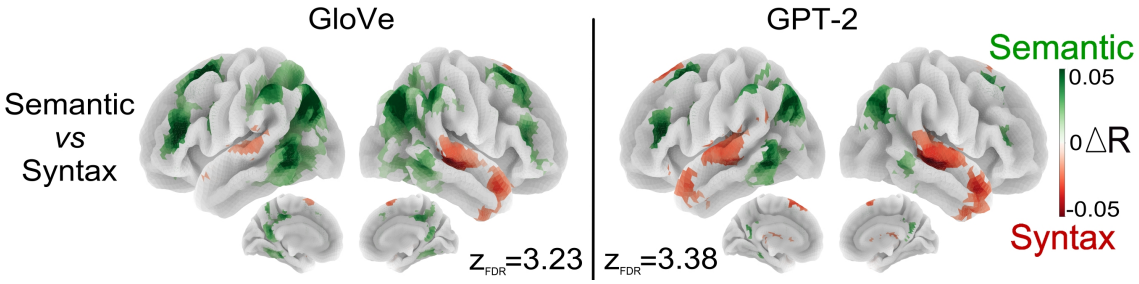


Figure A.6: **Comparison of the models trained on Semantic features with the models trained on Syntactic features.** Significant R score differences between the models trained on Semantic features and the models trained on Syntactic features. The brain regions that are better fitted by the former model appear in green, while the regions better fitted by the latter model appear in red. (All these maps represent voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with a FDR approach $p < 0.005$).

A.4.7 . Meta-Analysis based on Neurosynth

We used the *Neurosynth* database (https://github.com/neurosynth/neurosynth) to perform a meta-analysis of brain regions that appeared in fMRI articles containing the words 'syntactic' or 'semantic' in their abstract. Using a frequency threshold of 0.05, the keyword *semantic* yielded 626 articles, while *syntactic* yielded 128 articles.

The *meta.MetaAnalysis* function from the neurosynth package was then used to create association test maps for syntax and semantics. These maps display voxels that are reported more often in articles that mention the keyword than articles that do not. Such association test maps indicate whether or not there's a non-zero association between activation of the voxel in question and the use of a particular term in a study. We fused the maps associated to *syntactic* and *semantic*, thresholded with a False Discovery Rate set to 0.01, to produce Fig.6.6.

A.5 . General Discussion

In this section we compared the fitting performance of our custom-trained BERT models with pre-trained open-source BERT models. The comparison of 2-layer BERT models in Fig. A.7A showed that our model outperforms the model pre-trained by GOOGLE, with R score differences on the order of 0.02. The comparison of 4-layer BERT models in Fig. A.7B showed that our model is on par with the one pre-trained by GOOGLE. Finally, comparing our small 4-layer BERT model with the 12-layer GOOGLE BERT model (Fig. A.7C) showed that our model is slightly outperformed by the model pre-trained by GOOGLE, with R score differences on the order of 0.01.

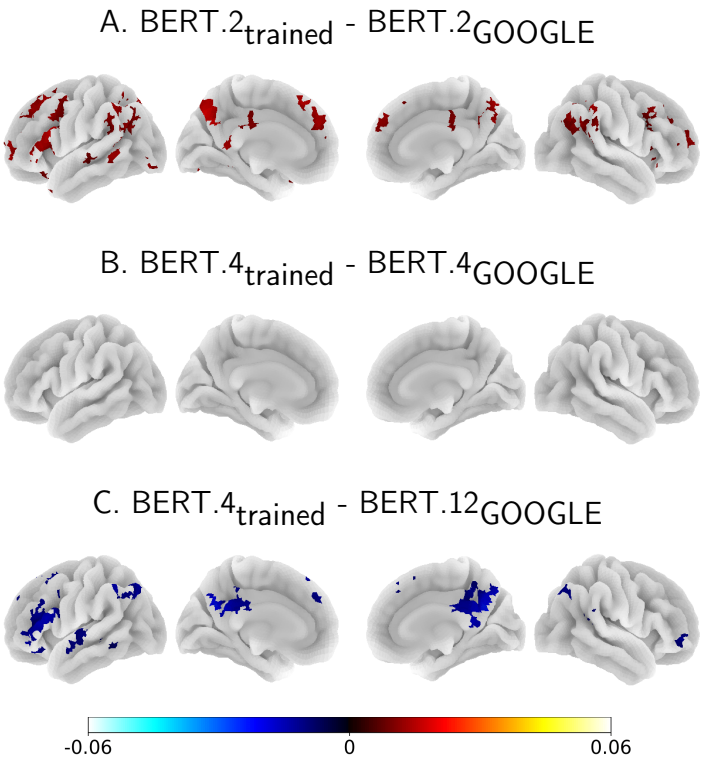


Figure A.7: **Comparison of the trained BERT models with off-the-shelf baselines.** To assess the performance of our trained models, we compared their ability to predict brain data with that of off-the-shelf models (`https://github.com/google-research/bert`). The 2 and 4-layers BERT models either significantly outperform the baseline or are on par. The 12-layers baseline, which is 3-times bigger than the 4-layers model, outperforms the latter in core regions of the language network, but only to a small extent. (All these maps represent voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with a FDR approach $p < 0.001$).

université
PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies
de l'information et de
la communication (STIC)

B - Synthèse

Titre: Déchiffrer les bases neurales de la compréhension du langage à l'aide de représentations linguistiques latentes

Mots clés: IRMf, Transformers, Traitement du Langage Naturel, Modèles linguistiques neuronaux, Modèles d'encodage, Apprentissage Profond

Synthèse: Au cours des dernières décennies, les modèles de langage (MLs) ont atteint des performances équivalentes à celles de l'homme sur plusieurs tâches. Ces modèles peuvent générer des représentations vectorielles qui capturent diverses propriétés linguistiques des mots d'un texte, telles que la sémantique ou la syntaxe. Les neuroscientifiques ont donc mis à profit ces progrès et ont commencé à utiliser ces modèles pour explorer les bases neurales de la compréhension du langage. Plus précisément, les représentations des ML calculées à partir d'une histoire sont utilisées pour modéliser les données cérébrales d'humains écoutant la même histoire, ce qui permet l'examen de plusieurs niveaux de traitement du langage dans le cerveau. Si les représentations du ML s'alignent étroitement avec une région cérébrale, il est probable que le modèle et la région codent la même information.

En utilisant les données cérébrales d'IRMf de participants américains écoutant l'histoire du Petit Prince, cette thèse 1) examine les facteurs influant l'alignement entre les représentations des modèles de langage et celles du cerveau, ainsi que 2) les limites de telles alignements. La comparaison de plusieurs modèles de langage pré-entraînés et personnalisés (GloVe, LSTM, GPT-2 et BERT) a révélé que les Transformers s'alignent mieux aux données d'IRMf que LSTM et GloVe. Cependant, aucun d'entre eux n'est capable d'expliquer tout le signal IRMf, suggérant des limites liées au paradigme d'encodage ou aux modèles de langage. En étudiant l'architecture des Transformers, nous avons constaté qu'aucune région cérébrale n'est mieux expliquée par une couche ou une tête d'attention spécifique. Nos résultats montrent que la nature et la quantité de données d'entraînement affectent l'alignement. Ainsi, les modèles pré-entraînés sur de petits ensembles de données ne sont pas efficaces pour capturer les activations cérébrales. Nous avons aussi montré que l'entraînement des modèles de langage influence leur capacité à s'aligner aux données IRMf et que la perplexité n'est pas un bon prédicteur de leur capacité à s'aligner. Cependant, entraîner les modèles de langage améliore particulièrement leur performance d'alignement dans les régions coeur de la sémantique, indépendamment de l'architecture et des données d'entraînement. Nous avons également montré que les représentations du cerveau et des modèles de langage convergent d'abord pendant l'entraînement du modèle avant de diverger l'une de l'autre.

Cette thèse examine en outre les bases neurales de la syntaxe, de la sémantique et de la sensibilité au contexte en développant une méthode qui peut sonder des dimensions linguistiques spécifiques. Cette méthode utilise des modèles de langage restreints en information, c'est-à-dire des architectures entraînées sur des espaces de représentations contenant un type spécifique d'information. Tout d'abord, l'entraînement de modèles de langage sur des représentations sémantiques et syntaxiques a révélé un bon alignement dans la plupart du cortex mais avec des degrés relatifs variables. La quantification de cette sensibilité relative à la syntaxe et à la sémantique a montré que les régions cérébrales les plus sensibles à la syntaxe sont plus localisées, contrairement au traitement de la sémantique qui reste largement distribué dans le cortex. Une découverte notable de cette thèse est que l'étendue des régions cérébrales sensibles à la syntaxe et à la sémantique est similaire dans les deux hémisphères. Cependant, l'hémisphère gauche a une plus grande tendance à distinguer le traitement syntaxique et sémantique par rapport à l'hémisphère droit.

Dans un dernier ensemble d'expériences, nous avons conçu une méthode qui contrôle les mécanismes d'attention dans les Transformers afin de générer des représentations qui utilisent un contexte de taille fixe. Cette approche fournit des preuves de la sensibilité au contexte dans la plupart du cortex. De plus, cette analyse a révélé que les hémisphères gauche et droit avaient tendance à traiter respectivement des informations contextuelles plus courtes et plus longues.

Bibliography

Katrin Amunts, Axel Schleicher, Uli Burgel, Hartmut Mohlberg, Harry B M Uylings, and Karl Zilles. Broca's region revisited: Cytoarchitecture and intersubject variability. *J. Comp. Neurol.*, 412(2):319–341, September 1999.

Kris Baetens, Ning Ma, Johan Steen, and Frank Van Overwalle. Involvement of the mentalizing network in social and non-social high construal. *Social Cognitive and Affective Neuroscience*, 9(6):817–824, June 2014. ISSN 1749-5016. doi: 10.1093/scan/nst048. URL https://doi.org/10.1093/scan/nst048.

Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, abs/2006.11477, 2020.

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.e5, 2017. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2017.06.041. URL https://www.sciencedirect.com/science/article/pii/S0896627317305937.

Jasmijn Bastings, Marco Baroni, Jason Weston, Kyunghyun Cho, and Douwe Kiela. Jump to better conclusions: SCAN both left and right. September 2018.

Elizabeth Bates and Frederic Dick. Language, gesture, and the developing brain. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 40(3):293–310, 2002. URL https://pubmed.ncbi.nlm.nih.gov/11891640/. Publisher: Wiley Online Library.

Elizabeth Bates and Brian MacWhinney. Functionalism and the competition model. In Brian MacWhinney and Elizabeth Bates, editors, *The Crosslinguistic Study of Sentence Processing*, pages 3–73. Cambridge University Press, 1989. URL https://www.researchgate.net/publication/230875840_Functionalism_and_the_Competition_Model/link/545a97170cf2c16efbbbc1d5/download.

Mark Jung Beeman and Christine Chiarello. *Right hemisphere language comprehension: Perspectives from cognitive neuroscience*. Psychology Press, 2013.

D K Bemis and L Pylkkänen. Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cereb. Cortex*, 23(8):1859–1873, August 2013.

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

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. ISSN 00359246. URL http://www.jstor.org/stable/2346101.

Ryan Bennett and Emily Elfner. The Syntax–Prosody Interface. *Annual Review of Linguistics*, 5(1):151–171, January 2019. ISSN 2333-9683, 2333-9691. doi: 10.1146/annurev-linguistics-011718-012503. URL https://www.annualreviews.org/doi/10.1146/annurev-linguistics-011718-012503.

Hans Berger. Über das elektrenkephalogramm des menschen. *Archiv f. Psychiatrie*, 108(3):407–431, June 1938.

Ellen Bialystok, Fergus I M Craik, and Gigi Luk. Bilingualism: consequences for mind and brain. *Trends Cogn. Sci.*, 16(4):240–250, April 2012.

Jeffrey R Binder and Rutvik H Desai. The neurobiology of semantic memory. *Trends Cogn. Sci.*, 15(11):527–536, November 2011a.

Jeffrey R. Binder and Rutvik H. Desai. The neurobiology of semantic memory. *Trends in Cognitive Sciences*, 15 (11):527–536, November 2011b. ISSN 13646613. doi: 10.1016/j.tics.2011.10.001. URL http://linkinghub.elsevier.com/retrieve/pii/S1364661311002142.

Jeffrey R. Binder, Rutvik H. Desai, William W. Graves, and Lisa L. Conant. Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex*, 19(12):2767–2796, 03 2009. ISSN 1047-3211. doi: 10.1093/cercor/bhp055. URL https://doi.org/10.1093/cercor/bhp055.

C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

Susan Bookheimer. Functional MRI of language: new approaches to understanding the cortical organization of semantic processing. *Annu. Rev. Neurosci.*, 25(1):151–188, March 2002.

L Boroditsky. Does language shape thought? mandarin and english speakers' conceptions of time. *Cogn. Psychol.*, 43(1):1–22, August 2001.

Gabriella Bottini, Rhiannon Corcoran, Roberto Sterzi, Eraldo Paulesu, Pietro Schenone, Pina Scarpa, Richard Frackowiak, and Chris Frith. The role of the right hemisphere in the interpretation of figurative aspects of language. a positron emission tomography activation study. *Brain : a journal of neurology*, 117 (Pt 6):1241–53, 01 1995. doi: 10.1093/brain/117.6.1241. URL https://www.researchgate.net/publication/15377772_The_role_of_the_right_hemisphere_in_the_interpretation_of_figurative_aspects_of_language_A_positron_emission_tomography_activation_study.

Geoffrey M. Boynton, Stephen A. Engel, Gary H. Glover, and David J. Heeger. Linear systems analysis of functional magnetic resonance imaging in human v1. *Journal of Neuroscience*, 16(13):4207–4221, 1996. ISSN 0270-6474. doi: 10.1523/jneurosci.16-13-04207.1996.

P Broca. Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). *Bulletin de la Société Anatomique*, 6:330–357, 1861.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Richard B. Buxton. *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511605505. URL https://radktob.files.wordpress.com/2017/05/richard_b-_buxton_introduction_to_functional_magbookzz-org1.pdf.

David Caplan, Nathaniel Alpert, and Gloria Waters. Effects of Syntactic Structure and Propositional Number on Patterns of Regional Cerebral Blood Flow. *Journal of Cognitive Neuroscience*, 10(4):541–552, July 1998. ISSN 0898-929X. doi: 10.1162/089892998562843. URL https://doi.org/10.1162/089892998562843. _eprint: https://direct.mit.edu/jocn/article-pdf/10/4/541/1931814/089892998562843.pdf.

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. Brains and algorithms partially converge in natural language processing. 2022. doi: 10.1038/s42003-022-03036-1. URL https://pubmed.ncbi.nlm.nih.gov/35173264/.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Disentangling Syntax and Semantics in the Brain with Deep Networks. In *ICML 2021 - 38th International Conference on Machine Learning*, Online conference, France, July 2021. URL https://hal.archives-ouvertes.fr/hal-03361421.

Claire H. C. Chang, Samuel A. Nastase, and Uri Hasson. Information flow across the cortical timescale hierarchy during narrative construction. *Proceedings of the National Academy of Sciences*, 119(51):e2209307119, December 2022. doi: 10.1073/pnas.2209307119. URL http://www.pnas.org/doi/full/10.1073/pnas.2209307119. Publisher: Proceedings of the National Academy of Sciences.

Po-Hsuan (Cameron) Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby, and Peter J Ramadge. A reduced-dimension fmri shared response model. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/b3967a0e938dc2a6340e258630febd5a-Paper.pdf.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. June 2014.

Andy Clark. *Mindware: An introduction to the philosophy of cognitive science.* Oxford University Press, 2 edition, 2000.

John Clarke. Squid fundamentals. In *SQUID Sensors: Fundamentals, Fabrication and Applications*, pages 1–62. Springer Netherlands, Dordrecht, 1996.

D Cohen. Magnetoencephalography: evidence of magnetic fields produced by alpha-rhythm currents. *Science*, 161 (3843):784–786, August 1968.

Ayanna Cooke, Edgar B. Zurif, Christian DeVita, David Alsop, Phyllis Koenig, John Detre, James Gee, Maria PinÃ£ngo, Jennifer Balogh, and Murray Grossman. Neural basis for sentence comprehension: Grammatical and short term memory components. *Human Brain Mapping*, 15(2):80–94, November 2001. ISSN 1065-9471. doi: 10.1002/hbm.10006. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6872024/.

Robert W. Cox. AFNI: software for the analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173, 1996. Publisher: Academic Press.

Fergus Craik and Robert Lockhart. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11:671–, 12 1972. doi: 10.1016/S0022-5371(72)80001-X.

Kamalaker Dadi, Gaël Varoquaux, Antonia Machlouzarides-Shalit, Krzysztof J. Gorgolewski, Demian Wassermann, Bertrand Thirion, and Arthur Mensch. Fine-grain atlases of functional modes for fmri analysis. *NeuroImage*, 221:117126, 2020. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2020.117126. URL https://www.sciencedirect.com/science/article/pii/S1053811920306121.

H. Damasio, D. Tranel, T. Grabowski, R. Adolphs, and A. Damasio. Neural systems behind word and concept retrieval. *Cognition*, 92(1-2):179–229, 2004. ISSN 0010-0277. doi: 10.1016/j.cognition.2002.07.001.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.286. URL https://aclanthology.org/2022.acl-long.286.

Peter Dayan and Laurence F Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems.* MIT press, 2005.

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. *The Journal of Neuroscience*, 37(27):6539–6557, July 2017. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3267-16.2017. URL http://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.3267-16.2017.

Gary S Dell. A spreading-activation theory of retrieval in sentence production. *Psychol. Rev.*, 93(3):283–321, 1986.

Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, Marilyn S Albert, and Ronald J Killiany. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, jul 2006.

Roberto Dessì and Marco Baroni. CNNs found to jump around more skillfully than RNNs: Compositional generalization in seq2seq convolutional networks. May 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv: 1810.04805.

Frederic Dick, Elizabeth Bates, Beverly Wulfeck, Jennifer Aydelott Utman, Nina Dronkers, and Morton Ann Gernsbacher. Language deficits, localization, and grammar: evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. *Psychological review*, 108(4):759, 2001. URL https://psycnet.apa.org/record/2001-18918-004. Publisher: American Psychological Association.

N F Dronkers. A new brain region for coordinating speech articulation. *Nature*, 384(6605):159–161, November 1996.

Nina F. Dronkers, Maria V. Ivanova, and Juliana V. Baldo. What Do Language Disorders Reveal about Brain–Language Relationships? From Classic Models to Network Approaches. *Journal of the International Neuropsychological Society : JINS*, 23(9-10):741–754, October 2017. ISSN 1355-6177. doi: 10.1017/S1355617717001126. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6606454/.

Erik Edwards, Srikantan S Nagarajan, Sarang S Dalal, Ryan T Canolty, Heidi E Kirsch, Nicholas M Barbaro, and Robert T Knight. Spatiotemporal imaging of cortical activation during verb generation and picture naming. *Neuroimage*, 50(1):291–301, March 2010.

Jeffrey Elman. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225, 1991. URL https://link.springer.com/article/10.1007/BF00114844.

David Embick. Features, syntax, and categories in the latin perfect. *Linguistic Inquiry*, 31(2):185–230, 2000. ISSN 00243892, 15309150. URL http://www.jstor.org/stable/4179104.

Daniel Everett. Cultural constraints on grammar and cognition in piraha: Another look at the design features of human language. *Language*, 46:621–646, 08 2005. doi: 10.1086/431525.

Evelina Fedorenko, Idan Blank, Matthew Siegelman, and Zachary Mineroff. Lack of selectivity for syntax relative to word meanings throughout the language network. *bioRxiv*, page 477851, 2020. URL https://www.sciencedirect.com/science/article/pii/S0010027720301670. Publisher: Cold Spring Harbor Laboratory.

Evelyn C. Ferstl and D. Yves von Cramon. The role of coherence and cohesion in text comprehension: an event-related fMRI study. *Cognitive Brain Research*, 11(3):325–340, June 2001. ISSN 0926-6410. doi: 10.1016/S0926-6410(01)00007-6. URL http://www.sciencedirect.com/science/article/pii/S0926641001000076.

Jerry Fodor. *The modularity of mind*. MIT press, 1983.

Angela D Friederici. The Brain Basis of Language Processing: From Structure to Function. *Physiol Rev*, 91:36, 2011.

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

Angela D. Friederici, Shirley-Ann RÃ$\frac{1}{4}$schemeyer, 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, 02 2003. ISSN 1047-3211. doi: 10.1093/cercor/13.2.170. URL https://doi.org/10.1093/cercor/13.2.170.

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, 01 2006. ISSN 1047-3211. doi: 10.1093/cercor/bhj106. URL https://doi.org/10.1093/cercor/bhj106.

Angela D. Friederici, Sonja A. Kotz, Sophie K. Scott, and Jonas Obleser. Disentangling syntax and intelligibility in auditory language comprehension. *Human Brain Mapping*, 31(3):448–457, August 2009a. ISSN 1065-9471. doi: 10.1002/hbm.20878. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6870868/.

Angela D. Friederici, Michiru Makuuchi, and JÃ¶rg Bahlmann. The role of the posterior superior temporal cortex in sentence comprehension. *NeuroReport*, 20(6):563–568, April 2009b. ISSN 0959-4965. doi: 10.1097/WNR.0b013e3283297dee. URL https://journals.lww.com/neuroreport/Fulltext/2009/04220/The_role_of_the_posterior_superior_temporal_cortex.6.aspx.

Angela D Friederici, Noam Chomsky, Robert C Berwick, Andrea Moro, and Johan J Bolhuis. Language, mind and brain. *Nature human behaviour*, 1(10):713–722, 2017.

Angela Dorkas Friederici. Neurobiology of Syntax as the Core of Human Language. *BIOLINGUISTICS*, 11, 2017. URL https://bioling.psychopen.eu/index.php/bioling/article/view/9093.

K J Friston, P Fletcher, O Josephs, A Holmes, M D Rugg, and R Turner. Event-related fMRI: characterizing differential responses. *Neuroimage*, 7(1):30–40, January 1998.

P. Garrard, E. Carroll, D. Vinson, and G. Vigliocco. Dissociation of lexical syntax and semantics: Evidence from focal cortical degeneration. *Neurocase*, 10(5):353–362, 2004. doi: 10.1080/13554790490892248. URL https://doi.org/10.1080/13554790490892248. PMID: 15788273.

Anne-Lise Giraud and David Poeppel. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, 15(4):511–517, Apr 2012. ISSN 1546-1726. doi: 10.1038/nn.3063. URL https://doi.org/10.1038/nn.3063.

Jozien B M Goense and Nikos K Logothetis. Neurophysiology of the BOLD fMRI signal in awake monkeys. *Curr. Biol.*, 18(9):631–640, may 2008.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Se Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines. *bioRxiv*, 2021. doi: 10.1101/2020.12.02.403477. URL https://www.biorxiv.org/content/early/2021/09/30/2020.12.02.403477.

Harold Goodglass. *Understanding aphasia.* Academic Press, 1993.

Elizabeth Bates Judith C Goodman. On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. *Lang. Cogn. Process.*, 12(5-6):507–584, October 1997.

Yosef Grodzinsky and Angela D Friederici. Neuroimaging of syntax and syntactic processing. *Curr. Opin. Neurobiol.*, 16(2):240–246, April 2006.

Yosef Grodzinsky and Andrea Santi. The battle for broca's region. *Trends in Cognitive Sciences*, 12(12):474–480, 2008. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2008.09.001. URL https://www.sciencedirect.com/science/article/pii/S1364661308002222.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1108. URL https://aclanthology.org/N18-1108.

Peter Hagoort. Nodes and networks in the neural architecture for language: Broca's region and beyond. *Current opinion in Neurobiology*, 28:136–141, 2014. URL https://pubmed.ncbi.nlm.nih.gov/25062474/. Publisher: Elsevier.

John Hale, Adhiguna Kuncoro, Keith Hall, Chris Dyer, and Jonathan Brennan. Text genre and training data size in human-like parsing. 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 5846–5852, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1594. URL https://aclanthology.org/D19-1594.

E Halgren, R D Walter, D G Cherlow, and P H Crandall. Mental phenomena evoked by electrical stimulation of the human hippocampal formation and amygdala. *Brain*, 101(1):83–117, March 1978.

Daniel A Handwerker, John M Ollinger, and Mark D'Esposito. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*, 21(4):1639–1651, apr 2004.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Ryuichiro Hashimoto and Kuniyoshi L Sakai. Specialization in the left prefrontal cortex for sentence comprehension. *Neuron*, 35(3):589–597, 2002. ISSN 0896-6273. doi: https://doi.org/10.1016/S0896-6273(02)00788-2. URL https://www.sciencedirect.com/science/article/pii/S0896627302007882.

Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664):1634–1640, mar 2004.

Olaf Hauk, Ingrid Johnsrude, and Friedemann Pulvermüller. Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, 41(2):301–307, 2004. URL http://www.sciencedirect.com/science/article/pii/S0896627303008389.

Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. The faculty of language: what is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579, November 2002.

John Hewitt and Christopher D. Manning. A Structural Probe for Finding Syntax in Word Representations. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, page 10.

G Hickok and D Poeppel. Towards a functional neuroanatomy of speech perception. *Trends Cogn. Sci.*, 4(4):131–138, April 2000.

Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nat. Rev. Neurosci.*, 8(5):393–402, May 2007.

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

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

John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55. URL https://ieeexplore.ieee.org/document/4160265.

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, April 2016. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature17637. URL http://www.nature.com/articles/nature17637.

Ray Jackendoff. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press UK, 2002. URL https://academic.oup.com/book/32834.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/f471223d1a1614b58a7dc45c9d01df19-Paper.pdf.

Mark Jenkinson, Christian F. Beckmann, Timothy E. J. Behrens, Mark W. Woolrich, and Stephen M. Smith. FSL. *NeuroImage*, 62(2):782–790, August 2012. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2011.09.015. URL http://www.sciencedirect.com/science/article/pii/S1053811911010603.

Mark Jung-Beeman. Bilateral brain processes for comprehending natural language. *Trends in Cognitive Sciences*, 9(11):512–518, November 2005. ISSN 13646613. doi: 10.1016/j.tics.2005.09.009. URL http://linkinghub.elsevier.com/retrieve/pii/S1364661305002718.

Marcel Adam Just and Patricia A Carpenter. *Psychology of reading and language comprehension*. Allyn & Bacon, Old Tappan, NJ, October 1986.

Ryuta Kinno, Mitsuru Kawamura, Seiji Shioda, and Kuniyoshi L. Sakai. Neural correlates of noncanonical syntactic processing revealed by a pictured sentence matching task. *Human Brain Mapping*, 29(9):1015–1027, October 2007. ISSN 1065-9471. doi: 10.1002/hbm.20441. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6871174/.

Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1249. URL https://www.aclweb.org/anthology/P18-1249.

S Knecht. Handedness and hemispheric language dominance in healthy humans. *Brain*, 123(12):2512–2518, December 2000.

Nikolaus Kriegeskorte and Pamela K Douglas. Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, 55:167–179, 2019. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2019.04.002. URL https://www.sciencedirect.com/science/article/pii/S0959438818301004. Machine Learning, Big Data, and Neuroscience.

Judith F Kroll and Ellen Bialystok. Understanding the consequences of bilingualism for language processing and cognition. *J. Cogn. Psychol. (Hove)*, 25(5):497–514, 2013.

Prantik Kundu, Souheil J. Inati, Jennifer W. Evans, Wen-Ming Luh, and Peter A. Bandettini. Differentiating BOLD and non-BOLD signals in fMRI time series using multi-echo EPI. *NeuroImage*, 60(3):1759–1770, April 2012. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2011.12.028.

Prantik Kundu, Valerie Voon, Priti Balchandani, Michael V. Lombardo, Benedikt A. Poser, and Peter A. Bandettini. Multi-echo fMRI: A review of applications in fMRI denoising and analysis of BOLD signals. *NeuroImage*, 154(2):59–80, July 2017. ISSN 10538119. doi: 10.1016/j.neuroimage.2017.03.033. URL http://linkinghub.elsevier.com/retrieve/pii/S1053811917302410.

Marta Kutas and Kara D Federmeier. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.*, 62(1):621–647, January 2011.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks, 2018.

Brenden M. Lake, Tal Linzen, and Marco Baroni. Human few-shot learning of compositional instructions, 2019.

Yair Lakretz, Germán Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. The emergence of number and syntax units in lstm language models. In *NAACL-HLT (1)*, 2019.

Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699, 2021.

Matthew A Lambon Ralph, Karen Sage, Roy W Jones, and Emily J Mayberry. Coherent concepts are computed in the anterior temporal lobes. *Proc. Natl. Acad. Sci. U. S. A.*, 107(6):2717–2722, February 2010.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. A natural language fmri dataset for voxelwise encoding models. 2022. doi: 10.1101/2022.09.22.509104. URL https://www.biorxiv.org/content/10.1101/2022.09.22.509104v1.

Yulia Lerner, Christopher J. Honey, Lauren J. Silbert, and Uri Hasson. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915, 2011. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.3684-10.2011. URL https://www.jneurosci.org/content/31/8/2906.

W. J Levelt. *Speaking: from intention to articulation*. The MIT Press, MIT, oct 1989. doi: https://doi.org/10.7551/mitpress/6393.001.0001.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, Nathan Spreng, Jonathan R. Brennan, Yiming Yang, Christophe Pallier, and John Hale. Le petit prince multilingual naturalistic fmri corpus. *Scientific Data*, 9, 2022. URL https://doi.org/10.1038/s41597-022-01625-7.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1eYHoC5FX.

Nikos K Logothetis. The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 357(1424):1003–1037, aug 2002.

Nikos K. Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, Jun 2008. ISSN 1476-4687. doi: 10.1038/nature06976. URL https://doi.org/10.1038/nature06976.

Nikos K. Logothetis and Brian A. Wandell. Interpreting the bold signal. *Annual Review of Physiology*, 66(1):735–769, 2004. doi: 10.1146/annurev.physiol.66.082602.092845. URL https://doi.org/10.1146/annurev.physiol.66.082602.092845. PMID: 14977420.

Nikos K. Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fmri signal. *Nature*, 412(6843):150–157, Jul 2001. ISSN 1476-4687. doi: 10.1038/35084005. URL https://doi.org/10.1038/35084005.

João Loula, Marco Baroni, and Brenden M Lake. Rearranging the familiar: Testing compositional generalization in recurrent networks. July 2018.

Renqian Luo, Fei Tian, Tao Qin, Enhong Chen, and Tie-Yan Liu. Neural architecture optimization, 2019.

Saima Malik-Moraleda, Dima Ayyash, Jeanne Gallée, Josef Affourtit, Malte Hoffmann, Zachary Mineroff, Olessia Jouravlev, and Evelina Fedorenko. An investigation across 45 languages and 12 language families reveals a universal language network. *Nat. Neurosci.*, 25(8):1014–1019, August 2022.

Raymond A. Mar. The neural bases of social cognition and story comprehension. *Annual review of psychology*, 62:103–134, 2011. URL https://pubmed.ncbi.nlm.nih.gov/21126178/.

David Marr. *Vision*. W. H. Freeman, 1982.

William Matchin and Gregory Hickok. The cortical organization of syntax. *Cerebral Cortex*, 30(3):1481–1498, Mar 2020. ISSN 1047-3211. doi: 10.1093/cercor/bhz180. URL https://doi.org/10.1093/cercor/bhz180.

William Matchin, Christopher Hammerly, and Ellen Lau. The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI. *cortex*, 88:106–123, 2017. Publisher: Elsevier.

B. M. Mazoyer, N. Tzourio, V. Frak, A. Syrota, N. Murayama, O. Levrier, G. Salamon, S. Dehaene, L. Cohen, and J. Mehler. The Cortical Representation of Speech. *Journal of Cognitive Neuroscience*, 5(4):467–479, October 1993. ISSN 0898-929X. doi: 10.1162/jocn.1993.5.4.467. URL https://doi.org/10.1162/jocn.1993.5.4.467. _eprint: https://direct.mit.edu/jocn/article-pdf/5/4/467/1932303/jocn.1993.5.4.467.pdf.

Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010. URL https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf.

Andrea Mechelli, Jenny T Crinion, Uta Noppeney, John O'Doherty, John Ashburner, Richard S Frackowiak, and Cathy J Price. Neurolinguistics: structural plasticity in the bilingual brain. *Nature*, 431(7010):757, October 2004.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. Toward a realistic model of speech processing in the brain with self-supervised learning, 2022. URL https://arxiv.org/abs/2206.01685.

Tom Mitchell, Svetlana Shinkareva, Andrew Carlson, Kai-Min Chang, Vincente Malave, Robert Mason, and Marcel Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320:1191–1195, 2008. URL https://www.science.org/doi/full/10.1126/science.1152876.

Francis Mollica, Matthew Siegelman, Evgeniia Diachek, Steven T. Piantadosi, Zachary Mineroff, Richard Futrell, and Evelina Fedorenko. High local mutual information drives the response in the human language network. *bioRxiv*, page 436204, 2018. URL https://www.biorxiv.org/content/10.1101/436204v1.full.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, May 2011. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.07.073. URL http://linkinghub.elsevier.com/retrieve/pii/S1053811910010657.

Samuel A. Nastase, Ariel Goldstein, and Uri Hasson. Keep it real: rethinking the primacy of experimental control in cognitive neuroscience,. *NeuroImage*, 222, 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.117254. URL https://doi.org/10.1016/j.neuroimage.2020.117254. Publisher: NeuroImage.

Samuel A. Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J. Honey, Yaara Yeshurun, Mor Regev, and et al. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, 8(1), 2021. doi: 10.1038/s41597-021-01033-3.

Sharlene D. Newman, Toshikazu Ikuta, and Thomas Burns. The effect of semantic relatedness on syntactic analysis: an fMRI study. *Brain and language*, 113(2):51–58, May 2010. ISSN 0093-934X. doi: 10.1016/j.bandl.2010.02.001. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2854177/.

S Ogawa, T M Lee, A R Kay, and D W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci. U. S. A.*, 87(24):9868–9872, dec 1990.

S Ogawa, D W Tank, R Menon, J M Ellermann, S G Kim, H Merkle, and K Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proc. Natl. Acad. Sci. U. S. A.*, 89(13):5951–5955, jul 1992.

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.

Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527, 2011. URL https://www.pnas.org/doi/10.1073/pnas.1018711108. Publisher: National Acad Sciences.

A Pascual-Leone, D Bartres-Faz, and J P Keenan. Transcranial magnetic stimulation: studying the brain-behaviour relationship by induction of 'virtual lesions'. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 354(1387):1229–1238, July 1999.

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 *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162, pages 17499–17516, 2022. URL https://arxiv.org/abs/2207.03380.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL http://aclweb.org/anthology/D14-1162.

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):963, March 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03068-4. URL https://www.nature.com/articles/s41467-018-03068-4. Number: 1 Publisher: Nature Publishing Group.

S E Petersen, P T Fox, M I Posner, M Mintun, and M E Raichle. Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature*, 331(6157):585–589, February 1988.

Ana Luisa Grilo Pinho, Lucie Hertz-Pannier, and Bertrand Thirion. "ibc", 2021.

David Poeppel, Karen Emmorey, Gregory Hickok, and Liina Pylkkänen. Towards a new neurobiology of language. *J. Neurosci.*, 32(41):14125–14131, oct 2012.

Russell A. Poldrack, Jeanette A. Mumford, and Thomas E. Nichols. *Handbook of Functional MRI Data Analysis*. Cambridge University Press, 2011. doi: 10.1017/CBO9780511895029. URL https://www.cs.mtsu.edu/~xyang/fMRIHandBook.pdf.

Chantel S Prat, Timothy A Keller, and Marcel Adam Just. Individual differences in sentence comprehension: A functional magnetic resonance imaging investigation of syntactic and lexical processing demands. *J. Cogn. Neurosci.*, 19(12):1950–1963, December 2007.

Cathy J Price. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage*, 62(2):816–847, August 2012.

Friedemann Pulvermüller. Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits. *Brain and Language*, 127(1):86–103, October 2013. ISSN 1090-2155. doi: 10.1016/j.bandl.2013.05.015.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. page 24, 2019.

Marcus E. Raichle. The brain's default mode network. *Annual review of neuroscience*, 38:433–447, 2015. URL https://pubmed.ncbi.nlm.nih.gov/25938726/. Publisher: Annual Reviews.

Marus Raichle. A brief history of human functional brain mapping. 12 2000. doi: 10.1016/B978-012692545-6/50004-0.

Matthew A Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T Rogers. The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.*, 18(1):42–55, January 2017.

M. Regev, C. J. Honey, E. Simony, and U. Hasson. Selective and Invariant Neural Responses to Spoken and Written Narratives. *Journal of Neuroscience*, 33(40):15978–15988, October 2013. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1580-13.2013. URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1580-13.2013.

Hugo Richard, Lucas Martin, Ana Lusa Pinho, Jonathan Pillow, and Bertrand Thirion. Fast shared response model for fMRI data. *arXiv:1909.12537 [cs, eess, q-bio]*, December 2019. URL http://arxiv.org/abs/1909.12537. arXiv: 1909.12537.

Elliott D Ross. Dominant language functions of the right hemisphere? *Arch. Neurol.*, 36(3):144, March 1979.

Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M. Lake. A benchmark for systematic generalization in grounded language understanding, 2020.

D E Rumelhart, G E Hinton, and R J Williams. Learning internal representations by error propagation. In *Readings in Cognitive Science*, pages 399–421. Elsevier, 1988.

Jake Russin, Jason Jo, Randall C. O'Reilly, and Yoshua Bengio. Compositional generalization in a deep seq2seq model by separating syntax and semantics, 2019. URL https://arxiv.org/abs/1904.09708.

Kuniyoshi L Sakai, Yasuki Noguchi, Tatsuya Takeuchi, and Eiju Watanabe. Selective priming of syntactic processing by event-related transcranial magnetic stimulation of broca's area. *Neuron*, 35(6):1177–1182, September 2002.

Andrea Santi and Yosef Grodzinsky. fmri adaptation dissociates syntactic complexity dimensions. *NeuroImage*, 51(4):1285–1293, 2010. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2010.03.034. URL https://www.sciencedirect.com/science/article/pii/S1053811910003216.

A. Schaefer, R. Kong, E.M. Gordon, T.O. Laumann, X.N. Zuo, A.J. Holmes, S.B. Eickhoff, and B.T.T. Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral Cortex*, 29:3095–3114, 2018. URL http://people.csail.mit.edu/ythomas/publications/2018LocalGlobal-CerebCor.pdf.

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. Artificial Neural Networks Accurately Predict Language Processing in the Brain. Technical report, June 2020. URL https://www.biorxiv.org/content/10.1101/2020.06.26.174482v1.

Einat Shetreet and Naama Friedmann. The processing of different syntactic structures: fmri investigation of the linguistic distinction between wh-movement and verb movement. *Journal of Neurolinguistics*, 27(1):1–17, 2014.

Matthew Siegelman, Idan A Blank, Zachary Mineroff, and Evelina Fedorenko. An attempt to conceptually replicate the dissociation between syntax and semantics during sentence comprehension. *Neuroscience*, 413:219–229, 2019. URL https://www.sciencedirect.com/science/article/pii/S0306452219304026. Publisher: Elsevier.

Erez Simony, Christopher J. Honey, Janice Chen, Olga Lositsky, Yaara Yeshurun, Ami Wiesel, and Uri Hasson. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, 7(1):12141, July 2016. ISSN 2041-1723. doi: 10.1038/ncomms12141. URL http://www.nature.com/articles/ncomms12141. Number: 1 Publisher: Nature Publishing Group.

David R. So, Chen Liang, and Quoc V. Le. The evolved transformer, 2019.

Roger Wolcott Sperry. Cerebral Organization and Behavior: The split brain behaves in many respects like two separate brains, providing new research possibilities. *Science*, 133(3466):1749–1757, 1961. URL https://pubmed.ncbi.nlm.nih.gov/17829720/. Publisher: American Association for the Advancement of Science.

Karin Stromswold, David Caplan, Nathaniel Alpert, and Scott Rauch. Localization of syntactic comprehension by positron emission tomography. *Brain and Language*, 52(3):452–473, 1996. ISSN 0093-934X. doi: https://doi.org/10.1006/brln.1996.0024. URL https://www.sciencedirect.com/science/article/pii/S0093934X96900243.

B E Swartz and E S Goldensohn. Timeline of the history of EEG and associated fields. *Electroencephalogr. Clin. Neurophysiol.*, 106(2):173–176, February 1998.

Kirsten Thomsen, Nikolas Offenhauser, and Martin Lauritzen. Principal neuron spiking: neither necessary nor sufficient for cerebral blood flow in rat cerebellum. *J. Physiol.*, 560(Pt 1):181–189, oct 2004.

Alexis Thual, Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced gromov-wasserstein, 2022.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *arXiv preprint arXiv:1905.11833*, 2019. URL https://arxiv.org/abs/1905.11833.

Mariya Toneva, Otilia Stretcu, Barnabas Poczos, Leila Wehbe, and Tom M Mitchell. Modeling task effects on meaning representation in the brain via zero-shot meg prediction. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5284–5295. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/38a8e18d75e95ca619af8df0da1417f2-Paper.pdf.

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals new aspects of meaning composition. *BioRxiv*, pages 2020–09, 2022.

Michael T. Ullman. Contributions of memory circuits to language: the declarative/procedural model. *Cognition*, 92(1):231–270, May 2004. ISSN 0010-0277. URL https://www.sciencedirect.com/science/article/pii/S0010027703002324.

Aditya R. Vaidya, Shailee Jain, and Alexander G. Huth. Self-supervised models of audio effectively explain human cortical responses to speech, 2022. URL https://arxiv.org/abs/2205.14252.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December 2017. URL http://arxiv.org/abs/1706.03762. arXiv: 1706.03762.

A L Vazquez and D C Noll. Nonlinear aspects of the BOLD response in functional MRI. *Neuroimage*, 7(2):108–118, feb 1998.

Gabriella Vigliocco. Language processing: The anatomy of meaning and syntax. *Current Biology*, 10(2): R78–R80, 2000. ISSN 0960-9822. doi: https://doi.org/10.1016/S0960-9822(00)00282-7. URL https://www.sciencedirect.com/science/article/pii/S0960982200002827.

J P Vignal, P Chauvel, and E Halgren. Localised face processing by the human prefrontal cortex: stimulation-evoked hallucinations of faces. *Cogn. Neuropsychol.*, 17(1):281–291, February 2000.

M Vigneau, V Beaucousin, P Y Hervé, H Duffau, F Crivello, O Houdé, B Mazoyer, and N Tzourio-Mazoyer. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *Neuroimage*, 30 (4):1414–1432, May 2006.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T. Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, Yoshiki Vázquez-Baeza, and SciPy 1.0 Contributors. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, Mar 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0686-2. URL https://doi.org/10.1038/s41592-019-0686-2.

Tor D Wager, Alberto Vazquez, Luis Hernandez, and Douglas C Noll. Accounting for nonlinear BOLD effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies. *Neuroimage*, 25(1): 206–218, jan 2005.

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), 2014a. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0112575. Publisher: Public Library of Science.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, Doha, Qatar, October 2014b. Association for Computational Linguistics. doi: 10.3115/v1/D14-1030. URL https://aclanthology.org/D14-1030.

Janet F Werker and Richard C Tees. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behav. Dev.*, 7(1):49–63, January 1984.

C Wernicke. *Der aphasische Symptomenkomplex: Eine psychologische Studie auf anatomischer Basis*. Cohn & Weigert, 1874.

Jiang Xu, Stefan Kemeny, Grace Park, Carol Frattali, and Allen Braun. Language in context: emergent features of word, sentence, and narrative comprehension. *NeuroImage*, 25(3):1002–1015, April 2005. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2004.12.013. URL http://www.sciencedirect.com/science/article/pii/S1053811904007748.

Emiliano Zaccarella, Lars Meyer, Michiru Makuuchi, and Angela D. Friederici. Building by Syntax: The Neural Basis of Minimal Linguistic Structures. *Cerebral Cortex*, 27(1):411–421, 10 2015. ISSN 1047-3211. doi: 10.1093/cercor/bhv234. URL https://doi.org/10.1093/cercor/bhv234.