université
PARIS-SACLAY

Language representations in deep learning algorithms and the brain

Représentations de langage dans les algorithmes d'apprentissage profonds et le cerveau

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580, Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat : Informatique
Graduate School : Informatique et sciences du numérique. Référent : ENS Paris-Saclay

Thèse préparée à l'**Inria Saclay (MIND)** et **Meta AI (FAIR)**, sous la direction de **Alexandre GRAMFORT**, Directeur de recherche à l'Inria Saclay et la co-supervision de **Jean-Rémi KING**, Research Scientist à Meta AI & Ecole Normale Superieure.

Thèse soutenue à Paris, le 10 mai 2023, par

Charlotte CAUCHETEUX

Composition du jury

Membres du jury avec voix délibérative

Stanislas DEHAENE Professeur, INSERM-CEA & Collège de France	Président
Evelina FEDORENKO Assistant professor, Massachusetts Institute of Technology (MIT)	Rapporteur & Examinatrice
Alexander HUTH Assistant professor, University of Texas at Austin	Rapporteur & Examinateur
Christophe PALLIER Directeur de recherche, INSERM-CEA	Examinateur
Yann LECUN Chief AI Scientist, Meta AI & Silver Professor, NYU	Examinateur

THESE DE DOCTORAT

NNT : 2023UPASG031

université PARIS-SACLAY

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

Title : Language representations in deep learning algorithms and the brain
Keywords : Natural Language Processing, Neuroscience, Deep Learning, fMRI

Abstract : Recent advances in artificial intelligence have led to the emergence of deep language models – like GPT-3 and ChatGPT – able to produce text that closely resembles that of humans. Such similarity raises questions about how the brain and deep models process language, the mechanisms they use, and the internal representations they construct. In this thesis, I compare the internal representations of the brain and deep language models, with the goal of identifying their similarities and differences.

To this aim, I analyze functional resonance imaging (fMRI) and magnetoencephalography (MEG) recordings of participants listening to and reading sentences, and compare them to the activations of thousands of language algorithms corresponding to these same sentences.

Our results first highlight high-level similarities between the internal representations of the brain and deep language models. We find that deep nets' activations significantly predict brain activity across subjects for different cohorts (>500 participants), recording modalities (MEG and fMRI), stimulus types (isolated words, sentences, and natural stories), stimulus modalities (auditory and visual presentation), languages (Dutch, English and French), and deep language models. This alignment is maximal in brain regions repeatedly associated with language, for the best-performing algorithms and for participants who best understand the stories. Critically, we evidence a similar processing hierarchy between the two systems. The first layers of the algorithms align with low-level processing regions in the brain, such as auditory areas and the temporal lobe, while the deep layers align with regions associated with higher-level processing, such fronto-parietal areas.

We then show how such similarities can be leveraged to build better predictive models of brain activity and better decompose several linguistic processes in the brain, such as syntax and semantics.

Finally, we explore the differences between deep language models and the brain's activations. We find that the brain predicts distant and hierarchical representations, unlike current language models that are mostly trained to make short-term and word-level predictions.

Overall, modern algorithms are still far from processing language in the same way that humans do. However, the linear correspondence between their inner workings and that of the brain provide an promising platform for better understanding both systems, and pave the way for building better algorithms inspired by the human brain.

université PARIS-SACLAY

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

Titre : Représentations de langage dans les algorithmes d'apprentissage profonds et le cerveau
Mots clés : Traitement Automatique du Langage Naturel, Neurosciences, Apprentissage Profond, IRMf

Résumé : Algorithmes et cerveau, bien que de nature extrêmement différentes, sont deux systèmes capables d'effectuer des tâches de langage complexes. En particulier, de récentes avancées en intelligence artificielle ont permis l'émergence d'algorithmes produisant des textes de qualité remarquablement similaire à ceux des humains (ChatGPT, GPT-3). De telles similarités interrogent sur la *façon* dont le cerveau et ces algorithmes traitent le langage, les *mécanismes* qu'ils utilisent et les *représentations internes* qu'ils construisent. Ma thèse consiste à comparer les représentations internes de ces deux systèmes, d'identifier leurs similitudes et leurs différences.

Pour ce faire, nous analysons les enregistrements par imagerie fonctionnelle (fMRI) et magnéto-encéphalographie (MEG) de participants écoutant et lisant des histoires, et les comparons aux activations de milliers d'algorithmes de langage correspondant à ces mêmes histoires.

Nos résultats mettent d'abord en évidence des similarités de haut niveau entre les représentations internes du cerveau et des modèles de langage. Dans une première partie, nous montrons que les activations des réseaux profonds prédisent linéairement l'activité cérébrale de sujets chez différents groupes ($>$ 500 participants), pour différentes modalités d'enregistrement (MEG et fMRI), modalités de stimulus (présentation auditive et visuelle), types de stimulus (mots isolés, phrases et histoires naturelles), langues (néerlandais et anglais) et modèles de langage. Cette correspondance est maximale dans les régions cérébrales souvent associées au langage, pour les algorithmes les plus performants et pour les participants qui comprennent le mieux les histoires. De plus, nous mettons en évidence une hiérarchie de traitement similaire entre les deux systèmes. Les premières couches des algorithmes sont alignées sur les régions de traitement de bas niveau dans le cerveau, telles que les zones auditives et le lobe temporal, tandis que les couches profondes sont alignées sur des régions associées à un traitement de plus haut niveau, notamment les zones fronto-pariétales.

Nous montrons ensuite, dans une seconde partie, comment de telles similarités peuvent aider à construire de meilleurs modèles prédictifs de l'activité cérébrale, et à décomposer plus finement dans le cerveau différents processus linguistiques tels que la syntaxe et la sémantique.

Enfin, dans une troisième partie, nous explorons les différences entre cerveau et algorithmes. Nous montrons que le cerveau prédit des représentations distantes et hiérarchiques, contrairement aux modèles de langage actuels qui sont principalement entraînés à faire des prédictions à court terme et au niveau du mot.

Dans l'ensemble, les algorithmes modernes sont encore loin de traiter le langage de la même manière que les humains le font. Cependant, les liens directs entre leur fonctionnement interne et celui du cerveau fournissent une plateforme prometteuse pour mieux comprendre les deux systèmes, et ouvre la voie à la construction de meilleurs algorithmes inspirés du cerveau.

Acknowledgements

J'aimerais remercier Jean-Rémi King et Alexandre Gramfort, sans qui, cela va de soi, rien de tout cela n'aurait été possible. J'ai énormément appris à vos côtés, humainement, scientifiquement et techniquement, et je vous en suis extrêmement reconnaissante. Je vous remercie pour votre bienveillance, votre écoute, votre enthousiasme et humilité. Je garderai avec moi précieusement ces enseignements dans la suite de mon parcours, et aurai, je l'espère, l'occasion de les partager à nouveau. Merci à Jean-Rémi de m'avoir fait confiance et de m'avoir initiée au domaine des neurosciences avec passion et pédagogie. Merci d'avoir été présent, même et surtout dans les moments difficiles, pour tes conseils techniques et non-techniques toujours constructifs, pragmatiques et pour surmonter les problèmes quand le doute survient, quelles que soient la noblesse ou la difficulté du problème rencontré. Merci à Alex pour la parfaite distance et le recul qui ont permis de guider cette thèse, pour tes conseils scientifiques ou de vie toujours pertinents, et pour la confiance que tu m'as accordée. Je tiens également à vous remercier pour votre disponibilité, même en période de deadlines, et pour votre humour en toute circonstance. Equipe de choc qui me manquera beaucoup. Je pourrais continuer longtemps mais aucun discours ne sera suffisant pour exprimer l'admiration et l'affection que j'ai pour mes deux directeurs de thèse. Je m'arrête donc ici et vous remercie profondément de m'avoir donné confiance, et pour tout ce que vous m'avez apporté. Un des éléments les plus marquants de ces trois années aura certainement été mon travail avec vous.

I would like to thank the members of my dissertation committee: Evelyna Fedorenko, Alexander Huth, Stanislas Dehaene, Christophe Pallier and Yann Lecun, for their willingness to serve as jury members and evaluate my work. I am honored to have the opportunity to present and discuss my research in their presence.

I am deeply grateful to Alexander and Evelyna for their invaluable contributions to the rapidly evolving field and for reviewing my manuscript and providing insightful comments. Their groundbreaking research and pioneering work have inspired me and many others, and I am thankful for their guidance and support.

Je remercie également Stanislas Dehaene pour ses cours au Collège de France, pour ses directions scientifiques inspirantes, pour m'avoir sensibilisée aux sciences cognitives et donné envie de faire de la recherche dans ce domaine. Ses travaux sur l'étude des nombres, les langages du cerveau (parole, musique et mathématiques) et la singularité de l'espèce humaine

m'ont particulièrement marquée et ont été déterminants dans mon envie de faire de la recherche, dès mes années d'école d'ingénieur.

Je remercie Christophe Pallier pour son soutien tout au long de cette thèse, sa bienveillance constructive, ses a priori toujours positifs, ses interrogations et réflexions scientifiques et philosophiques inspirantes qui dépassent le cadre strict de son domaine de recherche.

Je remercie Yann Lecun pour avoir accepté de faire partie du jury, pour avoir été artisan des succès et de la démocratisation de l'intelligence artificielle. J'ai eu la chance d'être née au bon moment pour pouvoir y participer. Je remercie également Yann pour les ponts qu'il a établis entre la recherche publique et l'industrie, ponts dont je bénéficie au quotidien, et pour avoir initié, avec Antoine Bordes, la création du bureau de FAIR à Paris.

I would like to thank Marco Baroni for introducing me to Jean-Remi and for his advice throughout this thesis, for instance, on discrete representations, as well as the organization of the evil meeting.

Je remercie Emmanuel Dupoux pour m'avoir accueillie lors de mon premier stage de recherche au CoML, et pour les discussions stimulantes que nous avons eues sur l'apprentissage du langage chez les enfants.

I would like to thank all those who have contributed to the creation and accessibility of the open databases that I have used, namely Narratives, MOUS, and Le Petit Prince. I am particularly grateful to Sam Nastase for guiding me in the exploration of the Narratives dataset.

I would like to express my gratitude to the entire open-source community, without whom my work would not have been possible, especially the scikit-learn, MNE, PyTorch, Nilearn and HuggingFace teams.

Je remercie mon laboratoire Inria et Bertrand Thirion pour m'avoir accueillie. Ma présence a été limitée mais j'ai apprécié tous les moments passés au laboratoire, aux conférences, aux off-sites et aux réunions du mardi après-midi. J'ai toujours été admirative de la culture d'humilité, de rigueur scientifique et technique, de défense de l'open source et de bienveillance qui habitait l'équipe.

Je remercie mon laboratoire industriel, Meta AI (anciennement Facebook AI Research), et Antoine Bordes pour m'avoir accueillie. Je suis consciente de la chance que j'ai eue de pouvoir mener à bien mes recherches avec une grande liberté, entourée de chercheurs de pointe en apprentissage profond, avec toutes les ressources nécessaires pour contribuer au domaine fascinant qu'est l'intelligence artificielle.

J'aimerais remercier plus spécifiquement mes fantastiques collègues de FAIR, l'Inria et l'ENS avec qui j'ai eu la chance de collaborer et échanger, Juliette M, Alex D., Lina, Virginie, Pierre O, Theo D, Omar C, Guillaume C, PA D, Jean T, Evrard, Leonard, Baptiste, Roberto, Jeremy R, Louis M, Guillaume L, Herve J, Armand J, Daniel, Francisco, Linnea, Corentin, Josephine, Alexis T, Hubert B, Hubert E, Marvin, Timothée D, Julia, Apolline, Benedicte, Gael V, Thomas M, Louis R, Cedric, Alexandre P, et tous les thésards et non thésards de FAIR, MIND et SODA.

Je remercie un certain nombre de personnes avec qui j'ai eu la chance de travailler avant ma thèse, Claire Lasserre, Jacob Leygonie, Thomas Clozel, Gilles Wainrib, Yann Hendel, Lukasz Bolikowski, Priscille Boissonnet, Benoit Audigier, Caroline Apra, Arthur Mensch, Patrick Jourdain et Rahma Chaabouni. Chacun role modèle a sa façon dont j'ai beaucoup appris.

Je souhaite également remercier un certain nombre de professeurs ayant eu un impact particulier lors de mes études. Je remercie mes professeurs de master Julie Josse et Erwan Le Pennec, pour leur pédagogie, leur goût pour l'application, et l'humilité avec laquelle ils ont transmis des concepts complexes à des étudiants n'ayant pas un cursus ingénieur. J'aimerais également remercier mes professeurs de philosophie, mathématiques et littérature Mme Manonellas, Mme Danon, Mr Cornilleau, Mme Boulay, et tous mes professeurs de classe préparatoire et de licence. Je remercie également Frédéric Laupies pour son ouvrage sur le thème de l'Espace, et les auteurs dont les textes sur ce thème ont accompagné mes réflexions en classe préparatoire, E. Kant and M. Merleau-Ponty. Je remercie particulièrement Pierre Cornilleau et Denis Pennequin pour avoir soutenu ma candidature à Polytechnique. Je remercie également tous mes professeurs de Polytechnique, école qui, en toute simplicité, a profondément changé ma vie. Plus récemment, je remercie les organisateurs et les intervenants du séminaire "The Representation of Language in Brains and Machines" au Collège de France, en particulier Stanislas Dehaene, Stéphane Mallat et Luigi Rizzi, qui ont apporté un éclairage inspirant à mon sujet de thèse.

Je remercie tous les musiciens et artistes qui rendent ma vie plus heureuse.

Je remercie mes ami.e.s et partenaires de course, Mathieu, Jean, Guillaume, Roberto, Arthur, Jean-Baptiste, Timothee, Louis, Tanguy, Marie, Priscille, avec qui j'ai eu la chance de partager des discussions scientifiques passionnantes, discussions non scientifiques tout aussi passionnantes, mais aussi des discussions pas passionnantes du tout mais très joyeuses.

Je remercie mes ami.e.s et partenaires de vie Lucie, Paola, Sophie, Leila, Priscille, Tanguy, Marie, Sebastian, Claudia, Antonin, Robin, Manon, Thomas, Alexis, Marie, Bruno, Loiseau, Isabelle, Romane, Savinien, Diane, Sonia, Louis, Charlotte, Elise, Claire, Audrey, Ophelie, team coeur et grand coeur, les veggies'up. Merci à tous d'être là. Love infini.

Un merci tout particulier à mes partenaires de confinement sur l'île à poule, Lucie, Thomas, Robin, Manon, qui ont rendu cette période difficile heureuse.

Merci à Gautier,

Enfin, merci à Lili, Paul, Louise, mes parents, Joséphine, Anna et à toute ma famille et belle famille.

Contents

List of Figures

Chapter 1

Introduction

1.1 What this thesis tries to address

"Language in algorithms,
A code of ones and zeros,
A digital tongue,
A way to communicate with the pros.

Language in the brain,
A network of neurons and synapses,
A biological tongue,
A way to communicate with the senses.

Two different worlds,
But both with the power to convey,
Thoughts, feelings, and ideas,
In their own unique way."

– by ChatGPT

ChatGPT, an algorithm created by OpenAI in November 2022, has the ability to write coherent, well-constructed and evocative text. It demonstrates compositional skill and uses

figures of style that are often associated with a certain level of abstraction. For example, it the poem above, ChatGPT makes an analogy between the "digital tongue" of the computer and the "biological tongue" of the brain. In general, the algorithm is able to translate, synthesize text, and write code in a way that is strikingly similar to that of a human.

Language in algorithms: a conceptual paradox

Algorithms have generated novel images and faces for several years, but witnessing language abilities in algorithms is particularly striking given the special nature of language. Language is a *powerful communication tool* that influences our thoughts, actions, that structures our cultures and societies. It is also a *hard scientific problem*: there is close to infinite ways of combining words into sentences, and meaning varies across cultures and contexts, presenting short-term and long-term dependencies (Chomsky, 1957; Bengio et al., 2001). Critically, while machines follow instructions, language has long been associated with *intelligence and thought* (Mahowald et al., 2023). Already in the 17^{th} century, René Descartes linked language with human thought. In a famous letter addressed to the Marquess of Newcastle[1], he argued that animals, because they lack the symbolic system of human language, cannot possibly think. Reciprocally, he argued that non-thinking machines won't be able to demonstrate language abilities, and thus proposed language as a test to distinguish machines from humans.

> *"If there was a machine shaped like our bodies which imitated our actions as much as is morally possible, we would always have two very certain ways of recognizing that they were not, for all their resemblance, true human beings. The first is that* ***they could never use words or other signs, composing them as we do in order to declare our thoughts to others****.*
>
> *For one can readily conceive that a machine might be made in such a way that it produces words, and even that it produces some words relevant to the corporeal actions that effect some change in its organs, e.g., that if one touches it in a certain place, it will ask what one wishes to say to it; and that if one touches it in another place, it will exclaim that one is hurting it, and the like. But* ***one cannot conceive that the machine could arrange words so diversely as to respond to the meaning of all that might be said in its presence, as even the most stupid human beings can do.*** *"*

[1] Letter to the Marquis de Newcastle, on November 23, 1646

– Discourse on Method, Part V. AT VI, 56
René Descartes
Gerald J. Massey's translation

More recently, and in line with René Descartes, Alan Turing proposed dialogue as a test of intelligence. He introduced the Turing test in 1950, in which evaluators engage in conversation with a machine, attempting to distinguish the machines' responses from that of a human. If the evaluators are unable to do so, the machine is said to have passed the test, and evidence a form of intelligence and reasoning (Turing, 1950). While controversial, the Turing test has shaped the way society conceive intelligence (Mahowald et al., 2023).

On the need to probe intermediate representations

Today, humans are no longer the only systems capable of arranging words into meaning (A. Wang et al., 2018, 2020). Algorithms are improving at the Turing test and challenge René Descartes' predictions. These advancements raise questions about the inner workings of the two systems; the mechanisms they use and the intermediate steps they follow. Some argue that algorithms lack understanding and simply rely on low-level statistical patterns (Marcus, 2020b). On the opposite spectrum, others have raised questions about the possibility of higher-level human computational processes such as consciousness existing in machines (Cerullo, 2022; Chalmers, 2023). Less radically, and without delving into the question of thoughts, one may wonder whether humans and algorithms share similar underlying mechanisms to process language.

To clarify the dissociation between the models' performance and its inner workings, we break down language processing systems into three levels of analysis[2] (Figure 1.1):

- the systems' *behaviour*: the sequences the system generates and its performance at language tasks (e.g. the accuracy at predicting a word from its context)
- the systems' *intermediate representations*: the intermediate steps to achieve such behaviour (e.g. building the syntactic tree of the sentence)

[2]We do not use Maar's taxonomy here (Marr & Poggio, 1976) because we investigate the properties underlying similar observable behaviours, representations and structures.

- the systems' *implementation*: their architectural and physical properties (e.g. the number of synapses, digital vs. biological units).

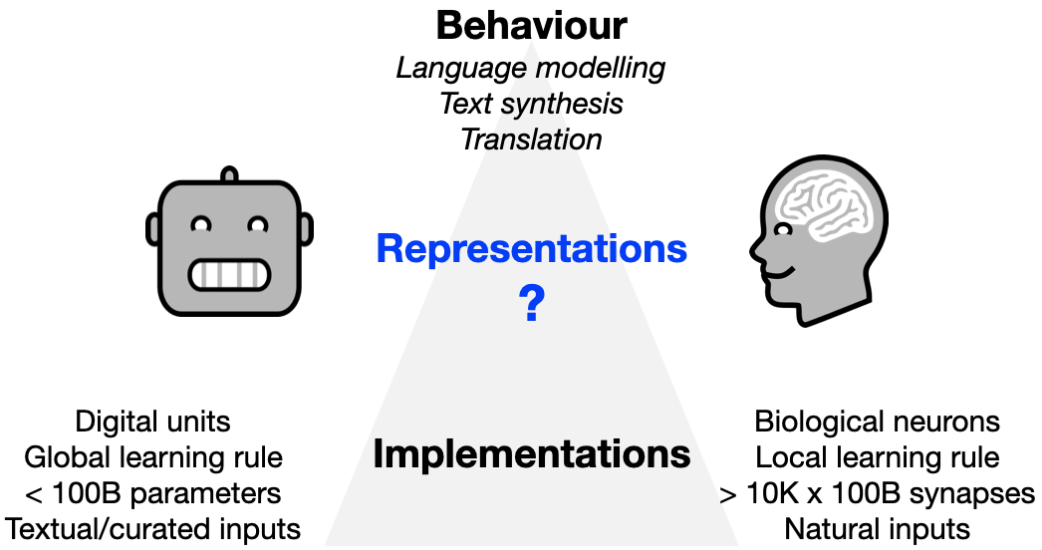


Figure 1.1: Problem statement. Artificial neural networks and the brain are two language processing systems able to generate, synthesize, translate text. While very different in their implementations, do they use similar intermediate representations to process language?

Multiple implementations can lead to the same behaviour. For instance, while artificial and biological neural networks are increasingly similar at the *behavioural* level, the are extremely different in their *implementation*. On the one hand, the brain is made of biological neurons, trained with a local learning rule, exposed to a variety of natural stimuli from senses that can be influenced through interaction with the world. On the other hand, the best performing language models are made of digital units trained with back-propagation, exposed to curated and textual inputs.

Similarly, multiple intermediate representations can lead to the same behaviour. Let's take the example of next-word prediction. Next-word prediction consists in predicting the next word, e.g. "time" given its previous context, e.g. "Once upon a [?]". Multiple systems can have the same *behaviour* and predict the word "time" while using different mechanisms. For instance, system A could search a database to look for the closest sentence in Wikipedia; system B could be hard-coded such that it always generates the word "time" after the word "upon a", system C could rely on shallow 4-grams statistics, while system D could build complex intermediate representations (e.g. constructing the syntactic structure of the sentence and the

expected semantic category). These four strategies can lead to the same answer "time", and hence different intermediate representations are built.

Question

Here, we do not focus on the systems' *behaviour* –the words they generate– nor on the systems' *implementation* –their architectural and physical properties– but on the *intermediate representations* they build to process language. Precisely, we address the following question:

Do artificial neural networks and the human brain build similar intermediate representations to process language?

Studying intermediate representations is challenging because both the human brain and deep algorithms are *black boxes* (Abnar et al., 2019). Artificial Neural Networks (ANNs) are made of billions of parameters, so it is difficult to *interpret* their explicit representations, and *disentangle* the properties they encode. On the other hand, computations in the brain can only be estimated through measurements and observations, and such high-dimensional and noisy measurements are hard to study. Despite these difficulties, decades of research have investigated language representations in algorithms and the brain independently, showing that deep neural networks build relevant semantic and syntactic spaces (Jawahar et al., 2019; Manning et al., 2020; Mahowald et al., 2023), and identifying in the brain the spatial and temporal dynamics underlying language processes (Hickok & Poeppel, 2007; Friederici, 2011; Pallier et al., 2011; Fedorenko et al., 2016). Here, we do not study language representations in artificial networks on the one side, and in the brain on the other side, but **directly quantify the similarity between the two**.

In the following sections, we give a non exhaustive overview of research that have investigated language representations in deep neural networks and the human brain independently, as well as recent approaches that directly quantifies the similarity bewteen deep networks' and brains' inner representations. Finally, we expose our approach and the contributions of the thesis.

1.2 Language representations in artificial neural networks

1.2.1 Language Modeling

Language can be formalized as a sequence of discrete random variables taking a finite set of values called a vocabulary. Modeling language is then learning the joint distribution of this sequence of random variables. A statistical model of language can be represented by the conditional probability of a word given its context (Bengio et al., 2001):

$$P(w_1^T) = \prod_{t=1}^{T} P(w_t | w_1^{t-1}) \ ,$$

where w_t is word t in the sequence, and w_i^j is the sequence of words $(w_i, \dots w_j)$ (Bengio et al., 2001). One can simplify the problem by leveraging the fact that temporally closer words in the word sequence are statistically more dependent, and reduce the context size to $k > 0$ words:

$$P(w_1^T) \approx \prod_{t=1}^{T} P(w_t | w_{t-k}^{t-1}) \ .$$

Learning P is then learning the distribution probability of the next word given its context. In practice, such probability is learnt using the cross-entropy loss, by maximizing the likelihood of the true word given its previous context:

$$\hat{\theta} = \text{argmin}_{\theta \in \Theta} \quad - \sum_{t=1}^{T} \log\big(P_\theta(w_t | w_{t-k}^{t-1})\big)$$

This is a hard optimization problem if we consider discrete random variables because of the curse of dimensionality. For example, if we use a vocabulary of $V = 10,000$ words for a sequence of $N = 10$ consecutive words, we get 10^{40} possible combinations. A solution to modeling this phenomenon is to project words into a continuous space and then learn the joint distribution in this new space. Two questions arise:

- Which continuous space should we choose? We will refer to vectors in this space as *lexical* representations, or word embeddings.
- How can we model the joint distribution in this new space? We will refer to vectors encoding interactions between multiple words as *contextual* representations.

In the followings, we precise different ways of computing lexical and contextual representations. Note that these two problems can be addressed simultaneously. For example, Bengio et al. (2001) proposed learning both the word embedding and the continuous probability distribution. This is also the case for most deep language models introduced subsequently (Devlin et al., 2019; Brown et al., 2020).

1.2.2 Lexical Representations

Architectures and training task

To build relevant lexical representations, Mikolov, Sutskever, et al. (2013) proposed the Word2Vec model. Each word is an index in a vocabulary. Each index corresponds to a vector representation which values are learned. The vector representation can be learnt in two ways (Figure 1.2):

- predicting a word from its context (continuous bag of words, or CBOW),
- predicting an adjacent word (skip-gram).

These tasks are classification tasks, with the number of classes being the number of words in the vocabulary. The model learns to represent each word as a unique vector, independently of its context.

Emergence of lexical semantics

Remarkably, this method allows for the emergence of a semantic representation space, meaning that two words close in the Word2Vec representation space are also semantically close (Figure 1.2B). This allows the model to capture relationships between words and to perform tasks such as analogy completion. Analogy is a way to evaluate "human-like" semantics in word embeddings. For instance, one commonly used benchmark is the Google Analogy Test Set, which consists of a list of analogy questions of the form "A is to B as C is to..." (Mikolov, Chen, et al., 2013). Here are a few examples of analogies that a Word2Vec model learns (Figure 1.2):

- "king" is to "queen" as "man" is to "woman" (queen = king - queen + woman)
- "Spain" is to "Madrid" as "France" is to "Paris" (madrid = paris - france + spain)
- "small" is to "little" as "big" is to "large" (little = large - big + small)
- "run" is to "jog" as "swim" is to "dive" (jog = swim - dive + run)

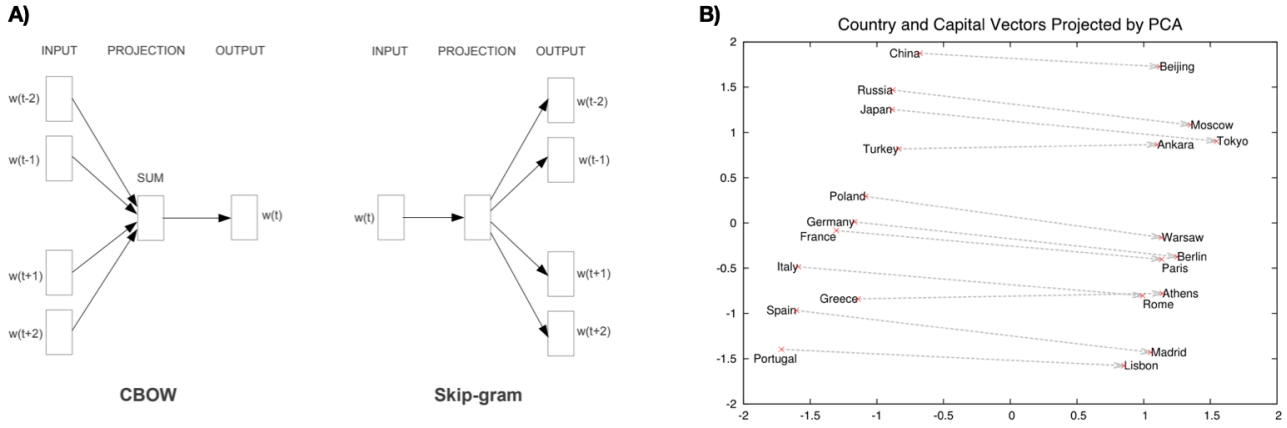


Figure 1.2: Learning lexical representations with Word2Vec. A) Architectures. In Continuous Bag Of Words (CBOW), the model predicts the current word based on context. In Skip-Gram, the model predicts the surrounding word given the current word. Extracted from (Mikolov, Chen, et al., 2013). **B)** PCA of the 1000-dimensional Skip-gram vectors of countries and their capital cities. Extracted from (Mikolov, Sutskever, et al., 2013).

Such findings are significant in two ways. First, despite their limitations, word embeddings provide a quantitative and objective characterization of lexical semantics: words that are close in the Word2Vec space are semantically similar. Second, they challenge traditional definitions of semantics, summarizing semantics to the co-occurrences of words in similar contexts.

Note on words versus other levels of vocabulary

Here, we use the term vocabulary to refer to a set of words. Other types of elementary units, such as characters, byte-pair encoding (BPE), and word-pieces, have been proposed as well. For example, BPE operates by iteratively replacing the most frequent byte (or character) pair in text with a single unused byte. This process is repeated until the desired vocabulary size is reached or the frequency of remaining byte pairs is below a certain threshold. In the following, we will use the term "word" to refer to elements in the vocabulary. Yet, most models are trained to process more granular units (Radford et al., 2019; Devlin et al., 2019; Brown et al., 2020).

1.2.3 Contextual Representations

The second issue is the integration of context. At each time step t, a contextual representation h_t is constructed, which includes information not only on the lexical representation of the word w_t, but also the representations of previous words $(w_{t-1}, \dots w_{t-k})$, with k being the context window defined beforehand. Long-Short-Term Memory (LSTMs) (Hochreiter & Schmidhuber,

1997), and Transformers (Vaswani et al., 2017) are two architectures addressing this problem and calculating the contextual representation h_t in different ways (Figure 1.3). In practice, the contextual representations h_t are learnt using different objective functions, such as next-word prediction and masked language modelling.

Architectures

Recurrent Neural Networks (RNNs) are recurrent architectures: they compute the contextual representations, or "hidden states" h_t based on the previous hidden state, h_{t-1}, and the current input, x_t.

$$h_t = g_\theta(h_{t-1}, x_t)$$

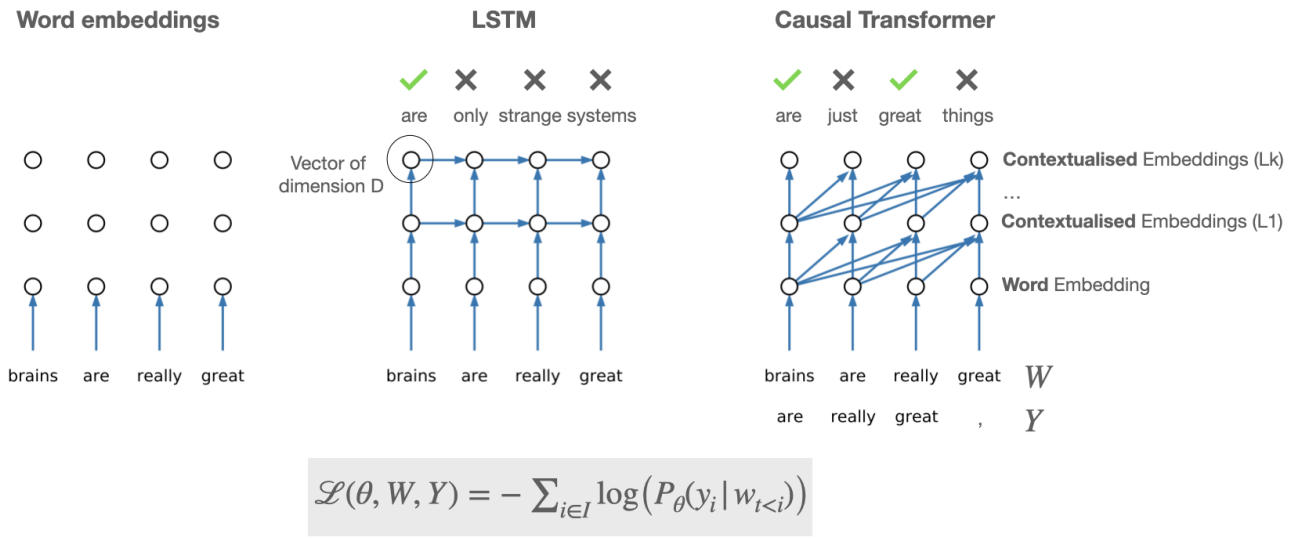
where g_θ is a non-linear function. **Long Short-Term Memory (LSTM)** networks are a particular case of RNNs that are designed to capture long-term dependencies in sequential data. LSTMs compute the current hidden state, h_t, based on the previous hidden state, h_{t-1}, the current input, x_t, and an additional "memory" cell, c_t, which is updated at each time step.

$$h_t = o_t \times \tanh(c_t)$$
$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t$$

with f_t, o_t, and i_t the forget, output, and input gates, $\tilde{c}_t$ the cell input activation vectors, each updated as follows:

$$f_t = \sigma(w_f[h_{t-1}; x_t] + b_f)$$
$$i_t = \sigma(w_i[h_{t-1}; x_t] + b_i)$$
$$o_t = \sigma(w_o[h_{t-1}; x_t] + b_o)$$
$$\tilde{c}_t = \tanh(w_c[h_{t-1}; x_t] + b_c)$$

where $\times$ is the element-wise product, σ is the sigmoid function, b and w represent the learned bias and weight of the corresponding gates. In theory, the contextual information can be carried forward throughout the sequence in the memory cell c_t. However, in practice, the gradient is said to "vanish" and the memory cell lack clear and retrievable information at the end of a long sentence. In addition, the model is recurrent and thus cannot leverage parallel computation. Transformer partly address these two issues.



$$\mathscr{L}(\theta, W, Y) = -\sum_{i \in I} \log\big(P_\theta(y_i \,|\, w_{t<i})\big)$$

Figure 1.3: Learning contextual representations with LSTMs and Transformers. Simplified computational graphs for lexical Word Embeddings (Mikolov, Sutskever, et al., 2013), causal LSTMs (Hochreiter & Schmidhuber, 1997) and causal Transformers (Vaswani et al., 2017). LSTMs combine contextual information using recurrent memory cells. Transformers combine contextual information using an attention mechanism that access all previous words in parallel. Only *causal* models are displayed here (i.e. models processing information from the left to the right).

Instead of computing the hidden state based on the previous hidden state, as in RNNs, **Transformers** compute contextual representations by directly accessing all inputs in a sequence of fixed length (Vaswani et al., 2017). They consist of a specific a contextual block called "self-attention" and a non-contextual feed-forward block (Figure 1.4). The self-attention block builds contextual representations following the equation:

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

where Q represents the queries, K represents the keys, and V represents the values. d is the dimension of the key. Q, K and V are three linear transformations of the input vectors $x = (x_{t_1}, \ldots, x_{t_n})$, with n the context length.

$$Q = W_q \cdot x$$
$$K = W_k \cdot x$$
$$V = W_v \cdot x$$

with W the weights corresponding to the keys, queries and values. The self-attention module can be interpreted as a weighted sum of input values V each weight being given by the similarity matrix $\text{softmax}(QK^T/\sqrt{d})$. The feed-forward block is applied position-wise, independently of the time step, and consists of two linear layers with a non-linearity between those.

Both LSTMs and Transformer networks generally consist of multiple contextual layers stacked onto a word embedding. The input x of layer $l+1$ is the output of layer l.

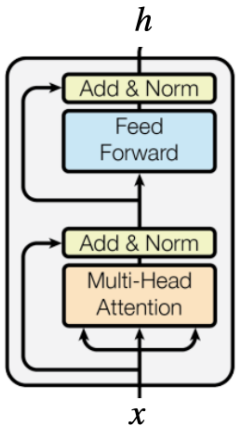


Figure 1.4: One Transformer layer. A Transformer "block", or "layer" consists of a self-attention module (orange), two fully connected layers (blue), skip connections and normalizations (yellow). We here call "hidden state" h the output of such transformer block. One Transformer model generally consists in multiple Transformer layers stacked onto a word embedding. Here, we only use encoder models, without cross-attention. Adapted from (Vaswani et al., 2017).

Since 2017, the best-performing language algorithms are based on the Transformer architecture. Some architectural improvements have been added such as the addition of a relative positional embedding (Transformer-XL, Dai et al. (2019)) or external memory that allows for the inclusion of longer contexts (Transformer-XL Dai et al. (2019), Compressive Transformer (Rae et al., 2019), Expirespan (Raikote, 2021)). However, most recent improvements in standard natural language processing benchmarks are driven by introducing new training tasks, increasing the size of the networks and training on more and higher quality data (A. Wang et al., 2018, 2020; Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020; Raffel et al., 2020).

Training tasks and datasets

Several training tasks have been proposed for learning general contextual language representations. **Language modeling** is the most straightforward task compared to our objective. It

consists of training the algorithm to predict the next word given its previous context. The famous models GPT-2 and GPT-3 developed by OpenAI are trained with language modeling objectives. **Masked language modeling** consists of predicting a masked word given its surrounding context. BERT and models derived from BERT (DistilBERT, RoBERTa (Liu et al., 2019)) are trained with masked language modeling. Other tasks include permutation language modeling in XLNet (Yang et al., 2020), span prediction in BART (Lewis et al., 2019), a combination of supervised and unsupervised objectives in T5 (Raffel et al., 2020), or adversarial approaches where a student discriminates between words generated by a teacher BERT model and real words (Clark et al., 2020).

The quantity, quality of training data, as well as the model size are key factors to learn relevant language representations. For example, the main differences between GPT-2 and GPT-3 lie in the models' sizes and the training data, yet GPT-3 systematically outperforms GPT-2 at various language tasks (Radford et al., 2021; Brown et al., 2020).

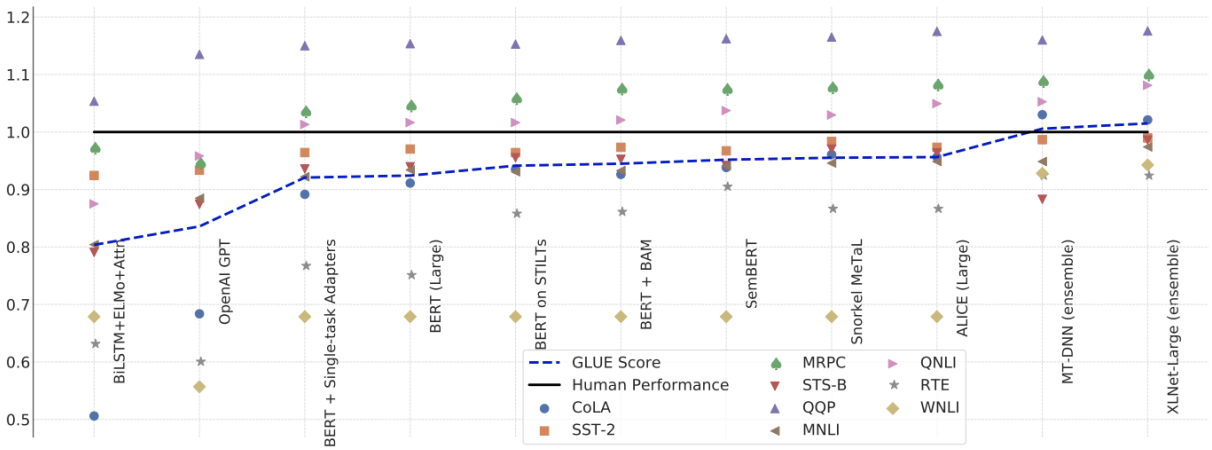


Figure 1.5: Deep networks' ability to generalize to multiple language tasks. Performance of eleven deep neural nets at GLUE, a natural language processing benchmark consisting of 9 tasks, including classifying a sentence as grammatically correct or incorrect (CoLA), as positive or negative (SST-2), classifying two sentences as semantically similar or not (STS-B, SST-Q) and identifying whether one sentence entails the other (MNLI, WNLI, RTE) (A. Wang et al., 2018). Performance is re-scaled to set human performance to 1. Extracted from (A. Wang et al., 2020).

Learning general representations of language

By utilizing specific architectures such as transformers and carefully curated training data, models can learn representations that are applicable to a wide range of language tasks. One way

to evaluate the generality of these representations is through fine-tuning, where a portion or all of the model's weights are further trained on a small set of annotated data for a different task. The performance of the fine-tuned model on the downstream task serves as an indicator of its ability to learn representations specific to that task. For example, models trained with masked language modeling and fine-tuned on downstream tasks have shown to excel at tasks such as sentiment analysis, part-of-speech tagging, named entity recognition, sentence classification, and summarization (A. Wang et al., 2018; Radford et al., 2019; A. Wang et al., 2020; Devlin et al., 2019; Brown et al., 2020; Lewis et al., 2019) (Figure 1.5).

Learning semantic and syntactic representations

Another way of assessing the generality of representations in language models is by directly analyzing the model's activations. In practice, a linear classifier can be trained to decode specific information from the activations, and the performance at this linear decoding task serves as an indicator of the type of representations encoded in the activations. For example, Jawahar et al. (2019) have shown that BERT contextual representations, specifically in middle layers, encode information such as tense, word-verb agreement, and syntactic complexity of sentences. In a different work, Manning et al. (2020) demonstrated that BERT activations contain sufficient information to reconstruct the syntactic dependency tree of sentences. Specifically, they utilize a linear projection that maps pairwise distances in the BERT activations to pairwise distances in the syntactic tree (Figure 1.6). This "structural probe" enables the authors to reconstruct a syntactic tree from BERT activations with a high degree of accuracy, achieving a correlation of 0.89 between the true and reconstructed depths and a correlation of 0.87 between the true and reconstructed pairwise distances.

1.2.4 Limitations of language models

Despite their performances, even the most recent large language models (LLMs) suffer from limitations.

Behavioural limitations.

- *Lack of Consistency*: LLMs have been trained on vast amounts of text data, but they still struggle with maintaining consistent behavior in generating text. LLMs are still sensitive to the way the prompt is provided. For instance, the query "Homeland premiered on [Y]" should yield the same answer as "Homeland originally aired on [Y]", whereas it is

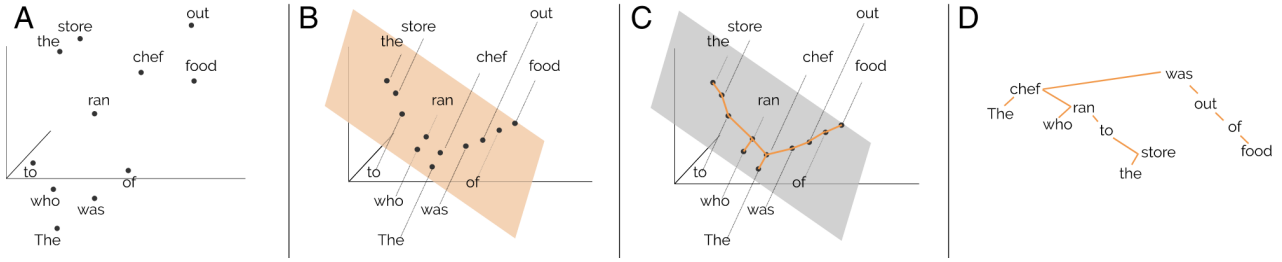


Figure 1.6: Learning syntactic representations. The structural probe method introduced in (Manning et al., 2020). The authors partly recover the syntactic dependency tree of the sentences from the sixteenth layer of a BERT model. **A.** Each of the words of the sentence "*The chef who ran to the store was out of food*" is internally represented in context as a vector. **B.** A structural probe finds a linear transform of that space under which squared L_2 distance between vectors best reconstructs tree path distance between words. **C.** Once in this latent space, the structure of the tree is globally represented by the geometry of the vector space, meaning words that are close in the space are close in the tree. **D.** In fact, the tree can be approximately recovered by taking a minimum spanning tree in the latent syntax space. Figure and caption are extracted from (Manning et al., 2020).

not systematically the case (Elazar et al., 2021). Similarly, a BERT-large model provides different answers to "Albania shares borders with [Y]" and "[Y] borders with Albania" (Figure 1.7).

- *Limited Memory*: LLMs are designed to process large amounts of text data and generate responses, but they have limited memory capacity. While several methods have been proposed to enhance LLMs with large attention spans and external memory (Raikote, 2021; Rae et al., 2019; Izacard et al., 2022), these may not be currently leveraged in the best generative models like ChatGPT, which *is said to* only access the 4,000 previous tokens[3].

- *Poor performance at Logic and Maths*: LLMs have been trained on text data and have not been specifically designed to perform well at logical and mathematical reasoning tasks. While they have shown some ability to perform simple arithmetic, they struggle with more complex logical and mathematical problems (Brown et al., 2020; Chowdhery et al., 2022; Jiang et al., 2022). Note that ChatGPT still shows proficiency in basic logical tasks like placing a green triangle to the left of a blue circle, outperforming other multi-modal models like DALLE (Ramesh et al., 2022) and StableDiffusion (Rombach et al., 2022) (Figure 1.8).

[3] https://help.openai.com/en/articles/6787051-does-chatgpt-remember-what-happened-earlier-in-the-conversation

- *Poor performance at commonsense reasoning*: The model lack some knowledge about real-world situations. For instance, when presented with a sentence like "The lawyer asked the witness a question, but he was reluctant to repeat it.", most humans would agree that "he" refers to "lawyer" and not "witness", but the model may not (Mahowald et al., 2023; Elazar et al., 2021), (Figure 1.9).

#	Subject	Object	Pattern #1	Pattern #2	Pattern #3	Pred #1	Pred #2	Pred #3
1	Adriaan Pauw	Amsterdam	[X] was born in [Y].	[X] is native to [Y].	[X] is a [Y]-born person.	Amsterdam	Madagascar	Luxembourg
2	Nissan Livina Geniss	Nissan	[X] is produced by [Y].	[X] is created by [Y].	[X], created by [Y].	Nissan	Renault	Renault
3	Albania	Serbia	[X] shares border with [Y].	[Y] borders with [X].	[Y] shares the border with [X]	Greece	Turkey	Kosovo
4	iCloud	Apple	[X] is developed by [Y].	[X], created by [Y].	[X] was created by [Y]	Microsoft	Google	Sony
5	Yahoo! Messenger	Yahoo	[X], a product created by [Y]	[X], a product developed by [Y]	[Y], that developed [X]	Microsoft	Microsoft	Microsoft
6	Wales	Cardiff	The capital of [X] is [Y] .	[X]'s capital, [Y].	[X]'s capital city, [Y].	Cardiff	Cardiff	Cardiff

Figure 1.7: Example of consistency errors from a BERT-large-cased model. Extracted from (Elazar et al., 2021). Pred #i correspond to Pattern #i. The answer depends on the phrasing of the pattern. If the model's prediction is accurate, it is colored blue. If the prediction is incorrect, it is colored red.

Algorithmic limitations. Second, there is evidence that the *way* LLMs process text is still different from the brain (Mahowald et al., 2023).

- *Unrealistic amounts of training data:* LLMs that achieve near-human performance are trained on much more data than a child is exposed to. For example, GPT-3 sees 1000x more language data than a 10-year-old human (Warstadt & Bowman, 2022).

- *Sensitivity to data curation.* While humans are continuously exposed with noisy and heterogeneous inputs, LLMs are highly sensitive to the quality of the data they are trained on: a few curated data to fine-tune the model on may yield better results that large amounts of non-currated data (Solaiman & Dennison, 2021).

Implementation limitations.

- *Power efficiency:* The brain surpasses large language models in power efficiency, even though it has a greater number of synaptic weights to update. To put it into perspective, training a model with 175 billion parameters such as GPT-3 would demand 1000 megawatt-hour of energy, whereas the brain with around 100 trillion synapses would only require 20 watt-hours, as reported by Zador et al. (2022).

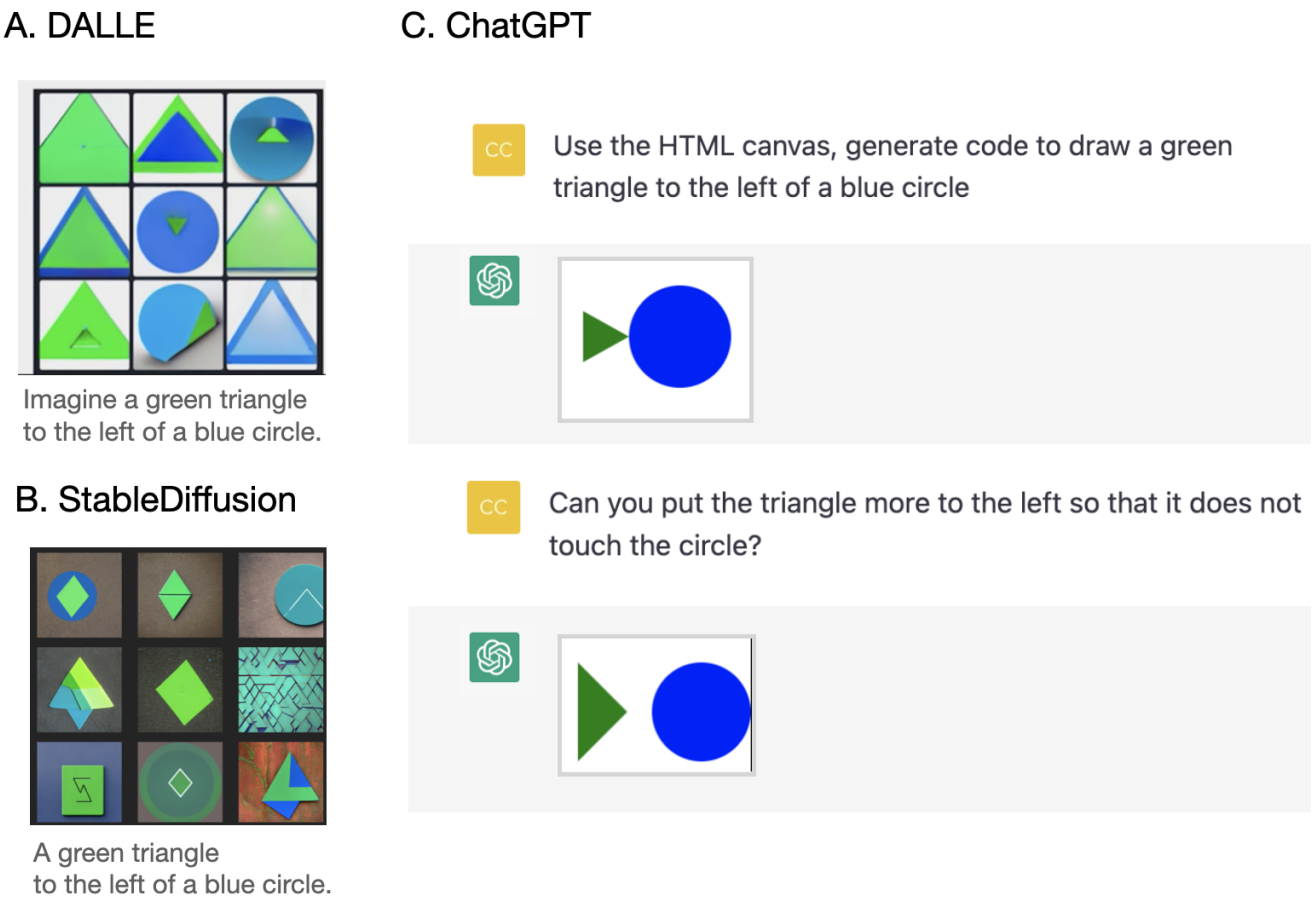


Figure 1.8: Drawing a green triangle to the left of a blue circle. The images were generated by DALLE (Ramesh et al., 2022), Stable Diffusion (Rombach et al., 2022), and ChatGPT when asked to generate a green triangle to the left of a blue circle. **A.** Example extracted from T. Desbordes PhD. presentation. **B.** Images generated on dreamstudio API.**C.** Images generated on ChatGPT API. ChatGPT's output was HTML code, we here display the corresponding HTML image.

Overall, despite their limitations, there is converging evidence that artificial neural networks build general representations of language, encoding both syntactic and semantic properties, and useful for the models to perform complex tasks such as text memorization and question answering.

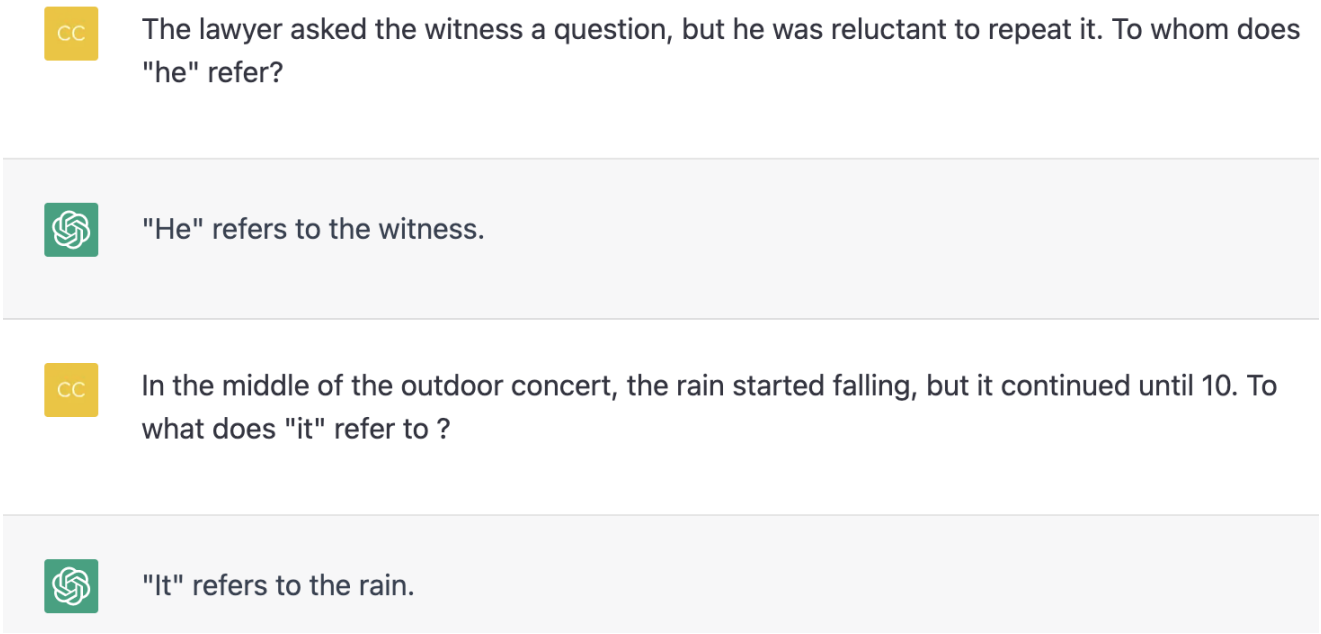


Figure 1.9: Example of non-intuitive generations from ChatGPT. Prompts from Levesque et al. (2012). Most humans would answer "lawyer" and "concert", but ChatGPT answers "witness" and "rain". Tested on ChatGPT website (06/02/2023).

1.3 Language representations in the brain

Unlike deep learning algorithms, mental representations in the brain are implicit and hardly accessible. They can only be estimated through physical observations or indirect and often noisy measurements. As a result, methodological advances in acquiring and analyzing brain signals have been crucial in the study of language processing in the brain.

1.3.1 Lesion studies

One of the oldest approaches to analyzing the neural basis of language is to compare the behaviors of patients whose brains have been damaged. The most famous examples of this approach are the studies of P. Broca and K. Wernicke, which gave rise to the Geschwind-Lichteim-Wernicke model (Lichteim, 1885; Geschwind, 1965; R. E. Graves, 1997). This model suggested that Broca's area, located in the frontal lobe, is responsible for language production, and damage to this area leads to difficulty in producing speech (expressive aphasia). Similarly, Wernicke's area, located in the temporal lobe, was proposed to be responsible for language comprehension, and damage to this area leads to difficulty in understanding speech (receptive

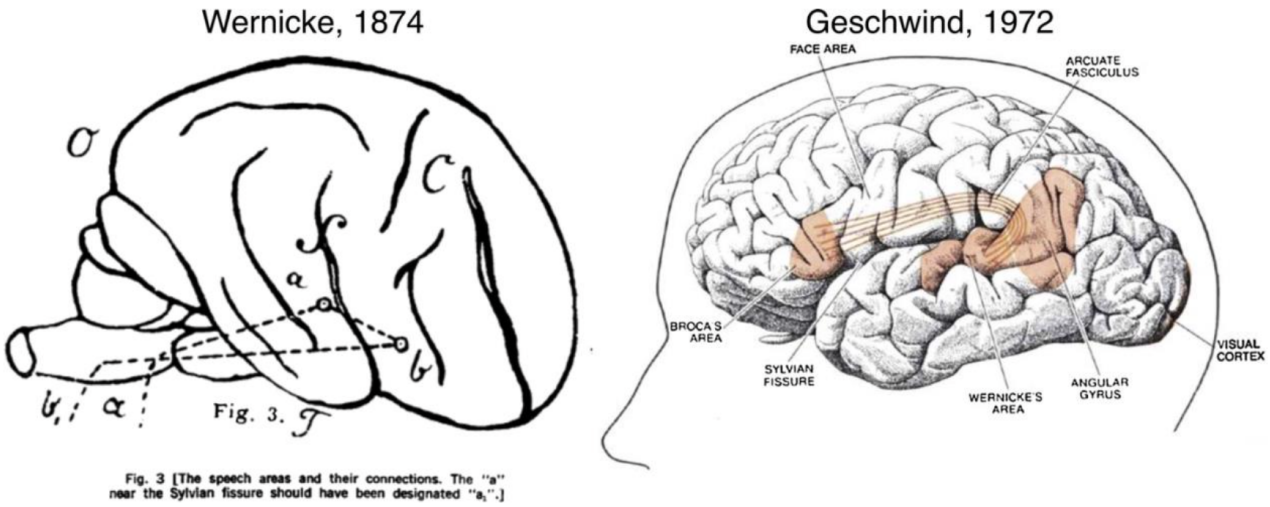


Figure 1.10: Wernicke and Geshwind models. Extracted from (Tremblay & Dick, 2016). On the left, the original Wernicke model represented in the right hemisphere. On the right, the Geschwind's updated classic model. As specified by the authors, the superior temporal gyrus is mislabeled as the angular gyrus, based on most anatomical definitions.

aphasia). This model dominated for much of the 20th century and was recently challenged by the advances of new techniques to record brain activity (Tremblay & Dick, 2016).

1.3.2 Controlled stimuli and contrast-based methods

In the early 1980s, the advent of new technologies such as intra-cortical electrophysiology, and non-invasive techniques such as functional magnetic resonance imaging (fMRI), positron emission tomography (PET), electroencephalography (EEG) and magnetoencephalography (MEG) allowed the recording of brain signals in real time. A common approach became to compare the brain responses of participants to different controlled stimuli. For example, Kutas & Federmeier (2011) showed that EEG responses to a semantic violation (e.g. "I like my coffee with cream and socks") have a larger negativity at t=400ms compared to brain responses to a control stimulus (e.g. "I like my coffee with cream and sugar"). The authors therefore concluded that the N400 (increased negativity at 400ms) was associated with this type of semantic violation. Similar approaches have been used in fMRI, analyzing brain responses to word lists vs. sentences, meaningful vs. meaningless sentences (Mazoyer et al., 1993; Humphries et al., 2006, 2007; Obleser et al., 2007; Friederici, 2011) and sentences of different syntactic properties (Fiebach et al., 2005; Makuuchi et al., 2009; Newman et al., 2010; Santi & Grodzinsky, 2010). For example, Pallier et al. (2011) recorded the fMRI responses to

sequences of varying syntactic complexity (constituent size) and semantic content (either words or pseudo-words). This allowed the authors to identify the brain regions involved in syntax, regardless of the semantic content of the phrase (Infero Frontal Gyrus and Superior Temporal Sulcus), and the regions responding to syntax only when presented to meaningful sentences (Temporo-Parietal Junction) (Figure 1.11).

Controlled studies offer a clear relationship between the hypothesis and experimental results, but suffer from several drawbacks (Jain et al., 2023), including:

- The effectiveness of controlled studies depends on the quality of the hypothesis being tested. Brain responses to controlled stimuli may not be representative of humans' natural language, and narrowing the focus to a single stimulus property can result in inaccurate conclusions that ignore interactions with other properties.
- They require the recordings of brain responses to each condition from each participants, which is costly.
- They lack flexibility. The need to design specific controls and measure results for each language property results in limited reuse of experimental data.
- Combining data from different controlled experiments can be challenging due to differences in methods, stimulus sets, and subjects (e.g. from different laboratories), and thus slows down the scientific process.

To address these limitations, there is a gradual shift of paradigm: from controlled stimuli and contrast-based methods toward natural stimuli and encoding models (Hamilton & Huth, 2018; Jain et al., 2023).

1.3.3 Toward natural stimuli and encoding methods

Semi-controlled stimuli. Instead of comparing brain responses to fixed-size, out-of-context phrases, studies analyze brain responses to phrases in context. For example, Lerner et al. (2011) analyze the fMRI recordings of participants in response to natural stories, and compare them to the brain responses of the same subjects when the words, phrases, and paragraphs were scrambled (Figure 1.12). This approach allows the authors to identify the brain regions involved in the processing of short, medium and long-term dependencies. Another example is that of Fedorenko et al. (2016). The authors compare intracranial responses to meaningful sentences,

Table 1. The stimuli were 12 items long sequences obtained by concatenating constituents of fixed sizes extracted from natural or jabberwocky right-branching sentences

Condition	Constituent size	Examples
c12	12 words	I believe that you should accept the proposal of your new associate
		I tosieve that you should begept the tropufal of your tew viroate
c06	6 words	the mouse that eats our cheese two clients examine this nice couch
		the couse that rits our treeve fow plients afomine this kice bloch
c04	4 words	mayor of the city he hates this color they read their names
		tuyor of the roty he futes this dator they gead their wames
c03	3 words	solving a problem repair the ceiling he keeps reading will buy some
		relging a grathem regair the fraping he meeps bouding will doy some
c02	2 words	looking ahead important task who dies his dog few holes they write
		troking ahead omirpant fran who mies his gog few biles they grite
c01	1 word	thing very tree where of watching copy tensed they states heart plus
		thang very gree where of wurthing napy gunsed they otes blart trus

In jabberwocky, all content words were replaced with pseudowords (*italics*). Examples are only illustrative, because the original stimuli were in French.

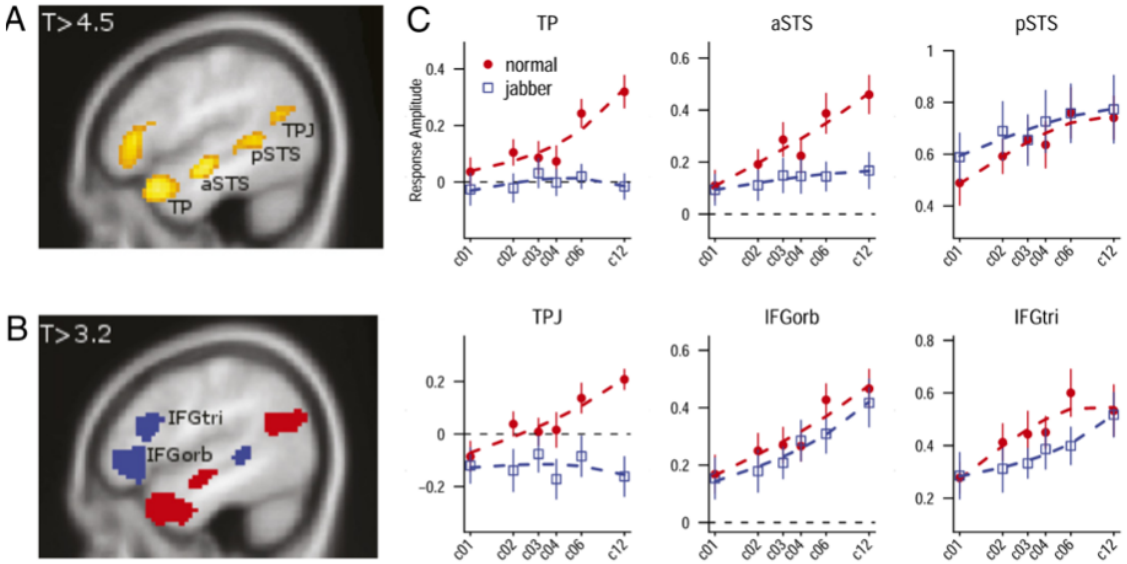


Figure 1.11: Using controlled stimuli and factorial designs to disentangle syntax and semantics. Extracted from (Pallier et al., 2011). The authors study the fMRI signals of 40 subjects reading sentences with varying syntactic and semantic properties. On the top, the 12 x 2 types of stimuli used (12 possible component sizes, either words or pseudo-words (jabberwocky, in italics). On the bottom, the regions with significant increase with constituent size with normal words (A) and in both the normal and pseudo-words (B, blue areas). In (C), the fMRI amplitude response varying with the constituent size, for normal words (red) or pseudo-words (blue), for different regions in the brain.

word lists, jabberwocky and non-word lists. The results show that the increase in activity is specific to sentence comprehension, and is not fully explained by responses to syntax or word meaning alone, evidencing compositional processes of sentence.

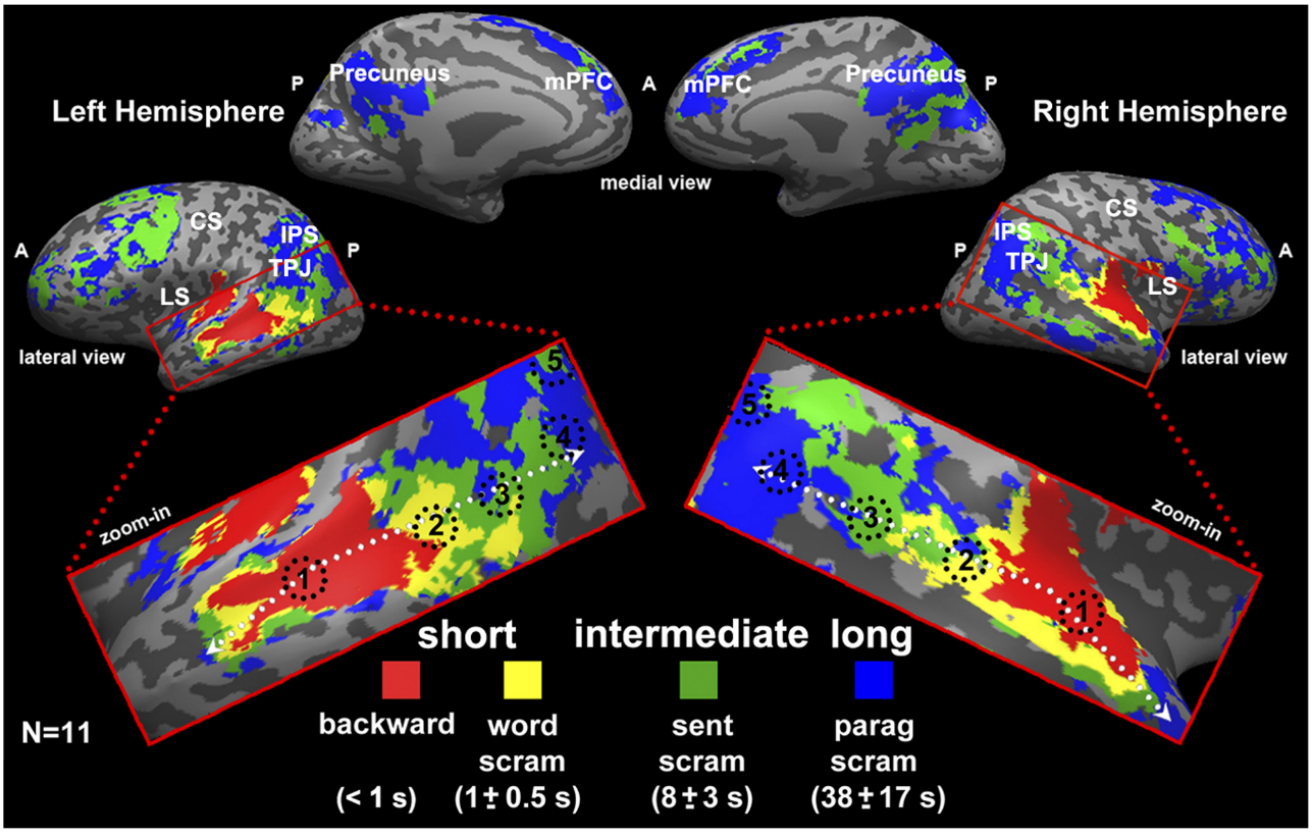


Figure 1.12: Using semi-controlled stimuli to analyse language processing in natural settings. Extracted from (Lerner et al., 2011). The authors study the fMRI activity of 15 participants in response to a 7 min natural story, the same story scrambled at the paragraph, sentence and word level, as well as the story played in reverse. Blue voxels are significant only in the paragraph scram condition, green voxels are significant in the sentence scram and paragraph conditions, yellow voxels are significant in the word, sentence and paragraph conditions and red voxels are significant in all conditions.

Natural stimuli and encoding models. While semi-controlled stimuli allow the analysis of sentences in context, they still require the recordings of subjects in multiple non-natural conditions (e.g. scrambling test and jabberwocky). Instead of using contrast-based methods to compare the brain responses to semi-controlled stimuli, studies have recently used *predictive models* of brain responses to *natural stimuli alone*. The idea is the following: participants attend to natural stimuli, one builds predictive models of their brain responses with constrained information as inputs. Finally, the model is validated on its ability to predict held out brain responses. For example, if an model based on a word's part-of-speech linearly predicts a particular voxel, it will be inferred that this voxel is involved in processing the word's part-of-speech. These models are called "encoding models". The approach is no longer to compare

brain responses to *multiple conditions*, but the ability *multiple models* to accurately predict brain responses to natural stimuli (Naselaris et al., 2011; Hamilton & Huth, 2018; King et al., 2018).

1.3.4 Deep encoding models of the sensory cortex

Deep learning algorithms as good encoding models? One challenge of encoding methods combined with natural stimuli is to find good candidate models. The question of what constitutes a good candidate model is not specific to the field of language, but has been investigated in vision research. In 2016, D. L. Yamins & DiCarlo (2016) proposed three minimal criteria for a good encoding model. Following their wordings, the three minimal criteria are:

- Stimulus-computability: The model should accept arbitrary stimuli within the general stimulus domain of interest;
- Mappability: The components of the model should correspond to experimentally definable components of the neural system;
- Predictivity: The units of the model should provide detailed predictions of stimulus-by-stimulus responses, for arbitrarily chosen neurons in each mapped area.

Deep learning models meet these criteria. They can process a wide range of stimuli within their training domain (stimulus-computability), they have a hierarchical processing architecture, so each layer is mappable to different brain regions (mappability); and finally, it is possible to evaluate the predictivity of each brain sensor (D. L. K. Yamins & DiCarlo, 2016). Thus, artificial neural networks have early been used to encode brain responses, particularly in vision research.

Visual cortex. Several works have early used distributed encoding features and Convolutional Neural Networks (CNNs) to predict brain responses to images (Kay et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011; Huth et al., 2012; Cadieu et al., 2014; D. L. K. Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Sermanet et al., 2014; Güçlü & Gerven, 2015; Eickenberg et al., 2017). For instance, in 2014, Yamins et al. examined CNNs trained on image classification tasks by comparing them to the neural responses of the Inferior Temporal (IT) visual cortex in two macaques (D. L. K. Yamins et al., 2014). They analyzed the brain activity and CNN activations in response to the same images and used a neural predictivity metric to measure the alignment between the two. Specifically, they calculated the r-squared value between predicted and actual brain responses after a linear projection was learned using separate data. The study

found that CNNs accurately predict IT responses to images and that different regions of the macaques' visual cortex are better explained by different layers of the network (Figure 1.13b-c). Importantly, they also demonstrated that architectures that perform well on high-level object recognition tasks also better predict IT responses (D. L. K. Yamins et al., 2014) (Figure 1.13a). Subsequent research extended these findings to human subjects using functional magnetic resonance imaging (fMRI). For example, Khaligh-Razavi & Kriegeskorte (2014) found that the human V1-V3 regions were best explained by the second layer of a CNN optimized for object recognition, while the IT region was best explained by the top layer (Figure 1.13e).

Auditory cortex. Similar approaches have also been applied in speech recognition, where A. J. E. Kell et al. (2018) found that CNNs accurately predict the human auditory cortex recorded using fMRI, with different layers of the network explaining different parts of the cortical hierarchy.

Benchmarks. These encoding approaches have the advantage of being more easily shared among research labs, in contrast to factorial designs. In an effort to promote replication and collaboration in the field, Schrimpf et al. (2018) introduced the "Brain Score" benchmark, which quantitatively compares the ability of deep learning algorithms to decode brain responses to images.

1.3.5 Deep encoding models of neural responses to language

Lexical responses. Distributed word representations were early used to encode brain activity. Mitchell et al. (2008) used co-occurences properties to build semantic embeddings of words, and used a linear encoding model to predict fMRI given the embeddings. This approach allowed the authors to predict fMRI responses to 60 words, including words absent from the training set. Using similar techniques, Huth, de Heer, et al. (2016) used lexical word embeddings to predict fMRI responses to *natural* stories, and found that word embeddings accurately predicted brain responses to speech (Figure 1.14). Critically, the authors derive a "semantic atlas" and identify the regions coding for specific semantic dimensions in the brain.

Contextual responses. Wehbe et al. (2014) went beyond lexical representations and used recurrent neural networks to build predictive models of MEG responses to visual sentences. They showed that MEG responses were partly accounted for by the contextual representations of RNNs, both before and during the word presentations. Jain & Huth (2018) later leveraged the

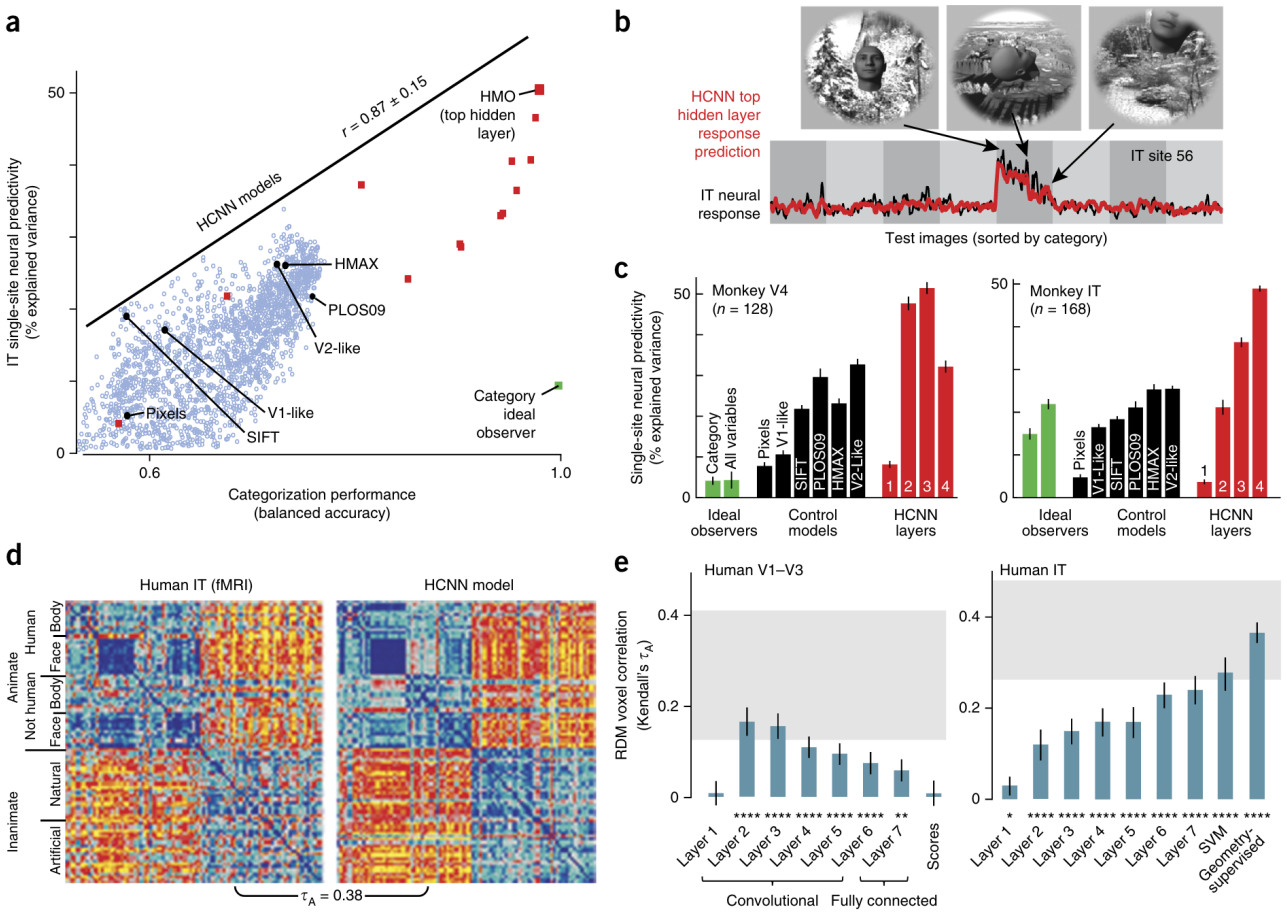


Figure 1.13: Convolutional neural nets encode the visual cortex. Extracted from (D. L. K. Yamins & DiCarlo, 2016). **a)** Performance of deep convolutional networks (CNNs) at image classification (x-axis) versus their neural predictivity, i.e. their ability to linearly encode the Inferior Temporal neuronal responses (IT) of monkeys, or (y-axis). **b)** Example of true monkey neural response (black) vs. the response predicted by the last layer of a deep CNN (red). **c)** Comparison of IT and V4 single-site neural predictivity. **d)** Representational dissimilarity matrices for human fMRI and deep CNN model (from low similarity in blue to high in yellow). **e)** Correlation between brain and deep net similarity matrices. d and e are adapted from (Khaligh-Razavi & Kriegeskorte, 2014).

advantages of LSTMs to investigate the effect of context on the predictability of fMRI responses. Precisely, they used a three-layer-LSTM and computed the brain score of its activations when fed with increasingly more context. They found that brain scores are sensitive to the context length up to around 15 sentences, and that the middle layer best encoded fMRI responses as opposed to word embeddings and the output layer (Figure 1.15).

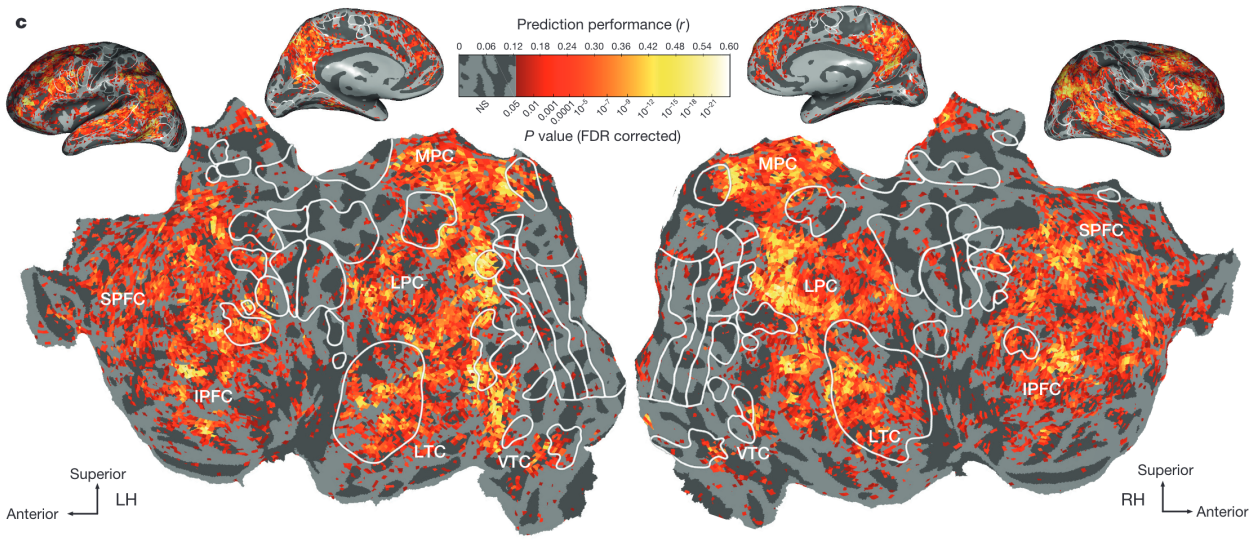


Figure 1.14: Lexical word embeddings encode brain responses to speech. Extracted from (Huth, de Heer, et al., 2016). Ability of a 985-dimensional word embedding to predict the fMRI responses to 2 hours of natural stories from seven subjects.

An active field of inquiry. My thesis was made possible by the rapid growth of works at the interface between AI and neuroscience, which expanded in tandem with my studies. This expansion has been supported by a growing open-source community, making deep language models and neuro-imaging datasets publicly available (Wolf et al., 2020; Nastase et al., 2020). Thus, during my PhD, and together with our works, several teams explored contextual language representations in deep learning algorithms and the brain. Multiple studies have shown a linear correlation between brain responses and deep language models' activations (Jat et al., 2019; Hollenstein et al., 2019; Schrimpf et al., 2021; Toneva, Stretcu, et al., 2020; Toneva, Mitchell, & Wehbe, 2020a,b; Toneva & Wehbe, 2019; Reddy & Wehbe, 2020; Sun et al., 2021; Anderson et al., 2021; S. Wang, Zhang, Wang, et al., 2020; Vaidya et al., 2022; Jain et al., 2023), with some investigating the factors that modulate this mapping (Caucheteux & King, 2022; Schrimpf et al., 2021; Antonello & Huth, 2022; Pasquiou et al., 2022; Goldstein et al., 2022). Some studies have reinforced the importance of next-word prediction in the mapping (Schrimpf et al., 2021; Goldstein et al., 2022), while others have challenged this notion (Pasquiou et al., 2022). Other studies have explored the effects of context (Jain & Huth, 2018), layer (Toneva & Wehbe, 2019; Jain & Huth, 2018; Vaidya et al., 2022), syntax, and semantics on brain mapping (Reddy & Wehbe, 2020; S. Wang, Zhang, Lin, & Zong, 2020; Pasquiou et al., 2023). Further details on these studies will be discussed later in the manuscript.

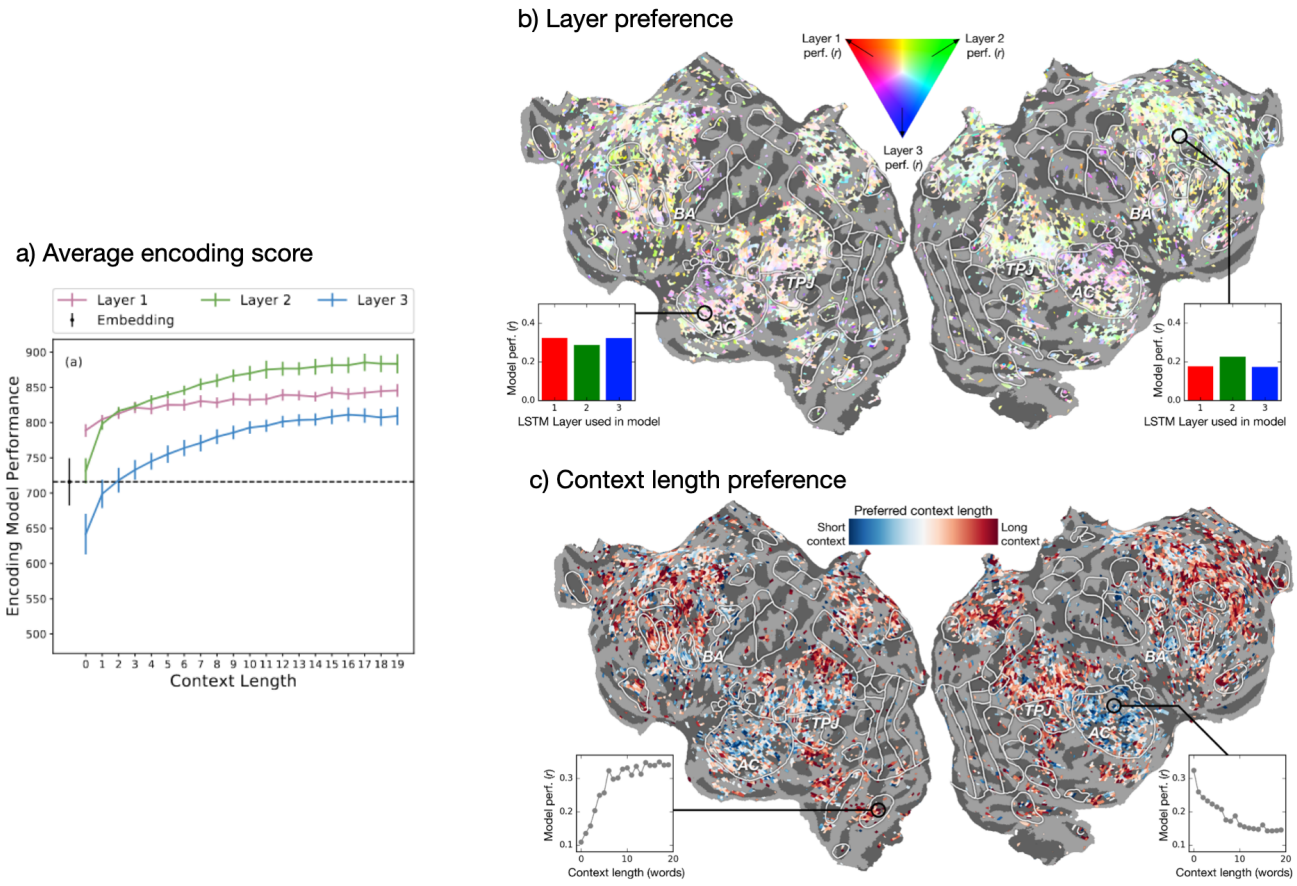


Figure 1.15: Contextual LSTMs encode brain responses to speech. Adapted from (Jain & Huth, 2018). The authors study the ability of an LSTM to predict fMRI responses to language, when the model is input with a varying amount of context. **a)** Ability of the model to predict brain responses (y-axis) as a function of the number of sentences in the context (x-axis), for each layer of the LSTM. **b-c)** Layer (b) and context length (c) maximizing the prediction performance.

1.3.6 Summary

Overall, at the beginning of my thesis, a large body of works had separately explored language representation in algorithms and the brain, demonstrating that deep neural networks create meaningful semantic and syntactic spaces (Jawahar et al., 2019; Manning et al., 2020; Mahowald et al., 2023), and clarifying the spatial and temporal dynamics behind language processes in the brain (Hickok & Poeppel, 2007; Kutas & Federmeier, 2011; Lerner et al., 2011; Fedorenko et al., 2016). A smaller body of works used linear encoding models based on artificial neural networks to predict brain responses to natural language. These works highlight similarities between word embeddings, the contextual layers of recurrent neural networks and brain acti-

vations (Wehbe et al., 2014; Huth, Lee, et al., 2016; Jain & Huth, 2018). The present manuscript extends this line of research, focusing on Transformer models, investigating multiple recording modalities (MEG and fMRI), large cohorts of English, French and Dutch participants (above 500 participants in total), as well as a large number of deep language models ($>$ 1000 language models). Furthermore, the present thesis investigates the *nature* of the shared representations (e.g. syntactic vs. semantic representations), and the *factors* modulating the similarity with the brain (e.g. language modeling performance and participants' level of comprehension), by re-training deep models *from scratch* to enable controlled comparisons. Critically, while most works focused on the *similarities* between the two systems, we investigate one major *difference*: the ability to predict long-range and hierarchical representations of the future, aligning with recent perspectives in artificial intelligence that emphasize the importance of hierarchical planning (LeCun, 2022).

In the following sections, we precise our general approach and the thesis contributions.

1.4 Approach

To study the similarity between brain and deep nets' representations of language, we build on previous research and directly quantify the similarity between brain and deep nets' activations in response to the same word sequences (Huth, de Heer, et al., 2016; Schrimpf et al., 2018; Jain & Huth, 2018). For example, in Figure 1.16, we input the algorithm with the sequence "not very happy", we extract the activations of each of its layers, and we compare these activations to the brain recordings of subjects who read or listen to the sequence "not very happy". Once we have extracted the two activation spaces in response to the same stimuli, we quantify their similarity using a relatively standard metric, which we will call here the "Brain Score", following D. L. K. Yamins et al. (2014); Schrimpf et al. (2018). Below, we precise the notions of representation, the brain score computation, and the methodological scope of our study.

1.4.1 Deep networks' activations

On the artificial networks side, we restrict our analysis to:

- i) **text**, to focus on language itself and not speech, i.e. sensory processing (except in Section 3.3);

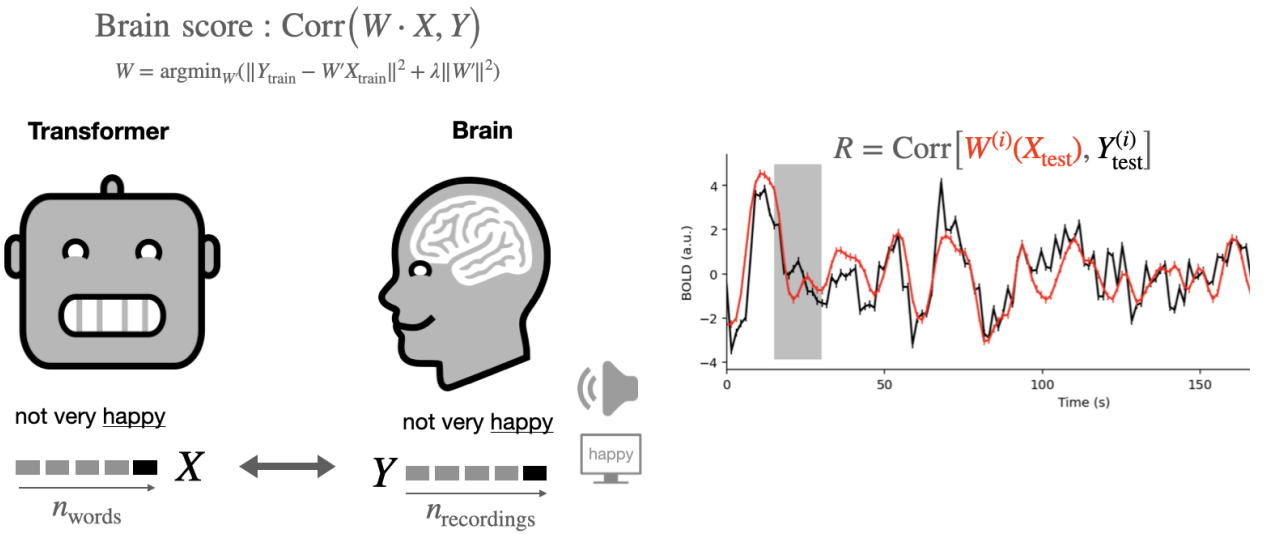


Figure 1.16: Approach. Participants either read or listen to sentences while their brain activity is recorded using functional MRI or MEG. We extract the corresponding activations from artificial neural network, from each layer. To quantify the similarity between the network activations X and the brain activations, approximated by the neuro-imaging recordings Y, we quantify the linear mapping between the two. Precisely, we fit a ℓ_2-penalized linear regression that predicts Y given X, and assess the Pearson's correlation between predicted and actual brain responses on held-out data. Such correlation score is hereafter called "Brain Score", following (Schrimpf et al., 2018). On the right, an example of the true vs. predicted fMRI response of one voxel.

- ii) **transformer-based models**, because they are the best-performing and most widely used architectures (Vaswani et al., 2017);
- iii) models trained with causal language modeling, i.e. predicting a word from its previous context, and masked language modeling, i.e. predicting a word given its surrounding context (both left and right).

In practice, we use the Huggingface (Wolf et al., 2020) or XLM implementation (Lample & Conneau, 2019) of GPT-2 (Radford et al., 2021) and BERT (Devlin et al., 2019). These models consist of twelve transformer blocks, called "layers", stacked onto a word embedding, i.e. a look-up table that assigns one vector for each vocabulary word, independently of its context.

As internal representations, we choose to study the **activations** of each layer i.e. the output of the activation function, after the multi-layer-perceptron, skip-connection and transformer layer (Figure 1.4). These are the contextual representations h_t introduced in section 1.2.3.

We thus restrict ourselves to the following methodological scope:

Language representations in transformer-based models are the activations of each layer, in response to text.

1.4.2 Brain activations

Language stimuli. We focus on **perceived** language: participants read or listen to either isolated words, isolated sentences or narratives.

Brain recordings. We define neural representations as information that is linearly readable from brain activations (Kriegeskorte et al., 2008; DiCarlo & Cox, 2007). To approximate brain activations, we use Magneto-Encephalographie (MEG) recordings, or the BOLD response measured with functional Magnetic Resonance Imaging (fMRI) (Figure 1.17).

- Magnetoencephalography measures the magnetic field induced by neurons' electrical currents. It has high temporal resolution (≈1 ms) but low spatial resolution (≈2 cm) (Baillet, 2017).
- BOLD-contrast fMRI measures the local fluctuations in blood oxygen induced by increased neuronal activity. Firing neurons require oxygen, which triggers local changes of deoxyhemoglobin and oxyhemoglobin. Such molecules have different magnetic properties and their relative volume can be estimated with fMRI. Yet, the blood oxygen flow is delayed compared to the neuronal activity, the measurements are thus a convolution of multiple sources (Figure 1.17b). fMRI has high spatial resolution (≈2 mm), but low temporal resolution (≈2 s) (Poldrack et al., 2011).

Datasets. Along this manuscript, we focus on two datasets. The Narratives dataset (Nastase et al., 2020), which contains the fMRI recordings of 345 native English speakers learning to short stories, from 7 min to 56 min, for a total of 4,6 hours of unique stimuli. The Mother Of Unification Studies (MOUS) dataset (Schoffelen et al., 2019), which contains both the fMRI and MEG recordings of 204 native Dutch speakers reading isolated words or sentences.

We thus restrict ourselves to the following methodological scope:

Language representations in the brain are information that is linearly readable from neuro-imaging recordings, in response to perceived words, sentences or narratives.

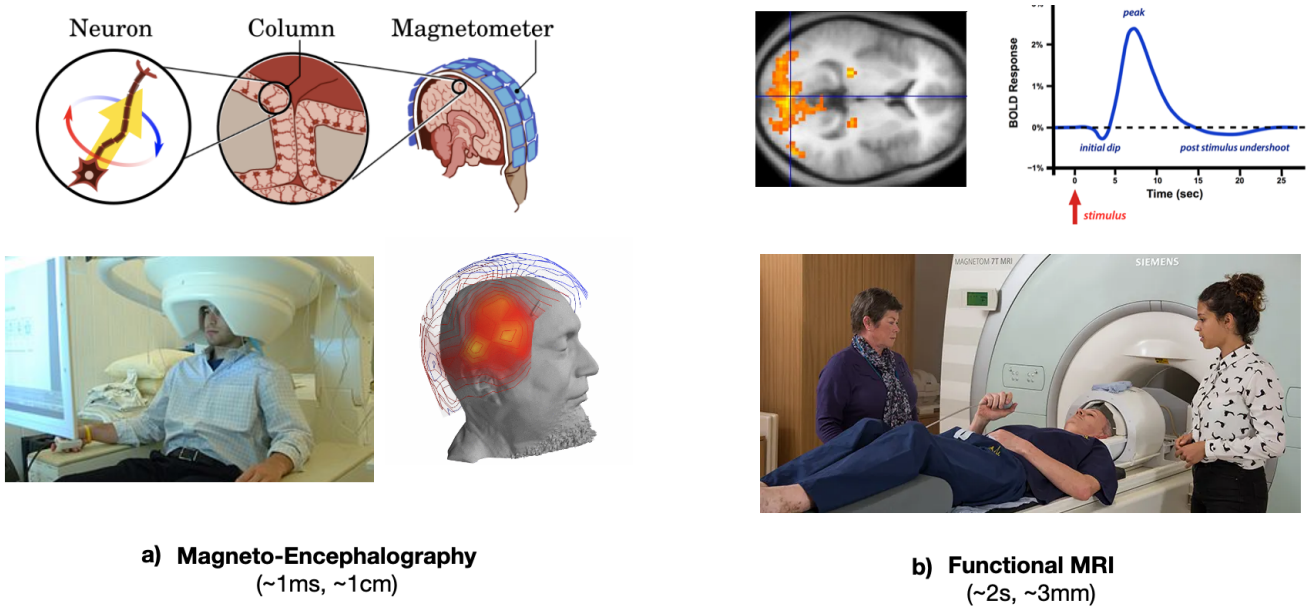


Figure 1.17: Neuro-imaging techniques used in this manuscript.

1.4.3 Quantifying similarity with the Brain Score

To assess the similarity between deep networks' and brain activations, we quantify the linear mapping between the two spaces, in response to the same language stimulus. Such an approach is motivated by previous works (Huth, Lee, et al., 2016; Schrimpf et al., 2018) and by the assumption that brain representations are linearly readable information from brain activity (DiCarlo & Cox, 2007; Kriegeskorte et al., 2008).

Precisely, we evaluate, for each subject s and sensor v (either channel for MEG or voxel for fMRI), the mapping between 1) the brain recordings $Y^{(s,v)}$ in response to the sentences and 2) the activations X of the deep network input with the textual transcripts of the same sentences. To this end, we fit a linear ridge regression W on a train set to predict the brain recordings given the network's activations. Finally, we evaluate this mapping by computing the Pearson correlation between predicted and actual brain recordings on a held out set:

$$\mathcal{R}^{(s,v)} : X \mapsto \text{Corr}\left(W \cdot X, Y^{(s,v)}\right) \quad , \tag{1.1}$$

with W the fitted linear projection, Corr Pearson's correlation, X the activations of one artificial neural network and $Y^{(s,v)}$ the brain recordings of one subject s at one sensor v, both elicited by the same held out stories.

Temporal alignment for fMRI. Following (Huth, de Heer, et al., 2016), we model the slow bold response thanks to a finite impulse response (FIR) model with five to six delays (e.g. from 0 to 9 seconds in the Narratives dataset where TR=1.5 seconds). Still following Huth, de Heer, et al. (2016), we sum the model activations of the words presented within the same TR, in order to match the sampling frequency of the fMRI and the language models. In the notations above, X consists in the deep networks activations after applying FIR.

Temporal alignment for MEG. In the case of MEG recordings (Section 2.1), there is more recordings than words. Thus, we epoch the recordings on the word onsets, fit and evaluate a ridge regression for each time step starting at the word onset.

1.5 Overview of the thesis

Do artificial neural networks and the human brain build similar intermediate representations to process language? We approach the question using transformer-based artificial neural networks, linear encoding models, and the neuro-imaging recordings of large cohorts of participants.

Chapter 2: High-level similarity in language representations

MEG and fMRI responses consistently correlate with deep networks' activations. In two first papers (Caucheteux & King, 2022; Caucheteux et al., 2022), we provide evidence of high-level similarities between brain and artificial networks' activations in response to language. We show that deep nets' activations significantly predict brain activity across subjects for different cohorts ($>$ 500 participants in total), different recording modalities (MEG and fMRI), stimulus types (isolated words, sentences and natural stories), stimulus modalities (auditory and visual presentation), languages (Dutch and English) and deep language models (models varying in training tasks, architectures and performances). This alignment is maximal in brain regions repeatedly associated with language, in the intermediate layers of algorithms, and for the best performing algorithms (i.e., those that best predict a word from its context) (Caucheteux & King, 2022; Caucheteux et al., 2022).

Next-word prediction primarily impacts the brain score. We then ask *why* artificial neural networks' activations correlate with brain responses (Kanwisher et al., 2023). To this aim, we study $>$ 1,000 transformer models varying in training task, architectural parameters and

performance, and compute the brain score of each layer of each network. We find that the depth of the representation as well as the model's language modeling perplexity are the primary factors driving the brain score, above other architectural parameters (e.g. number of layers and dimensionality). Interestingly, we find a non-monotonic relationship between the brain score and the performance of the network: the best-performing neural networks slightly diverge from brain-like representations of language, whereas the networks are still improving on their training task (Caucheteux & King, 2022).

Participant's comprehension affects the brain score. Finally, we investigate whether the participant's level of comprehension affects the similarity. We study the fMRI responses of 100 participants to natural stories and show that subjects who understand stories the best (their comprehension level is assessed using a 30-question questionnaire at the end of each story) are those whose brain activations are the most similar to those of deep networks. Such an effect is stronger for the contextual representations of transformers, is sensitive to very long-term dependencies (above 700 words) and peaks in areas previously associated with high-level semantics (Angular, Supramarginal gyri and the Precuneus) (Caucheteux et al., 2022).

Chapter 3: Leveraging the similarity to decompose the content, temporal and spatial organization of language representations in the brain

In this chapter, we illustrate how encoding models combined with artificial networks can be used to decompose natural language processes in the brain.

Syntax and Semantics. In a first paper, we introduce a simple method to decompose the activations of language models into their syntactic and semantic components. We apply such methods to the activations of GPT-2 and identify the regions involved in lexical and compositional syntax and semantics in the fMRI recordings of 345 participants. (Caucheteux et al., 2021a).

Temporal hierarchy. In a second paper, we compare the findings of Lerner et al. (2011) obtained with factorial designs to our results obtained using linear encoding methods combined with deep language models. We show that encoding methods applied to natural stimuli are able recover the results of Lerner et al., and extend those to precise the processing of short-to-long range dependencies in the brain of 345 participants (Caucheteux et al., 2021b).

Spatial hierarchy. Finally, in a third paper, we compute the brain scores of each layer of Wav2Vec2, a model based on speech and find that the layer hierarchy of Wav2Vec2 maps onto the cortical hierarchy: the first layers of the model are aligned with lower-level processing regions in the brain, and deep layers with regions associated with higher-level processing. Furthermore, training the exact same model on different languages (English, French, Mandarin) and comparing its activations to the fMRI responses of native English, French and Mandarin participants provide some preliminary evidence of acoustic, speech and language specifity in the brain. (Millet et al., 2022).

Chapter 4: Improving the similarity through hierarchical predictions

The similarity between deep networks and the brain remains partial: the "Brain Score" is low, and dialogue, question-answering, and text-generation algorithms are still imperfect. How to build algorithms more similar to brain activity? We explore this question in a last paper (Caucheteux et al., 2023), and show that algorithms predict short-term and word-level representations, unlike the brain, which predicts long-term and contextual representations of the future. Furthermore, we show that fine-tuning GPT-2 to predict longer-term and more contextual representations increase its similarity with the brain, specially in Supramarginal and Angular gyri (Caucheteux et al., 2023).

1.6 Publications included in the thesis

Chapter 2

- Caucheteux, C., & King, J.-R. 2022. Brains and algorithms partially converge in natural language processing. *Nature Communications Biology*
- Caucheteux, C., Gramfort, A., & King, J.-R. 2022. Deep language algorithms predict semantic comprehension from brain activity. *Nature Scientific Reports*.

Chapter 3

- Caucheteux, C., Gramfort, A., & King, J.-R. 2021. Disentangling syntax and semantics in the brain with deep networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021).*

- Caucheteux, C., Gramfort, A., & King, J.-R. 2021. Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. In *Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP 2021).*

- Millet*, J., Caucheteux*, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., King, J.-R. 2022. Toward a realistic model of speech processing in the brain with self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS 2022).*

Chapter 4

- Caucheteux, C., Gramfort, A., & King, J.-R. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*.

1.7 Publication *not* included in the thesis

- Defossez, A., Caucheteux, C., Rapin, J., Kabeli, O., King, J.-R. 2023. Decoding speech from non-invasive brain recordings. *Under Review*.

Chapter 2

High-level similarity in language representations

2.1 Brains and algorithms partially converge in Natural Language Processing

2.1.1 Abstract

Deep learning algorithms trained to predict masked words from large amount of text have recently been shown to generate activations similar to those of the human brain. However, what drives this similarity remains currently unknown. Here, we systematically compare a variety of deep language models to identify the computational principles that lead them to generate brain-like representations of sentences. Specifically, we analyze the brain responses to 400 isolated sentences in a large cohort of 102 subjects, each recorded for two hours with functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG). We then test where and when each of these algorithms maps onto the brain responses. Finally, we estimate how the architecture, training, and performance of these models independently account for the generation of brain-like representations. Our analyses reveal two main findings. First, the similarity between the algorithms and the brain primarily depends on their ability to predict words from context. Second, this similarity reveals the rise and maintenance of perceptual, lexical, and compositional representations within each cortical region. Overall, this study shows that modern language algorithms partially converge towards brain-like solutions, and thus delineates a promising path to unravel the foundations of natural language processing.

2.1.2 Introduction

Deep learning algorithms have recently made considerable progress in developing abilities generally considered unique to the human species (Turing, 2009; Chomsky, 2006; Dehaene et al., 2018). Language transformers, in particular, can complete, translate, and summarize texts with an unprecedented accuracy (Vaswani et al., 2017; Devlin et al., 2019; Lample & Conneau, 2019; Brown et al., 2020). These advances raise a major question: do these algorithms process words and sentences like the human brain?

Recent neuroimaging studies suggest that they might – at least partially (Lakretz et al., 2019; B. Lake & Baroni, 2018; Hale et al., 2021; B. M. Lake & Murphy, 2021; Marcus, 2018). First, word embeddings – high dimensional dense vectors trained to predict lexical neighborhood (Bengio et al., 2001; Mikolov, Chen, et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017) – have been shown to linearly map onto the brain responses elicited by words presented either in isolation (Mitchell et al., 2008; Anderson et al., 2019; Sassenhagen & Fiebach, 2019) or within narratives (Caucheteux et al., 2022; Oota et al., 2018; Abnar et al., 2019; Ruan et al., 2016; Brodbeck et al., 2018; Gauthier & Ivanova, 2018; Wehbe et al., 2014; Schrimpf et al., 2021; Caucheteux et al., 2021a,b; Goldstein et al., 2022). Second, the contextualized activations of language transformers improve the precision of this mapping, especially in the prefrontal, temporal and parietal cortices (Jain & Huth, 2018; Athanasiou et al., 2018; Toneva & Wehbe, 2019). Third, specific computations of deep language models, such as the estimations of word surprisal (i.e. the probability of a word given its context) and the parsing of syntactic constituents have been shown to correlate with evoked related potentials (Heilbron & Chait, 2018; J. R. Brennan & Pylkkänen, 2017; Hale et al., 2018; Goldstein et al., 2022) and functional Magnetic Resonance Imaging (fMRI) (Caucheteux et al., 2021a; Hale et al., 2018). However, the above studies remain fragmentary: first, most only analyze a small number of subjects (although see (Caucheteux et al., 2022, 2021b,a)). Second, most studies only explore the spatial but not the temporal properties of the brain responses to language (although (Toneva & Wehbe, 2019; Goldstein et al., 2022)).

More critically, the principles that lead a deep language models to generate brain-like representations remain largely unknown. Indeed, past studies only investigated a small set of pretrained language models that typically vary in dimensionality, architecture, training objective, and training corpus. The inherent correlations between these multiple factors thus prevent identifying those that lead algorithms to generate brain-like representations.

To address this issue, we systematically compare a wide variety of deep language models in light of human brain responses to sentences (Figure 2.1). Specifically, we analyze the brain

activity of 102 healthy adults, recorded with both functional magnetic resonance imaging (fMRI) and source-localized magneto-encephalography (MEG). During these two 1 h-long sessions the subjects read isolated Dutch sentences composed of 9 to 15 words (Schoffelen et al., 2019). After quantifying the signal-to-noise ratio of the brain responses (Figure 2.2), we train a variety of deep learning algorithms, extract their responses to the very same sentences and compare their ability to linearly map onto the fMRI and MEG brain recordings. Finally, we assess how the training, the architecture, and the word-prediction performance independently explains the brain-similarity of these algorithms and localize this convergence in both space and time.

We find that (1) a variety of deep learning algorithms linearly map onto the brain areas associated with reading (Figure 2.3), (2) the best brain-mapping are obtained from the middle layers of deep language models and, critically, we show that (3) whether an algorithm maps onto the brain primarily depends on its ability to predict words context (Figure 2.4).

2.1.3 Results

Shared brain responses to words and sentences across subjects.

Before comparing deep language models to brain activity, we first aim to identify the brain regions recruited during the reading of sentences. To this end, we (i) analyze the average fMRI and MEG responses to sentences across subjects and (ii) quantify the signal-to-noise ratio of these responses, at the single-trial single-voxel/sensor level.

As expected (Fedorenko et al., 2020; Dehaene & Cohen, 2011; Hagoort & Indefrey, 2014; Hickok & Poeppel, 2007), the average fMRI and MEG responses to words reveals a hierarchy of neural responses originating in V1 around 100 ms and continuing within the left posterior fusiform gyrus around 200 ms, the superior and middle temporal gyri, as well as the pre-motor and infero-frontal cortices between 150 and 500 ms after word onset (Supplementary Note 6.1.1, Figure 2.2a).

To quantify the proportion of these brain responses that depend on the specific content of sentences, we fit, for each subject separately, a shared response model across subjects (or noise-ceiling, see Methods, Supplementary Note 6.1.2, Supplementary Table S1, Figure 2.2b-d). We then assess the accuracy of this model with a Pearson R correlation (hereafter referred to as 'brain score' following (D. L. K. Yamins et al., 2014)) between the true and the predicted brain responses to held-out sentences, using a five-fold cross-validation. Finally, we assess the statistical significance of these brain scores with a two-sided Wilcoxon test across subjects, after testing for multiple comparison using False Discovery Rate (FDR) across voxels (see

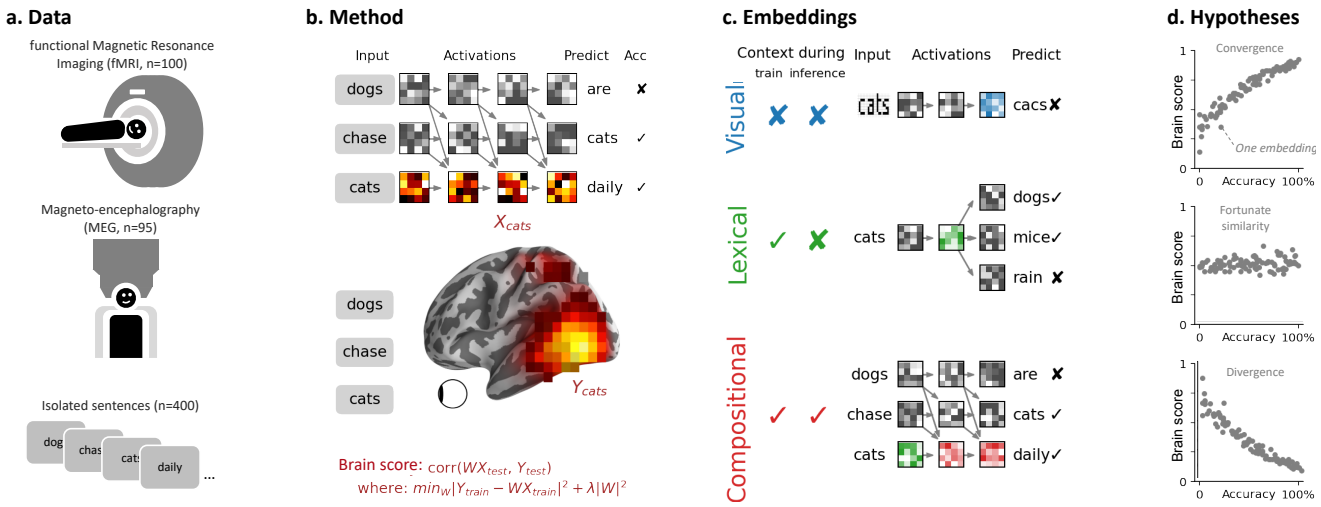


Figure 2.1: Approach. **a**. Subjects read isolated sentences while their brain activity was recorded with fMRI and MEG (Schoffelen et al., 2019). **b**. To compute the similarity between a deep language model and the brain, we (1) fit a linear regression W from the model's activations X to predict brain responses Y and (2) evaluate this mapping with a correlation between the predicted and true brain responses to held-out sentences Y_{test}. **c**. We consider different types of embedding depending on whether they vary with neighboring words during training and/or during inference. Visual embeddings refer, here, to the activations of a deep convolutional neural network trained on character recognition. Lexical embeddings refer, here, to the non-contextualized activations associated with a word independently of its context. Here, we use the word-embedding layer of language transformers (bottom green), as opposed to algorithms like Word2Vec (Mikolov, Sutskever, et al., 2013) (middle, green). Compositional embeddings refer, here, to the context-dependent activations of a deep language model (see Supplementary Note 6.1.4 for a discussion of our terminology). **d**. The three panels represent three hypotheses on the link between deep language models and the brain. Each dot represents one embedding. Algorithm are said to *converge* to brain-like computations if their performance (x-axis: i.e. accuracy at predicting a word from its previous context) indexes their ability to map onto brain responses to the same stimuli (i.e. y-axis: brain score). High-dimensional neural networks can, in principle, capture relevant information (Bingham & Mannila, 2001; Frankle & Carbin, 2019) and thus lead to a fortunate similarity with brain responses, and event a systematic divergence.

Methods). Our shared response model confirms that the brain network classically associated with language processing elicits representations specific to words and sentences (Mitchell et al., 2008; Fedorenko et al., 2016; Huth, de Heer, et al., 2016).

Deep language models reveal the hierarchical generation of language representations in the brain.

Where and when are the language representations of the brain similar to those of deep language models? To address this issue, we extract the activations (X) of a visual, a word and a com-

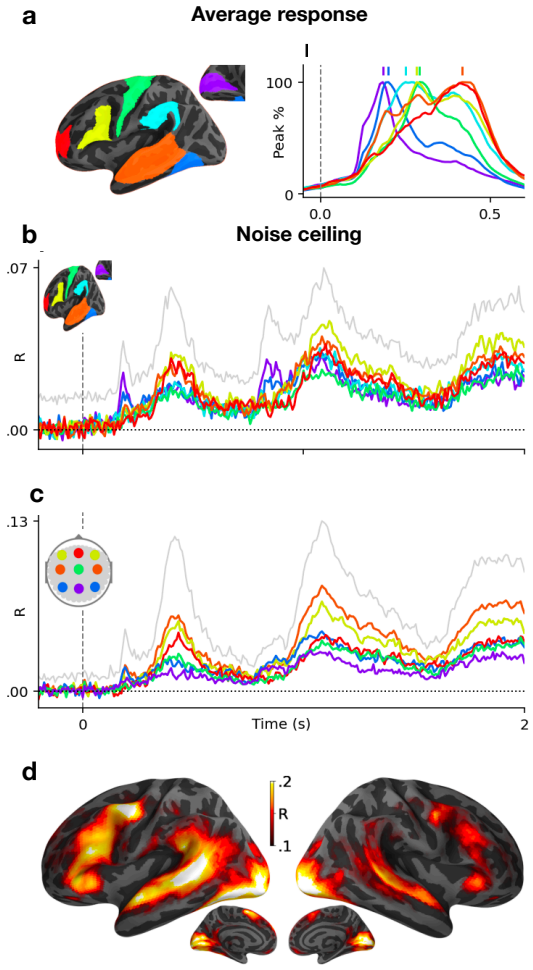


Figure 2.2: Average and shared response modeling (or noise ceiling). **a**. Grand average MEG source estimates to word onset (t=0 ms) for 7 regions typically associated with reading (V1: purple, M1: green, fusiform gyrus: dark blue, supramarginal gyrus: light blue, superior temporal gyrus: orange, infero-frontal gyrus: yellow and fronto-polar gyrus: red), normalized to their peak response. Vertical bars indicate the peak time of each region. **b**. MEG shared response model (or noise ceilings), approximated by predicting brain responses of a given subject from those of all other subjects. Colored lines depict the mean noise ceiling in each region of interest. The grey line depicts the best noise ceiling across sources. **c**. Same as (d) in sensor space. **d**. Share response model of fMRI recordings.

positional embedding (Figure 2.1d) and evaluate the extent to which each of them maps onto the brain responses (Y) to the same stimuli. To this end, we fit, for each subject independently, an ℓ_2-penalized regression (W) to predict single-sample fMRI and MEG responses for each voxel/sensor independently. We then assess the accuracy of this mapping with a brain-score similar to the one used to evaluate the shared response model.

Overall, the brain scores of these trained models are largely above chance (all $p < 10^{-9}$,

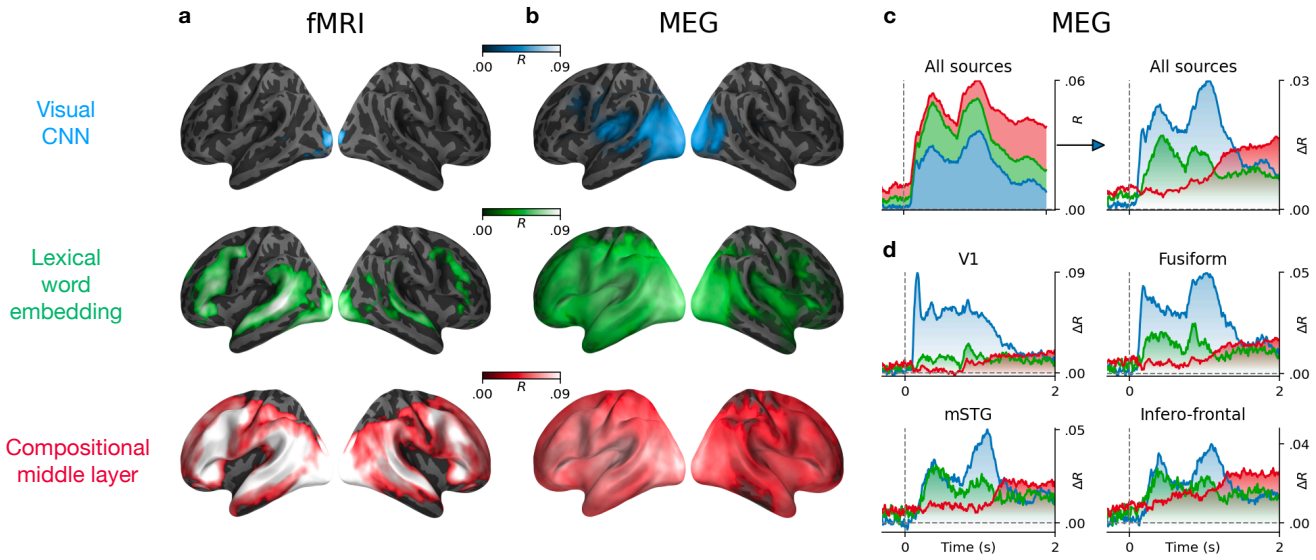


Figure 2.3: Brain-score comparison across embeddings. Lexical and compositional representations (see (Supplementary Note 6.1.4 for the definition of compositionality) can be isolated from (i) the word embedding layer (green) and (ii) one middle layer (red) of a typical language transformer (here, the ninth layer of a 12-layer causal transformer), respectively. We also report the brain scores of a convolutional neural network trained on visual character recognition (blue) to account for low-level visual representations. **a**. Mean (across subjects) fMRI scores obtained with the visual, word , and compositional embeddings. All colored regions display significant fMRI scores across subjects (n=100) after FDR correction. **b**. Mean MEG scores averaged across all time samples and subjects (n=95 subjects). **c**. Left: mean MEG scores averaged across all sensors. Right: mean MEG gains averaged across all sensors: *i.e.* the gain in MEG score of one level relative to the level below (blue: R[visual]; green: R[word] - R[visual]; red: R[compositional] - R[word]). **d**. Mean MEG gains in four regions of interest. For the raw scores (without subtraction), see Supplementary Figure S6. For the distribution of scores across channels and voxels, see Supplementary Figure S4.

Figure 2.4a and e). The modest correlation values are consistent with the high level of noise in single-sample single-voxel/channel neuroimaging data (Figure 2.2b-d). For example, fMRI and MEG scores reach R=.048 and R=.041, respectively, for the compositional embedding, which is close to and even exceeds our shared response model (fMRI: R=.060, MEG: R=.020, Figure 2.2).

In fMRI, the brain scores of the visual embedding peak in the early visual cortex (V1) (mean brain scores across voxels: $R = .022 \pm .003$, $p < 10^{-11}$). By contrast, the brain scores of lexical embedding peak in the left superior temporal gyrus ($R = .052 \pm .004$, $p < 10^{-13}$) as well as in the inferior temporal cortex and middle frontal gyrus ($R = .053 \pm .003$, $p < 10^{-15}$) and are significant across the entire language and reading network (Figure 2.3b). Finally, the brain scores of the compositional embedding are significantly higher than those of lexical of embeddings in the superior temporal gyrus ($\Delta R = .012 \pm .001$, $p < 10^{-16}$), the angular gyrus

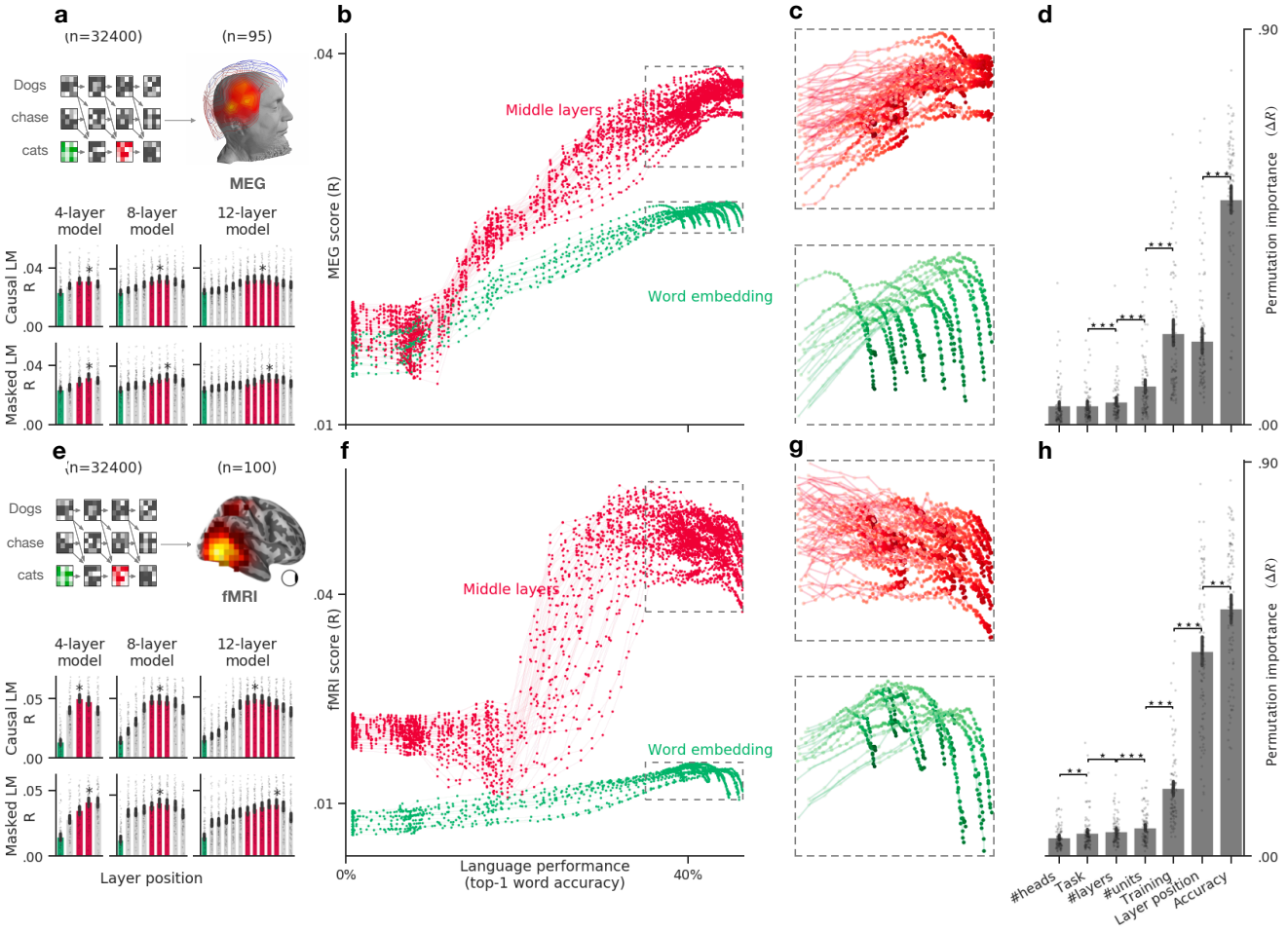


Figure 2.4: Language transformers tend to converge towards brain-like representations. a. Bar plots display the average MEG score (across time and channels) of six representative transformers varying in tasks (causal vs masked language modeling) and depth (4-12 layers). The green and red bars correspond to the word-embedding and middle layers, respectively. The star indicates the layer with the highest MEG score. **b**. Average MEG scores (across subjects, time, and channels) of each of the embeddings (dots) extracted from 18 *causal* architectures, separately for the input layer (word embedding, green) and the middle layers (red). **c**. Zoom of (b), focusing on the best neural networks (i.e. word-prediction accuracy ¿35 %). The results reveal a plateau and/or a divergence of the middle and input layers. **d**. Permutation importance quantifies the extent to which each property of the language transformers specifically contribute to making its embeddings more-or-less similar to brain activity (ΔR). All properties (training task. dimensionality *etc.*) significantly contribute to the brain scores ($\Delta R > 0$, all $p < 0.0001$ across subjects). Ordered pairwise comparisons of the permutation scores are marked with a star ('*': $p < .05$, '**': $p < .01$, '***': $p < .001$). **e-h**. Same as a-d, but evaluated on fMRI recordings. All error bars are the 95% confidence intervals across subjects (n=95 for MEG, n=100 for fMRI).

(ΔR =.010±.001, $p < 10^{-16}$), the infero-frontal cortex (ΔR =.016±.001, $p < 10^{-16}$) and the dorsolateral prefrontal cortex (ΔR =.012±.001, $p < 10^{-13}$). While these effects are lateralized

(left hemisphere versus right hemisphere: $\Delta R = .010 \pm .001$, $p < 10^{-14}$), they are significant across a remarkably large number of bilateral areas (Figure 2.3b). Lexical and compositional embeddings accurately predict brain responses in the early visual cortex. This result is not necessarily surprising: language embeddings encode features (e.g. position of words in the sentence, beginning/end of the sentence) that correlate with visual information (words are flashed at a screen, and the sentences are separated by pauses). Critically, the gain (ΔR) of these embeddings remain very small, suggesting that this effect is mainly driven by the covariance between low- and high-level representations of words.

Tracking the sequential generation of language representations over time and space.

To characterize the dynamics of these brain representations, we perform the same analysis using source-localized MEG recordings. The resulting brain scores are consistent with – although less spatially precise than – the above fMRI results (Figure 2.3c, average brain score between 0 and 2 s). For clarity, Figure 2.3d plot the gain in MEG scores: i.e. the difference of prediction performance between i) word and visual embeddings (green) and ii) the difference between compositional and word embedding (red). The brain scores of the visual embedding peak around 100 ms in V1 ($R = .008 \pm .002$, $p < 10^{-3}$), and rapidly propagate to higher-level areas (Figure 2.3D). The gain achieved by the word embedding can be observed in the left posterior fusiform gyrus around 200 ms and peaks around 400 ms and in the left temporal and frontal cortices. Finally, the gain achieved by the compositional embedding is observed in a large number of bilateral brain regions, and peaks around one second after word onset (Figure 2.3c and d).

After that period, brain areas outside the language network, such as area V1, appear to be better predicted by word and compositional embeddings than by visual ones (e.g between visual and word in V1: $\Delta R = .016 \pm .002$, $p < 10^{-10}$). These effects could thus reflect feedback activity (Seydell-Greenwald et al., 2020) and explain why the corresponding fMRI responses are better accounted for by word and compositional embeddings than by visual ones.

Together with Supplementary Figure S1, these results show with unprecedented spatio-temporal precision, that the brain-mapping of our three representative embeddings automatically recovers the hierarchy of visual, lexical, and compositional representations of language in each cortical region.

Compositional embeddings best predict brain responses.

What computational principle leads these deep language models to generate brain-like activations? To address this issue, we generalize the above analyses and evaluate the brain scores of 36 transformer architectures (varying from 4 to 12 layers, each ranging from 128 to 512 dimensions, and each benefiting from 4 to 8 attention heads), trained on the same Wikipedia dataset either with a causal language modeling (CLM) or a masked language modeling task (MLM). While causal language models are trained to predict a word from its previous context, masked language models are trained to predict a randomly masked word from its both left and right context.

Overall, we observe that the corresponding brain scores largely vary as a function of the relative depth of the embedding within the language transformer. Specifically, both MEG and fMRI scores follow an inverted U-shaped pattern across layers for all architectures (Figure 2.4a and e): the middle layers systematically outperform the output (fMRI: $\Delta R = .011 \pm .001$, $p < 10^{-18}$, MEG: $\Delta R = .003 \pm .0005$, $p < 10^{-13}$) and the input layers (fMRI: $\Delta R = .031 \pm .001$, $p < 10^{-18}$, MEG: $\Delta R = .009 \pm .001$, $p < 10^{-17}$). For simplicity, we refer to 'middle layers' as the layers $l \in [n_{\text{layers}}/2, 3n_{\text{layers}}/4]$ in Figure 2.4a and e. This result confirms that the intermediary representations of deep language transformers are more brain-like than those of the input and output layers (Toneva & Wehbe, 2019).

The emergence of brain-like representations predominantly depends on the algorithm's ability to predict missing words.

The above findings result from trained neural networks. However, recent studies suggest that random (i.e. untrained) networks can significantly map onto brain responses (A. J. E. Kell et al., 2018; Schrimpf et al., 2021; Millet & King, 2021). To test whether brain mapping specifically and systematically depends on the language proficiency of the model, we assess the brain scores of each of the 32 architectures trained with 100 distinct amounts of data. For each of these training steps, we compute the top-1 accuracy of the model at predicting masked or incoming words from their contexts. This analysis results in 32,400 embeddings, whose brain scores can be evaluated as a function of language performance, i.e. the ability to predict words from context (Figure 2.4 b and f).

We observe three main findings. First, random embeddings systematically lead to significant brain scores across subjects and architectures. The mean fMRI score across voxels is $R = .019 \pm .001$, $p < 10^{-16}$. The mean MEG score across channels and time sample is $R = .018$

$\pm$.0008, $p < 10^{-16}$. This result suggests that language transformers partially map onto brain responses independently of their language abilities.

Second, brain scores strongly correlate with language accuracy in both MEG (R =.77 Pearson's correlation on average $\pm$.01 across subjects) and fMRI (R =.57 $\pm$.02, Figure 2.4b and c). The correlation is highest for middle (fMRI: R =.81$\pm$.02; MEG: R=.86$\pm$.01) than input (fMRI: R=.39$\pm$.03; MEG: R=.73$\pm$.02) and output layers (fMRI: R =.63$\pm$.03; MEG:R=.78$\pm$.02). Beta coefficients for each particular layer and architecture are displayed in Supplementary Figure S1a and b. Furthermore, single-voxel analyses show that this correlation between brain score and language performance is driven mainly by the superior temporal sulcus and gyrus for the embedding layer (mean R=.52$\pm$.06) and is widespread for the middle layers, exceeding a correlation of R=.85 in the superior temporal sulcus, infero-frontal, fusiform and angular gyri (Supplementary Figure S1c). Overall, this result suggests that the better language models are at predicting words from context, the more their activations linearly map onto those of the brain.

Third, the highest brain scores are not achieved by the very best language transformers (Figure 2.4c and g). For instance, CLM transformers best map onto MEG (R=.039) and fMRI (R=.056) when they reach a language performance of 43% and 32%, respectively. By contrast, the very best transformers reach a language accuracy of 46%, but have significantly smaller brain scores (Figure 2.4c and g).

Architectural and training factors impact brain scores too.

Language performance co-varies with the amount of training as well as with several architectural variables. To disentangle the contribution of each of these variables to the brain scores, we perform a permutation feature importance analysis. Specifically, we train a Random Forest estimator (Breiman, 2001) to predict the average brain scores (across voxels or MEG sensors) of each subject independently, given the layer of the representation, the architectural properties (number of layers, dimensionality, attention head), task (CLM, MLM), amount of training (number of steps) and language performance (top-1 accuracy) of the transformer. Permutation feature importance then estimates the unique contribution of each feature in explaining the variability of brain scores across models (Pedregosa et al., 2011; Breiman, 2001). The results confirm that language performance is the most important factor that drives brain scores (Figure 2.4d-h). This factor supersedes other covarying factors such as the amount of training, and the relative position of the embedding with regard to the architecture ('layer position'): ΔR=.56$\pm$.01 for fMRI, ΔR=.51$\pm$.02 for MEG. Nevertheless, these other factors contribute significantly to the prediction of brain scores ($p < 10^{-16}$ across subjects for all variables).

Overall, these results show that the ability of deep language models to map onto the brain primarily depends on their ability to predict words from the context, and is best supported by the representations of their middle layers.

2.1.4 Discussion

Do deep language models and the human brain process sentences in the same way? Following a recent methodology (D. L. K. Yamins et al., 2014; Tang et al., 2018; Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte et al., 2008; Güçlü & Gerven, 2015; Eickenberg et al., 2017; D. L. K. Yamins & DiCarlo, 2016; Saxe et al., 2021; A. J. E. Kell et al., 2018; Huth, de Heer, et al., 2016; Wehbe et al., 2014), we address this issue by evaluating whether the activations of a large variety of deep language models linearly map onto those of 102 human brains. Our study provides two main contributions.

First, our work complements previous studies (Jain & Huth, 2018; Athanasiou et al., 2018; Toneva & Wehbe, 2019; Wehbe et al., 2014; Heilbron et al., 2022; Goldstein et al., 2022; Schrimpf et al., 2021) and confirms that the activations of deep language models significantly map onto the brain responses to written sentences (Figure 2.3). This mapping peaks in a distributed and bilateral brain network (Figure 2.3a and b) and is best estimated by the middle layers of language transformers (Figure 2.4a and e). The notion of representation underlying this mapping is formally defined as linearly-readable information. This operational definition helps identify brain responses that any neuron can differentiate – as opposed to entangled information which would necessitate several layers before being usable (Minsky & Papert, 1969; Cadieu et al., 2014; Kriegeskorte et al., 2008; King & Dehaene, 2014; U. Cohen et al., 2020).

Furthermore, the comparison between visual, lexical, and compositional embeddings precise the nature and dynamics of these cortical representations. In particular, our results shows with unprecedented spatio-temporal precision that early visual responses (¡150 ms) are quasi-entirely accounted for by visual embeddings, and then transmitted to the posterior fusiform gyrus, which switches from visual to lexical representations around 200 ms (Movie 2). This finding strengthens the claim that this area is responsible for orthographic and morphemic computations (Dehaene & Cohen, 2011; Hermes et al., 2017; Woolnough et al., 2020). Then, around 400 ms, word embeddings predict a large fronto-temporo-parietal network which peaks in the left temporal gyrus; these word representations are then maintained for several seconds (Mitchell et al., 2008; Jain & Huth, 2018; Toneva & Wehbe, 2019; Sassenhagen & Fiebach, 2019).

This result not only confirms the wide spread distribution of meaning in the brain (Huth, de Heer, et al., 2016; Price, 2010), but also reveals its remarkably long-lasting nature.

Finally, compositional embeddings peak in the brain regions associated with high-level language processing such as the infero-frontal and the anterior temporal cortices as well as the superior temporal cortex and the temporal-parietal junction (Pallier et al., 2011; Hickok & Poeppel, 2007; J. R. Brennan & Pylkkänen, 2017). We confirm that these left-lateralized representations are significant in both hemispheres (Fedorenko et al., 2010; Cogan et al., 2014). Critically, MEG suggests that these compositional effects become dominant and clearly bilateral long after word onset (>800 ms). We speculate that this surprisingly late responses may be due to the complexity of the sentences used in the present study, which may slow down compositional computations.

At this stage, however, these three levels representations remain coarsely defined. Further inspection of artificial (Manning et al., 2020; Lakretz et al., 2019) and biological networks (Caucheteux et al., 2021a; Reddy & Wehbe, 2020; Hale et al., 2021) remains necessary to further decompose them into interpretable features. In particular, it will be important to test whether the converging representations presently identified solely correspond to well-known linguistics phenomena as our supplementary analyses suggest (Supplementary Figure S2 and Supplementary Note 6.1.3), or, on the contrary, whether they correspond to unknown language structures.

Second, our study shows that the similarity between deep language models and the brain primarily depends on their ability to predict words from their context. Specifically, we show that language performance is the most contributing factor explaining the variability of brain scores across embeddings (Figure 2.4d and h). Analogous results have been reported in both vision and audition research, where best deep learning models tend to best map onto brain responses (D. L. K. Yamins et al., 2014; D. L. K. Yamins & DiCarlo, 2016; Schrimpf et al., 2018; A. J. E. Kell et al., 2018; Schrimpf et al., 2021). In addition, our results are consistent with the findings of Schrimpf et al. (Schrimpf et al., 2021) reported simultaneously to ours. Together, these results suggest that deep learning algorithms converge – at least partially – to brain-like representations during their training. This result is not trivial: the representations that are optimal to predict masked or future words from large amounts of text could have been very distinct from those the brain learns to generate.

The mapping between deep language models and brain recordings reaches very low correlation values. This phenomenon is expected: i) neuroimaging is notoriously noisy and ii) we analyze and model here single-sample responses of single-voxel/sensor. However, the resulting brain scores are i) highly significant (all $p < 10^{-9}$ on average across both all fMRI voxels and MEG sensors), including when compared to a permutation baseline (Supplementary Figure S3), and ii) in the same order of magnitude than a baseline shared-response model (or noise ceiling, Figure 2.2) as well as previous reports (e.g. (Huth, de Heer, et al., 2016), before correcting for the noise ceiling). Besides, we generally report brain scores averaged across all voxels or MEG channels, even though many brain areas do not strongly respond to language (Figure 2.2). Critically, the link between brain scores and language performance is strong: the correlation between the language performance and brain scores is above R= .90 for MEG and R= .80 for fMRI (Supplementary Figure S1). Nevertheless, it is clear that improving the the signal-to-noise ratio, for instance by using increasingly large datasets (Caucheteux et al., 2021b; Millet & King, 2021; Nastase et al., 2020; Caucheteux et al., 2022) will be critical to precisely characterize the nature of brain representations.

Permutation feature importance shows that several factors such as the amount of training and the architecture significantly impact brain scores. This finding contributes to a growing list of variables that lead deep language models to behave more-or-less similarly to the brain. For example, Hale et al. (Hale et al., 2018) showed that the amount and the type of corpus impact the ability of deep language parsers to linearly correlate with EEG responses. The present work complements this finding by evaluating the full set of activations of deep language models. It further demonstrates that the key ingredient to make a model more brain-like is, for now, to improve its language performance.

The conclusion that deep networks converge towards brain-like representations should be qualified: we show that the brain scores of the very best models tend to ultimately decrease with language performance, especially in fMRI (Figure 2.4g). We speculate that this phenomenon (also observed in vision (Schrimpf et al., 2018)) may rise because transformers overfit an inappropriate objective. Specifically, while there is growing evidence that the human brain does predict words from context (Keller & Mrsic-Flogel, 2018; Heilbron et al., 2022; Goldstein et al., 2022), this learning rule may not fully account for the complex (and potentially various) tasks performed by the brain (e.g. long-range (L. Wang, 2021; Lee et al., 2021) and hierarchical predictions (K. J. Friston & Stephan, 2007)).

This discrepancy adds to the long-list of differences between deep language models and the brain: whereas the brain is trained (i) with a recurrent architecture and (ii) on a relatively small amount of grounded sentences, transformers are trained (i) with a massively feedforward architecture and (ii) on huge text databases (Brown et al., 2020) (note that, given large-enough spaces, feedforward transformers may actually implement computations similar to recurrent networks (Ramsauer et al., 2021)). Consequently, while the similarity between deep networks and the brain provide a stepping stone to unravel the foundation of natural language processing, identifying the remaining differences between these two systems remains, by far, the major challenge to build algorithms that learn and think like humans (Brown et al., 2020; Baroni, 2020; B. M. Lake et al., 2016; Zellers et al., 2019).

2.1.5 Methods

Deep language transformers

To model word and sentence representations, we trained a variety of transformers (Vaswani et al., 2017), and input them with the same sentences that the subject read. Transformers consist of multiple contextual transformer layers stacked onto one non-contextualized word embedding layer (a look-up table). Following the standard implementation (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2019), the word embedding layer is trained simultaneously with the contextual layers: the weights of the word embedding vary with the training, and so do their activations in response to fixed inputs. Thus, one representation can be extracted from each (contextual or non-contextual) layer. We always extract activation in a causal way: for example, given the sentence 'THE CAT IS ON THE MAT', the brain response to 'ON' would be solely compared to the activations of the transformer input with 'THE CAT IS ON', and extracted from the 'ON' contextualized embeddings. Word embeddings and contextualized embeddings were generated for every word, by generating word sequences from the three previous sentences. We did not observe qualitatively different results when using shorter or longer contexts. It is to be noted that the sentences were isolated, and were not part of a narrative.

In total, we investigated 32 distinct architectures varying in their dimensionality ($\in [128, 256, 512]$), number of layers ($\in [4, 8, 12]$), attention heads ($\in [4, 8]$), and training task (causal language modeling and masked language modeling). While causal language transformers are trained to predict a word from its previous context, masked language transformers predict randomly masked words from a surrounding context. We froze the networks at ≈ 100 training stages (log distributed between 0 and 4,5M gradient updates, which corresponds to ≈ 35 passes over

the full corpus), resulting in 3,600 networks in total, and 32,400 word representations (one per layer). The training was early-stopped when the networks' performance did not improve after 5 epochs on a validation set. Therefore, the number of frozen steps varied between 96 and 103 depending on the training length.

The algorithms were trained using XLM implementation (Lample & Conneau, 2019). No hyper-parameter tuning was performed. Following (Lample & Conneau, 2019), each algorithm was trained each on 8 GPUs using early stopping with training perplexity criteria, 16 streams per batch, 128 words per stream, epoch size of 200 000 streams, 0.1 dropout, 0.1 attention dropout, gelu activation, inverse (sqrt) adam optimizer with learning rate 0.0001, 0.01 weight decay, on the same Wikipedia corpus of 278,386,651 words (in Dutch) extracted using WikiExtractor (Attardi, 2015) and pre-processed using Moses tokenizer (Koehn et al., 2007), with punctuation. We restricted the vocabulary to the 50,000 most frequent words, concatenated with all words used in the study (50,341 vocabulary words in total). These design choices enforce that the difference in brain scores observed across models cannot be explained by differences in corpora and text preprocessing.

To evaluate the language processing performance of the networks, we computed their performance (top-1 accuracy on word prediction given the context) using a test dataset of 180,883 words from Dutch Wikipedia. The list of architectures and their final performance at next-word prerdiction is provided in Supplementary Table S2.

For clarity, we dissociate:

- The architectures (e.g one transformer with 12 layers): there are 36 transformer architectures here (18 CLM and 18 MLM).
- The models: one architecture, frozen at one particular learning step. Since we use 100 learning steps, there are 36 x 100 = 3,600 networks here.
- The embeddings: one word representation extracted from a network, at one particular layer. Since the number of layers varies with the architecture (twelve networks with 5, twelve networks with 9 and twelve networks with 13 twelve layers, including the non contextualized word embedding), there are 12 x (5 + 9 + 13) = 324 representations per step, so 324 x 100 = 3,400 word embeddings in total.

Visual convolutional neural network

To model visual representations, every word presented to the subjects was rendered on a gray 100 x 32 pixel background with a centered black Arial font, and input to a VGG network pretrained to recognize words from images (Baek et al., 2019), resulting in an 888-dimensional embedding. Specifically, this model was trained on real pictures of single words taken in naturalistic settings (e.g. ad, banner).

This embedding was used to replicate and extend previous work on the similarity between visual neural network activations and brain responses to the same images (e.g. (D. L. K. Yamins et al., 2014; Kriegeskorte et al., 2008; Güçlü & Gerven, 2015)).

Neuroimaging protocol

For all the analyses, we used the open-source dataset released by Schoffelen and colleagues (Schoffelen et al., 2019), gathering the functional magnetic resonance imaging (fMRI) and magneto-encephalography (MEG) recordings of 204 native Dutch speakers (100 males), aged from 18 to 33 years. Here, we focused on the 102 right-handed speakers who performed a reading task while being recorded by a CTF magneto-encephalography (MEG) and, in a separate session, with a SIEMENS Trio 3T Magnetic Resonance scanner (Schoffelen et al., 2019).

Words (in Dutch) were flashed one at a time with a mean duration of 351 ms (ranging from 300 to 1400 ms), separated with a 300 ms blank screen, and grouped into sequences of 9 - 15 words, for a total of approximately 2,700 words per subject. Sequences were separated by a 5 s-long blank screen. We restricted our study to meaningful sentences (400 distinct sentences in total, 120 per subject). The exact syntactic structures of sentences varied across all sentences. Roughly, sentences were either composed of a main clause and a simple subordinate clause, or contained a relative clause. Twenty percent of the sentences were followed by a yes/no question (e.g. "Did grandma give a cookie to the girl?") to ensure that subjects were paying attention. Questions were not included in the dataset, and thus excluded from our analyses. Sentences were grouped into blocks of five sequences. This grouping was used for cross-validation to avoid information leakage between the train and test sets.

Magnetic Resonance Imaging (MRI)

Structural images were acquired with a T1-weighted magnetization-prepared rapid gradient-echo (MP-RAGE) pulse sequence. The full acquisition details, available in (Schoffelen et al., 2019), are summarized here simplicity: TR=2,300 ms, TE=3.03 ms, 8 degree flip-angle, 1

slab, slice-matrix size=256×256, slice thickness=1 mm, field of view=256 mm, isotropic voxel-size=1.0×1.0×1.0 mm. Structural images were defaced by Schoffelen and colleagues. Preprocessing of the structural MRI was performed with Freesurfer (Fischl, 2012), using the `recon-all` pipeline and a manual inspection of the cortical segmentations, realigned to 'fsaverage'. Region-of-interest analyses were selected from the PALS Brodmann's Area atlas (Van Essen, 2005) and the Destrieux atlas (Destrieux et al., 2010).

Functional images were acquired with a T2*-weighted functional echo-planar blood oxygenation level-dependent (EPI-BOLD) sequence. The full acquisition details, available in (Schoffelen et al., 2019), are summarized here for simplicity: TR=2.0 seconds, TE=35ms, flip angle=90 degrees, anisotropic voxel size=3.5×3.5×3.0 mm extracted from 29 oblique slices. fMRI was preprocessed with fMRIPrep with default parameters (Esteban et al., 2019). The resulting BOLD times series were detrended and de-confounded from 18 variables (the 6 estimated head-motion parameters ($\texttt{trans}_{x,y,z}$, $\texttt{rot}_{x,y,z}$) and the first 6 noise components calculated using anatomical CompCorr (Behzadi et al., 2007) and 6 DCT-basis regressors using nilearn's `clean_img` pipeline and otherwise default parameters (Abraham et al., 2014). The resulting volumetric data lying along a 3mm line orthogonal to the mid-thickness surface were linearly projected to the corresponding vertices. The resulting surface projections were spatially decimated by 10, and are hereafter referred to as voxels, for simplicity. Finally, each group of 5 sentences was separately and linearly detrended. It is noteworthy that our cross-validation never splits such groups of five consecutive sentences between the train and test sets. Two subjects were excluded from the fMRI analyses because of difficulties in processing the metadata, resulting in 100 fMRI subjects.

Magneto-encephalography (MEG)

The MEG time series were preprocessed using MNE-Python and its default parameters except when specified (Gramfort et al., 2013). Signals were band-passed filtered between 0.1 and 40 Hz filtered, spatially corrected with a Maxwell Filter, clipped between the 0.01^{st} and 99.99^{th} percentiles, segmented between -500 ms to +2,000 ms relative to word onset and baseline-corrected before t=0. Reference channels and non-MEG channels were excluded from subsequent analyses, leading to 273 MEG channels per subject. We manually co-referenced (i) the skull segmentation of subjects' anatomical MRI with (ii) the head markers digitized before MEG acquisition. A single-layer forward model was generated with the Freesurfer-wrapper implemented in MNE-Python (Gramfort et al., 2013). Due to the lack of empty-room recordings, the noise covariance matrix used for the inverse operator was estimated from the zero-centered 200 ms of baseline

MEG activity preceding word onset. Subjects' source space inverse operators were computed using a dSPRM. The average brain responses displayed in Figure 2.1d were computed as the square of the average evoked related field across all words for each subject separately, averaged across subjects, and finally divided by their respective maxima, to highlight temporal differences. Seven subjects were excluded from the MEG analyses because of difficulties in processing the metadata, resulting in 92 usable MEG recordings.

Shared response model: Brain → Brain mapping

To estimate the amount of explainable signal in each MEG and fMRI recording, we trained and evaluated, through cross-validation, a linear mapping model W to predict the brain responses of a given subject to each sentence Y from the aggregated brain responses of all other subjects who read the same sentence X. Specifically, five cross-validation splits were implemented across 5-sentence blocks with scikit-learn GroupKFold (Pedregosa et al., 2011). For each word of each sentence i, all but one subject who read the corresponding sentence were averaged with one another to form a template brain response: $x_i \in \mathbb{R}^n$ with n the number of MEG channels or fMRI voxels, as well as a target brain response $y_i \in \mathbb{R}^n$ corresponding to the remaining subject. X and Y were normalized (mean=0, std=1) across sentences for each spatio-temporal dimension, using a robust scaler clipping below and above the 0.01^{st} and 99.99^{th} percentiles, respectively. A linear mapping $W \in \mathbb{R}^{n \times n}$ was then fit with a ridge regression to best predict Y from X on the train set:

$$W = (X_{\text{train}}^T X_{\text{train}} + \lambda I)^{-1} X_{\text{train}}^T Y_{\text{train}} \tag{2.1}$$

with λ the $l2$ regularization parameter, chosen amongst 20 values log-spaced between 10^{-3} and 10^8 with nested leave-one-out cross-validation for each dimension separately (as implemented in (Pedregosa et al., 2011)). Brain predictions $\hat{Y} = WX$ were evaluated with a Pearson correlation on the test set:

$$R = \text{Corr}(Y_{\text{test}}, \hat{Y}_{\text{test}}) \tag{2.2}$$

For the MEG source noise estimate, the correlation was also performed after source projection:

$$R = \text{Corr}(KY_{\text{test}}, K\hat{Y}_{\text{test}}) \tag{2.3}$$

with $K \in \mathbb{R}^{n \times m}$ the inverse operator projecting the n MEG sensors onto m sources. Correlation scores were finally averaged across cross-validation splits for each subject, resulting in one correlation score ('brain score') per voxel (or per MEG sensor/time sample) per subject.

Brain score and similarity: Network → Brain mapping

To estimate the functional similarity between each artificial neural network and each brain, we followed the same analytical pipeline used for noise ceiling, but replaced X with the activations of the deep learning models. Specifically, using the same cross-validation, and for each subject separately, we trained a linear mapping $W \in \mathbb{R}^{o,n}$ with o the number of activations, to predict brain responses Y from the network activations X. X was normalized across words (mean=0, std=1).

To account for the hemodynamic delay between word onset and the BOLD response recorded in fMRI, we used a finite impulse response (FIR) model with five delays (from 2 to 10 seconds) to build X^* from X. W was found using the same ridge regression described above, and evaluated with the same correlation scoring procedure. The resulting brain correlation scores measure the linear relationship between the brain signals of one subject (measured either by MEG or fMRI) and the activations of one artificial neural network (e.g a word embedding). For MEG, we simply fit and evaluated the model activations X at each time sample independently.

In principle, one may orthogonalize low-level representations (e.g. visual features) from high-level network models (e.g. language model), to separate the specific contribution of each type of model. This is because middle layers have access to the word-embedding layer, and can, in principle, simply copy some of its activations. Similarly, word embedding can implicitly contain visual information: e.g. frequent words tend to be visually smaller than rare ones. In our case, however, the middle layers of transformers were much better than word embeddings, which were much better than visual embeddings. To quantify the gain ΔR achieved by a higher-level model M_1 (e.g. the middle layers of a transformer) and a lower level model M_2 (e.g. a word embedding) we thus simply compared the difference of their encoding scores:

$$\Delta R_{M_1} = R_{M_1} - R_{M_2} \tag{2.4}$$

Results are consistent when using different orthogonalization methods (Supplementary Figure S5).

Convergence analysis

All neural networks but the visual CNN were trained from scratch on the same corpus (as detailed in the first Methods section). We systematically computed the brain scores of their activations on each subject, sensor (and time sample in the case of MEG) independently. For

computational reasons, we restricted model comparison on MEG encoding scores to ten time samples regularly distributed between $[0, 2]$s. Brain scores were then averaged across spatial dimensions (i.e. MEG channels or fMRI surface voxels), time samples, and subjects to obtain the results in Figure 2.4. To evaluate the convergence of a model, we computed, for each subject separately, the correlation between (1) the average brain score of each network and (2) its performance or its training step (Figure 2.4 and Supplementary Figure S1). Positive and negative correlations indicate convergence and divergence, respectively. Brain scores above 0 before training indicate a fortuitous relationship between the activations of the brain and those of the networks.

Permutation feature importance

To systematically quantify how the architecture, language accuracy, and training of the language transformers impacted their ability to linearly map onto brain activity, we fitted, for each subject separately, a Random Forest across the models' properties to predict their brain scores, using scikit-learn's `RandomForest` (Breiman, 2001; Pedregosa et al., 2011). Specifically, we input the following features to the random forest: the training task (causal language modeling "CLM" vs. masked language modeling "MLM"), the number of attention heads $\in [4, 8]$, the total number of layers $\in [4, 8, 12]$, dimensionality $\in [128, 256, 512]$, training step (number of gradient updates, $\in [0, 4.5M]$), language modeling accuracy (top-1 accuracy at predicting a masked word) and the relative position of the representation (a.k.a 'layer position', between 0 for the word-embedding layer, and 1 for the last layer). The performance of the Random Forest was evaluated for each subject separately with a Pearson correlation R using five-split cross-validation across models.

"Permutation feature importance" summarizes how each of the covarying properties of the models (their task, architecture, etc.) specifically impacts the brain scores (Breiman, 2001). Permutation feature importance was implemented with scikit-learn (Pedregosa et al., 2011) and is summarized with ΔR: the decrease in R when shuffling one feature (using 50 repetitions). For each subject, we reported the average decrease across the cross-validation splits (Figure 2.4). The resulting scores (ΔR) are expected to be centered around 0 if the corresponding feature does not impact the brain scores , and positive otherwise.

Statistics and Reproducibility

To estimate the robustness of our results, we systematically performed second-level analyses across subjects. Specifically, we applied Wilcoxon signed-rank tests across subjects' estimates to

evaluate whether the effect under consideration was systematically different from the chance level. The p-values of individual voxel/source/time samples were corrected for multiple comparisons, using a False Discovery Rate (Benjamini/Hochberg) as implemented in MNE-Python (Gramfort et al., 2013) (we use the default parameters). Error bars and $\pm$ refer to the standard error of the mean (SEM) interval across subjects.

Brain parcellation

In Figure 2.3, we focus on particular regions of interest using the Brodmann's areas from the PALS parcellation of freesurfer (Fischl, 2012). The superior temporal gyrus (BA22) is split into its anterior, middle and posterior parts to increase granularity. For clarity, we rename certain areas as specified in Table 2.1.

Label	Corresponding Brodmann's areas
V1	BA17
Fusiform	BA37
Angular	BA39
aSTG	BA22-anterior
mSTG	BA22-middle
pSTG	BA22-posterior
Supramarginal	BA40
Infero-frontal	BA44 / BA45 / BA47
Fronto-polar	BA10
Temporo-polar	BA38

Table 2.1: Brain parcellation. Taxonomy used to label the regions of interest in the brain following the PALS Brodmann's Area atlas (Van Essen, 2005)

Ethics

These data were provided (in part) by the Donders Institute for Brain, Cognition, and Behaviour after having been approved by the local ethics committee (CMO – the local "Committee on Research Involving Human Subjects" in the Arnhem-Nijmegen region). As stated in the original paper (Schoffelen et al., 2019), "In the informed consent procedure, [the subjects] explicitly consented for the anonymized collected data to be used for research purposes by other researchers. [..] The study was approved by the local ethics committee (CMO – the local

"Committee on Research Involving Human Subjects" in the Arnhem-Nijmegen region) and followed guidelines of the Helsinki declaration."

2.2 Deep language algorithms predict semantic comprehension from brain activity

2.2.1 Abstract

Deep language algorithms, like GPT-2, have demonstrated remarkable abilities to process text, and now constitute the backbone of automatic translation, summarization and dialogue. However, whether these models encode information that relates to human comprehension still remains controversial. Here, we show that the representations of GPT-2 not only map onto the brain responses to spoken stories, but they also predict the extent to which subjects understand the corresponding narratives. To this end, we analyze 101 subjects recorded with functional Magnetic Resonance Imaging while listening to 70 min of short stories. We then fit a linear mapping model to predict brain activity from GPT-2's activations. Finally, we show that this mapping reliably correlates ($\mathcal{R} = 0.50, p < 10^{-15}$) with subjects' comprehension scores as assessed for each story. This effect peaks in the angular, medial temporal and supramarginal gyri, and is best accounted for by the long-distance dependencies generated in the deep layers of GPT-2. Overall, this study shows how deep language models help clarify the brain computations underlying language comprehension.

2.2.2 Introduction

In less than two years, language transformers like GPT-2 have revolutionized the field of natural language processing (NLP). These deep learning architectures are typically trained on very large corpora to complete partially-masked texts, and provide a one-fit-all solution to translation, summarization, and question-answering tasks (Radford et al., 2019; Devlin et al., 2019; Yang et al., 2020). These advances raise a major question: do these algorithms process language like the human brain? Recent studies suggest that they partially do: the hidden representations of various deep neural networks have shown to linearly predict single-sample fMRI (Caucheteux et al., 2021b; Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022; Caucheteux et al., 2021a; Hale et al., 2021; Anderson et al., 2021; Sun et al., 2021), MEG (Toneva & Wehbe, 2019; Caucheteux & King, 2022), and intracranial responses to spoken and written texts (Goldstein et al., 2022; Schrimpf et al., 2021).

However, whether these models encode, retrieve and pay attention to information that specifically relates to behavior in general, and to comprehension in particular remains controversial (Nie et al., 2020; Lakretz et al., 2021; Hupkes et al., 2020; B. M. Lake & Murphy,

2021; Linzen & Baroni, 2021; McClelland et al., 2020; Marcus, 2020a). This issue is all-the-more relevant that the behavior of deep language models remains challenged by complex questions, including subject-verb agreement (Lakretz et al., 2021; Linzen & Baroni, 2021; Hupkes et al., 2020), causal reasoning (Marcus, 2020a; B. M. Lake & Murphy, 2021), story generation, text summarization as well as dialogue and question answering (Holtzman et al., 2020; Wiseman et al., 2017; Thakur et al., 2021; Raffel et al., 2020; Krishna et al., 2021).

To explore the relationship between comprehension and the representations of GPT-2, we compare GPT-2's activations to the functional Magnetic Resonance Imaging of 101 subjects listening to 70min of seven short stories. We first quantify this similarity with a "brain score" (M) (D. L. K. Yamins et al., 2014; Huth, de Heer, et al., 2016). We then evaluate how brain scores systematically vary with — and thus predict — semantic comprehension, as individually assessed by a questionnaire at the end of each story. Finally, by decomposing and manipulating GPT-2's processes, we identify (1) the brain regions, (2) the levels of representations (phonological, lexical, compositional), and (3) the attentional gating that specifically relates to this prediction.

The alignment identified between behavior, brain activations and the representations of GPT-2 suggest that comprehension relies on a specific computational hierarchy, whereby the auditory cortices integrate information over short time windows, and the fronto-parietal areas combine supra-lexical information over long time windows.

2.2.3 Results

GPT-2's activations linearly map onto fMRI responses to spoken narratives. To assess whether GPT-2 generates similar representations to those of the brain, we analyze the Narratives dataset: 101 subjects listening to seven short stories while their brain activity is recorded with fMRI. Note that subjects do not necessarily listen to the same stories (Figure 2). First, we evaluate, for each voxel, subject and narrative independently, whether the fMRI responses can be predicted from a linear combination of GPT-2's activations (Figure 2.5A). We summarize the precision of this mapping with a brain score $\mathcal{M}$: i.e. the correlation between the true fMRI responses and the fMRI responses linearly predicted, with cross-validation, from GPT-2's responses to the same narratives (cf. Methods).

To mitigate the spatial resolution of fMRI and the necessity to correct voxel analyses for multiple comparisons, we here report either 1) the average brain scores across voxels or 2) the average score within each region of interest ($n = 314$, following an automatic subdivision of the Destrieux atlas (Destrieux et al., 2010), cf. Supplementary Note 6.2.1), and correct statistical tests

for multiple comparisons across the brain regions. Consistent with previous findings (Toneva & Wehbe, 2019; Jain & Huth, 2018; Caucheteux & King, 2022; Schrimpf et al., 2018), these brain scores are significant over a distributed and bilateral cortical network, and peak in middle- and superior-temporal gyri and sulci, as well as in the supra-marginal and the infero-frontal cortex (Toneva & Wehbe, 2019; Jain & Huth, 2018; Caucheteux & King, 2022) (Figure 2.5B).

By separately analyzing the activations of each layer of GPT-2, we confirm that middle layers best map onto the brain (Figure 2.5C), as previously reported (Jain & Huth, 2018; Toneva & Wehbe, 2019; Caucheteux & King, 2022). For clarity, the following analyses focus on the activations extracted from the eighth layer, i.e. the layer with the highest brain score on average across voxels (Figure 2.5C). However, the results generalize to other contextual layers of GPT-2 (Supplementary Note 6.2.5, Supplementary Figure S10).

The brain predictions of GPT-2 correlate with semantic comprehension. Does the linear mapping between GPT-2 and the brain reflect a fortunate correspondence (Caucheteux & King, 2022)? Or, on the contrary, does it reflect similar representations of high-level semantics (Caucheteux et al., 2021a)? To address this issue, we correlate these brain scores to the level of comprehension of the subjects, assessed for each subject-story pair. On average across all voxels, this correlation reaches $\mathcal{R} = 0.50$ ($p < 10^{-15}$, Figure 2.5D, as assessed across subject-story pairs with the Pearson's test provided by SciPy (Virtanen et al., 2020)). This correlation is significant across a wide variety of the bilateral temporal, parietal and prefrontal cortices typically linked to language processing (Figure 2.5E). Together, these results suggest that the shared representations between GPT-2 and the brain reliably vary with semantic comprehension.

Low-level processing only partially accounts for the correlation between comprehension and GPT-2's mapping Low-level speech representations typically vary with attention (Mesgarani & Chang, 2012; L. Cohen et al., 2021), and could thus, in turn, influence down-stream comprehension processes. Consequently, one can legitimately wonder whether the correlation between comprehension and GPT-2's brain mapping is simply driven by variations in low-level auditory processing. To address this issue, we evaluate the predictability of fMRI given low-level phonological features: the word rate, phoneme rate, phonemes, stress and tone of the narrative (cf. Methods). The corresponding brain scores correlate with the subjects' understanding ($\mathcal{R} = 0.17, p < 10^{-2}$) but considerably less than the brain scores of GPT-2 ($\Delta\mathcal{R} = 0.32$). These low-level correlations with comprehension peak in the left superior temporal cortex

(Figure 2.5F). Overall, this result suggests that the link between comprehension and GPT-2's brain mapping may be partially explained by – but not reduced to – the variations of low-level auditory processing.

The reliability of high-level representations best predict comprehension Is the correlation between comprehension and GPT-2's mapping driven by a *lexical* process and/or by an ability to meaningfully combine words? To tackle this issue, we compare the correlations obtained from GPT-2's word embedding (i.e. layer 0) to those obtained from GPT-2's eighth layer, i.e. a contextual embedding. On average across voxels, the correlation with comprehension is 0.12 lower with GPT-2's word embedding than with its contextual embedding. An analogous analysis, comparing word embedding to phonological features is displayed in Figure 2.5F. Strictly lexical effects (word-embedding *versus* phonological) peak in the superior-temporal lobe and in pars triangularis. By contrast, higher-level effects (GPT-2 eighth layer *versus* word-embedding) peak in the superior-frontal, posterior superior-temporal gyrus, in the precuneus and in both the triangular and opercular parts of the inferior frontal gyrus – a network typically associated with high-level language comprehension (Lerner et al., 2011; Pallier et al., 2011; Fedorenko et al., 2016; Friederici, 2011; Hickok & Poeppel, 2007; Caucheteux & King, 2022). Together, these model comparisons suggest that GPT-2 best predicts how brain responses to speech vary with comprehension.

Comprehension effects are mainly driven by individuals' variability The variability in comprehension scores could result from exogeneous factors (e.g. some stories may be harder to comprehend than others for GPT-2) and/or from endogeneous factors (e.g. some subjects may better understand specific texts because of prior knowledge). To address this issue, we fit a linear mixed model to predict comprehension scores given brain scores, specifying the narrative as a random effect (cf. Supplementary Note 6.2.2). The fixed effect of brain score (shared across narratives) is highly significant: $\beta = 0.04, p < 10^{-29}$, cf. Supplementary Note 6.2.2). However, the random effect (slope specific to each single narrative) is not ($\beta < 10^{-2}$, $p > 0.11$). We also replicate the main analysis (Figure 2.5D) within each single narrative: the correlation with comprehension reaches 0.76 for the 'Sherlock' story and is above 0.40 for every story (cf. Supplementary Note 6.2.3). Overall, these analyses confirm that the link between GPT-2 and semantic comprehension is best accounted for by an endogeneous factor: i.e. individual differences in comprehension scores.

Decomposing the brain regions, levels of representation and attention distances underlying comprehension Can GPT-2 be further decomposed to identify the mechanisms responsible for generating representations that both (i) map with the human brain and (ii) predict subjects' comprehension? To address this issue, we investigate the links between (1) short- and long-range attentional gating, (2) the depth of the representation and (3) brain and comprehension scores. Specifically, we compute both of these scores for different layer k when restricting their attention span to different distances d (i.e. layers $k' \leq k$ only access the d previous words). By systematically and independently varying k and d, we can compute β_{distance} and β_{layer}: the two coefficients that indicate how brain scores and comprehension scores vary across layers and attentional spans, respectively. Precisely, a positive β_{distance} indicates that scores are sensitive to long-range dependencies. On the contrary, a null β_{distance} indicates that scores are not sensitive to long-range-dependencies. Similarly, a positive β_{layer} indicates that deep layers have better scores than shallow layers, while a negative β_{layer} indicates that shallow layers have better scores than deep layers.

Our results are three-fold. First, both the brain score ($\mathcal{M}$) and the comprehension scores ($\mathcal{R}$) increase with the attention span ($\beta_{\text{distance}} > 0$, $p^M < 10^{-14}$ for brain scores, $p^R = .01$ for comprehension scores) as well as with the depth of the representation ($\beta_{\text{layer}} > 0$, $p^M < 10^{-4}$, $p^R = .001$). The gain in scores obtained with attention to distant context is observed even up to the most distant items (e.g. between distance $\approx 1,000$ and 300 words: $\Delta R > 0$, $p^M < 10^{-4}$, $p^R = .02$, Figure 2.6A).

Second, the attention span primarily impacts the brain scores and the comprehension scores of the middle layers (difference between layer 8 and layer 12: $\Delta\beta_{\text{distance}} = .001$, $p^M < 10^{-8}$ for brain scores, $\Delta\beta_{\text{distance}} = .03$, $p^R = .005$ for comprehension scores, Figure 2.6AD). Interestingly, and to our surprise, restricting the attention span of the first layers improved their ability to predict comprehension (e.g. for the first layer, difference between scores with an attention of 10 words and full attention $\Delta R = .06$, $p = .004$, Figure 2.6D). This unexpected result suggests that language transformers could be made more similar to the brain by increasing the attention span as a function of depth.

Finally, brain regions commonly associated with high-level comprehension are better predicted by the deep representations of past words, and their corresponding brain scores and comprehension scores are relatively strongly modulated by long-distance attention (e.g. in angular gyrus: $\beta_{\text{layer}} = .14 > 0$, $p = .002$, $\beta_{\text{distance}} = .03 > 0$, $p = .016$ for comprehension scores). On the contrary, low-level acoustic regions are best predicted by the shallow layers of the network, and are, in comparison, little altered by long-distance dependencies (e.g. for the

comprehension scores in Heschl gyrus, $\beta_{\text{layer}} = -.076 < 0$, $p = .004$, $\beta_{\text{distance}} = -.014 < 0$, $p = .012$).

Overall, our analysis suggests that comprehension depends on a hierarchy of neural representations, whereby the first areas of the language network deploys shallow and short-span attention processes, while the fronto–parietal network relies on compositional and long-span attention processes. Interestingly, our analysis also highlights that shortening the attention span of lower-layers makes them more brain-like, and could perhaps thus provide a useful inductive bias to these algorithms.

2.2.4 Discussion

Our analyses reveal a reliable correlation between story comprehension and the degree to which language transformers like GPT-2 maps onto brain responses to the corresponding story. Furthermore, the systematic comparison, decomposition and manipulation of such language models allow us to decompose (1) the brain regions (2) the level of representation (sub-lexical, lexical, supra-lexical) and (3) the attentional gating (i.e. the short- or long-range retrieval of past stimuli) that relate to the comprehension of complex narratives.

These findings complement prior work on the brain bases of comprehension in three major ways. First, a number of qualitative theories describe how words may be combined into meaningful representations (Hagoort et al., 2009; Hagoort, 2013; Hagoort & Indefrey, 2014; Bornkessel-Schlesewsky & Schlesewsky, 2006, 2013; Hickok & Poeppel, 2007; Ullman, 2001; Friederici, 2011). For example, the Memory, Unification and Control model (MUC) distinguishes three types of computations and links them to the temporal lobe, Broca area and the rest of the prefrontal lobe, respectively. Similarly, the extended Argument Dependency Model (eADM) proposes that the ventral and the dorsal streams of the auditory pathway compute time-independent and time-dependent unifications, respectively. Our results support an analogous division of acoustics, lexical and compositional representations in the language areas. However, we reveal a slightly different functional anatomy: the early areas of the language network, located around the auditory cortices, deploy sub-lexical and shallow representations thanks to short attention spans. By contrast, the fronto–parietal network tracks and unifies very distant contexts to current words (Figure 2.5F). How these cortical areas communicate with the hippocampus and retrieve words from long-term memory remains an exciting direction for future studies (Lu et al., 2022).

Second, several quantitative approaches have been proposed to investigate comprehension, either with "model-free" methods based on inter-subject correlation (e.g. (Lerner et al., 2011; Fedorenko et al., 2016; Dehghani et al., 2017)) or "model-based" methods based on word vectors (Broderick et al., 2020). For example, Lerner et al. analyzed the fMRI activity of subjects listening to either normal texts or texts scrambled at the word, sentence or paragraph level (Lerner et al., 2011). While brain activity correlated across subjects in the primary and secondary auditory areas even when the input was heavily scrambled (and thus poorly comprehensible), the bilateral infero-frontal and temporo-parietal cortex only correlated across subjects when sentences and/or paragraphs were not scrambled (and thus comprehensible). Broderick et al. used a similar design to investigate electro-encephalography (EEG) responses to variably scrambled versions of the same story (Broderick et al., 2020), as well as the EEG responses to speech played in reverse and in noise (Broderick et al., 2018). Consistently with our results, they showed that the mapping between word embeddings' and the EEG activity varies with comprehension as manipulated by these various protocols. Our results thus complement these findings by showing (1) the brain regions where GPT-2's predictions vary with subject's comprehension, and (2) what type of representations these features relate to: comprehension appears here to depend on a hierarchy of neural representations, whereby the first areas of the language network deploys shallow and short-span-attention processes, while the fronto–parietal network relies on compositional and long-span-attention processes.

Finally, previous analyses have investigated the role of attention in the brain (Sabri et al., 2008; Kok et al., 2012; Toneva & Wehbe, 2019). We complement these studies by (1) showing that very-long term attention affects brain scores (even above 1,000 words), (2) identifying the brain regions that are sensitive to long vs. short attention spans, and(3) investigating the interactions between attention span, the ability to generate brain-like representations, and one behavioral metric: comprehension.

Interestingly, some regions, like the angular and supramarginal gyri, present a modest brain score and nevertheless strongly predict comprehension. How can one interpret such dissociation? We propose that deep neural networks encode a variety of features, ranging from low- to high-level representations. While some of these features may relate to general language processing (e.g. short-range information about words), others may specifically relate and thus predict comprehension (e.g. long-range dependencies). In this view, the regions that are best predicted by GPT-2's representations (e.g. Heschel's gyrus) need not be identical to those that best predict comprehension (e.g. Angular gyrus). Our ablation studies fit this

hypothesis: the auditory cortices are marked by high brain scores but low comprehension scores (Figure 2.5G) and indeed appear to encode short-range and shallow representations – i.e. features that presumably only indirectly relate to the comprehension of a narrative (Figure 2.6). By contrast the angular gyrus demonstrates a high comprehension score (Figure 2.5G) and indeed appears to encode long-range dependencies and deep representations – i.e. features that presumably relate to the latent structures of narratives, and from which comprehension should depend (Figure 2.6).

Overall, the present study suggests that GPT-2 retrieves information that relates to human comprehension, thus strengthening previous works that study the similarities between deep language models and the brain (Caucheteux et al., 2021b; Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022; Caucheteux et al., 2021a; Hale et al., 2021; Anderson et al., 2021; Sun et al., 2021; Goldstein et al., 2022). For instance, several studies showed that deep nets' encoding accuracy correlated with the level of semantic and syntactic information of their activations (Sun et al., 2021), as well as their ability to predict a word from context (Schrimpf et al., 2021; Caucheteux & King, 2022). We complement these results and show that the encoding accuracy of GPT-2 correlates with the level of understanding of the subjects, as assessed with comprehension questionnaires. Interestingly, our analysis also highlights that shortening the attention span of lower-layers would make them more brain-like. Thus, these results contribute to revealing remaining functional differences between brains and language models, and could thus help guide the development of modern algorithms (Toneva & Wehbe, 2019; Caucheteux et al., 2023).

The relationship between GPT-2's representations and human comprehension remains to be qualified, however. First, we restrict the challenging and composite notion of semantic comprehension to an empirical definition: i.e. the extent to which subjects understand a narrative, as assessed by a questionnaire presented at the end of each story. We acknowledge that comprehension spans a very diverse set of conditions, ranging from scientific writing to newspapers, which are not presently tested.

Second, our results remain solely based on correlations. Supplementary analyses suggest that GPT-2's brain scores may be partially explained by – but not reduced to – attentional processes (Supplementary Note 6.2.8). Yet, the factors that causally influence comprehension, such as attention, prior knowledge, working memory capacity, and language complexity are not controlled here and should thus be explicitly examined and manipulated in future work. In

particular, it would be interesting to evaluate how working memory capacity, cognitive control, vocabulary, as well as an continuous-monitoring of subjects' attention separately contribute to the fluctuation of comprehension and specifically account for the link between GPT-2 and the brain. Similarly, the study of inter-individual differences could further help modeling specific cognitive deficits associated with comprehension such as dyspraxia, dyslexia or autistic syndrome. However, such investigation would likely require large amounts of data, and thus a dedicated effort (Marek et al., 2022).

Third, we find that the long-distance representations of GPT-2 middle layers specifically account for comprehension in associative cortices, while the short-distance information encoded in the shallow layers account for comprehension in lower-level brain regions. However, what these features actually represent remains largely unknown. Previous studies have shown that language transformers explicitly represent syntactic (Manning et al., 2020; Lakretz et al., 2021) and semantic features (Lakretz et al., 2021). Similarly, Manning et al. showed that syntactic trees appear to be encoded by the distances between contextualized word embedding (Manning et al., 2020). Clarifying the nature of word embeddings remains an important direction to explore (e.g. syntactic vs. semantic (Gauthier & Levy, 2019; Reddy & Wehbe, 2020; Sun et al., 2021; Caucheteux et al., 2021a).

Finally, although highly significant, and significantly better than alternative models (Supplementary Figure S9), the brain-scores of GPT-2 are relatively low (Huth, de Heer, et al., 2016; Fedorenko et al., 2016; Toneva & Wehbe, 2019). This phenomenon is largely expected: we fit and evaluate the brain mapping at the single-TR single-voxel level and across all brain voxels to avoid selection biases. Nonetheless, these brain scores reach up to 32% of the noise ceiling (Supplementary Note 6.2.4, Supplementary Figure S8). This indicates that while GPT-2 may be our best model of language representations in the brain, it remains far from fully capturing those of complex narratives.

The comparison between brains, behavior and deep nets was originally introduced in vision research (D. L. K. Yamins & DiCarlo, 2016). The present study strengthens this approach and clarifies the links between GPT-2 and the brain. Specifically, we show that GPT-2's mapping correlates with comprehension up to $\mathcal{R} = 0.50$. This result is both promising and limited: on the one hand, we reveal that the similarity between deep nets and the brain non-trivially relates to a high-level cognitive process. On the other hand, half of the comprehension variability remains unexplained by this algorithm.

This limit is expected: several studies demonstrate that current deep language models fail to capture several aspects critical to comprehension (Marcus, 2020a; B. M. Lake & Murphy, 2021): they (i) often fail to generalize beyond the training distribution (Baroni, 2020), (ii) do not perfectly capture deep syntactic structures (Manning et al., 2020; Lakretz et al., 2021) and (iii) remain relatively poor at summarizing texts, generating stories and answering questions (Holtzman et al., 2020; Wiseman et al., 2017; Thakur et al., 2021). Furthermore, GPT-2 is only trained with textual data and does not situate objects in a grounded environment that would capture their real-world interactions (Bisk et al., 2020; McClelland et al., 2020). These limits may be temporary, however: the latest models appear to be more robust to out-of-distribution sampling (Brown et al., 2020) and trained on multimodal data (Radford et al., 2021; Ramesh et al., 2021).

Together, these elements thus suggest that modern language algorithms like GPT-2 offer a promising basis to unravel the brain and computational signatures of comprehension. Vice versa, by highlighting the similarities and remaining differences between deep language models and the brain, our study reinforces the mutual relevance of neuroscience and AI.

2.2.5 Methods

Our analyses rely on the "Narratives" dataset (Nastase et al., 2020), composed of the brain signals, recorded using fMRI, of 345 subjects listening to 27 narratives. The dataset is publicly available and the methods were performed in accordance with relevant guidelines and regulations.

Narratives and comprehension score Among the 27 stories of the dataset, we selected the seven stories for which subjects were asked to answer a comprehension questionnaire at the end, and for which the answers varied across subjects (more than ten different comprehension scores across subjects), resulting in 70 min of audio stimuli in total, from four to 19 minutes per story (Figure 2.7). Questionnaires were either multiple-choice, fill-in-the blank, or open questions (answered with free text) rated by humans (Nastase et al., 2020). Here, we used the comprehension score computed in the original dataset which was either a proportion of correct answers or the sum of the human ratings, scaled between 0 and 1 (Nastase et al., 2020). It summarizes the comprehension of one subject for one narrative (specific to each (narrative, subject) pair).

Brain activations The brain activations of the 101 subject who listened to the seven selected narratives were recorded using fMRI. As suggested in the original paper (Nastase et al., 2020), pairs of (subject, narrative) were excluded because of noisy recordings, resulting in 237 pairs in total.

All seven studies used a repetition time (TR) of 1.5 seconds. As stated in the orginal paper (Nastase et al., 2020), the "Merlin", "Sherlock", "Slumlord" and "Reach for the Stars" datasets were collected on a 3T Siemens Magnetom Skyra (Erlangen, Germany) with a 20-channel phased-array head coil using the following acquisition parameters. "Functional BOLD images were acquired in an interleaved fashion using gradient-echo echo-planar imaging (EPI) with an in-plane acceleration factor of 2 using GRAPPA. The full acquisition details are summarized here for simplicity: TR/TE = 1500/28 ms, flip angle = 64 degrees, bandwidth = 1445 Hz/Px, in-plane resolution = 3x3mm, slice thickness = 4 mm, matrix size = 64x64, FoV = 192x192 mm, 27 axial slices with roughly full brain coverage and no gap, anterior–posterior phase encoding, prescan normalization, fat suppression. At the beginning of each run, three dummy scans were acquired and discarded by the scanner to allow for signal stabilization.

The "Pie Man (PNI)" (pieman-pni) "Running from the Bronx"(bronx), "I Knew You Were Black" (black) and "The Man Who Forgot Ray Bradbury"(forgot) datasets were collected on the same 3T Siemens Magnetom Prisma with a 64-channel head coil using different acquisition parameters. Functional images were acquired in an interleaved fashion using gradient-echo EPI with a multiband acceleration factor of 3 using blipped CAIPIRINHA and no in-plane acceleration: TR/TE 1500/31 ms, flip angle = 67degrees, bandwidth = 2480 Hz/Px, in-plane resolution = 2.5x2.5mm, slice thickness 2.5 mm, matrix size = 96x96, FoV = 240x 240 mm, 48 axial slices with full brain coverage and no gap, anterior–posterior phase encoding, prescan normalization, fat suppression, three dummy scans."

GPT-2 activations GPT-2 (Radford et al., 2019) is a high-performing neural language model trained to predict a word given its previous context (it does not have access to succeeding words), given millions of examples (e.g Wikipedia texts). It consists of multiple Transformer modules (twelve, each of them called "layer") stacked on a non-contextual word embedding (a look-up table that outputs a single vector per vocabulary word) (Radford et al., 2019). Each layer l can be seen as a nonlinear system that takes a sequence of w words as input, and outputs a contextual vector of dimension (w, d), called the "activations" of layer l ($d = 768$). Intermediate layers were shown to better encode syntactic and semantic information than input and output layers (Jawahar et al., 2019), and to better map onto brain activity (Toneva & Wehbe, 2019;

Caucheteux & King, 2022). Here, we show that the *eighth* layer of GPT-2 best predicts brain activity 2.5C. We thus select the eighth layer of GPT-2 for our analyses. Our conclusions remain unchanged with other intermediate-to-deep layers of GPT-2 (from 6^{th} to 12^{th} layers).

In practice, the narratives' transcripts were formatted (replacing special punctuation marks such as "–" and duplicated marks "?." by dots), tokenized using GPT-2 tokenizer and input to the GPT-2 pretrained model provided by Huggingface (Wolf et al., 2020). The representation of each token is computed separately using a sliding context window of 1024 tokens. For instance, to compute the representation of the third token of the story, we input GPT-2 with the third, second and first token, and then extract the activations corresponding to the third token. Similarly, to compute the activations of the 1500^{th} token, we input the model with the word 1500 and the 1023 words before. Overall, the activations of every word w_k are computed by inputting the model with the word w_k and the 1023 previous tokens (at most), and then extracting the activations corresponding to w_k. The procedure results in a vector of activations of size (w, d) with w the number of tokens in the story and d the dimensionality of the model. There are fewer fMRI scans than words. Thus, the activation vectors between successive fMRI measurements are summed to obtain one vector of size d per measurement. To match the fMRI measurements and the GPT-2 vectors over time, we used the speech-to-text correspondences provided in the fMRI dataset (Nastase et al., 2020).

Linear mapping between GPT-2 and the brain For each (subject, narrative) pair, we measure the mapping between i) the fMRI activations elicited by the narrative and ii) the activations of GPT-2 (layer eight) elicited by the same narrative. To this end, a linear spatiotemporal model is fitted on a train set to predict the fMRI scans given the GPT-2 activations as input. Then, the mapping is evaluated by computing the Pearson correlation between predicted and actual fMRI scans on a held out set I:

$$\mathcal{M}^{(s,w)} : I \mapsto \mathcal{L}\left(f \circ g(X^{(w)})_{i \in I}, (Y_i^{(s,w)})_{i \in I}\right) \tag{2.5}$$

With $f \circ g$ the fitted estimator (g: temporal and f: spatial mappings), $\mathcal{L}$ Pearson's correlation, $X^{(w)}$ the activations of GPT-2 and $Y^{(s,w)}$ the fMRI scans of subjects s, both elicited by the narrative w.

In practice, f is a ℓ_2-penalized linear regression, following scikit-learn implementation (Pedregosa et al., 2011). The regularization parameter is chosen for each voxel separately using nested cross validation on the train set. Specifically, we use scikit-learn's RidgeCV

estimator with built-in leave-one-sample-out cross-validation, with ten possible regularization parameters log-spaced between 10^{-1} and 10^{8}, one hyper-parameter being selected for each voxel independently. g is a finite impulse response (FIR) model with 5 delays, where each delay sums the activations of GPT-2 input with the words presented between two TRs. For each (subject, narrative) pair, we split the corresponding fMRI time series into five contiguous chunks using scikit-learn cross-validation. The procedure is repeated across the five train (80% of the fMRI scans) and disjoint test folds (20% of the fMRI scans). Pearson correlations are averaged across folds to obtain a single score per (subject, narrative) pair. This score, denoted $\mathcal{M}(X)$ in Figure 2.5A, measures the mapping between the activations space X and the brain of one subject, elicited by one narrative.

Phonological features To account for low-level speech processing, we computed the alignment (Equation (2.5)) between the fMRI brain recordings Y and phonological features X: the word rate (of dimension $d = 1$, the number of words per fMRI scan), the phoneme rate ($d = 1$, the number of phonemes per fMRI scan) and the concatenation of phonemes, stresses and tones of the words in the stimuli (categorical feature, $d = 117$). The latter phonological features are provided in the original dataset, and computed using Gentle [1]. The 117 dimensions are the combination of phonetic categories, stresses and tones. We use 40 English phonemes in the corpus, and 4 possible tones, which results in 40 x 4 = 160 possible categories. Some categories are never pronounced here. If we ignore these categories, this results in 117 categories, and thus 117 dimensions after one-hot encoding.

Voxel-level and ROI-level analyses All of the first-level analyses are performed at the voxel level (computation of the mapping scores $\mathcal{M}$ of equation (2.5), in blue in Figure 2.5). We then average these effects either (1) within each brain region (Figure 2.5B, E, F and G) or (2) across the whole brain (Figure 2.5C and D). From these average values, we compute the correlation with comprehension (in red in Figure 2.5). This approach mitigates the localization of the effect and the statistical correction for multiple comparisons.

Significance Significance was either assessed by using either (i) a second-level Wilcoxon test (two-sided) across subject-narrative pairs, testing whether the mapping (one value per pair) was significantly different from zero (Figure 2.5B), or (ii) by using the first-level Pearson

[1] https://github.com/lowerquality/gentle

p-value provided by SciPy (Virtanen et al., 2020) (Figure 2.5D-G). In Figure 2.5B, E, F, p-values were corrected for multiple comparison (2×142 ROIs) using False Discovery Rate (Benjamin/Hochberg) (Gramfort et al., 2013).

Data availability The Narratives dataset (Nastase et al., 2020) is publicly available on the OpenNeuro (`https://openneuro.org/datasets/ds002345/versions/1.1.4`) and Datalad platforms (`http://datasets.datalad.org/?dir=/labs/hasson/narratives`).

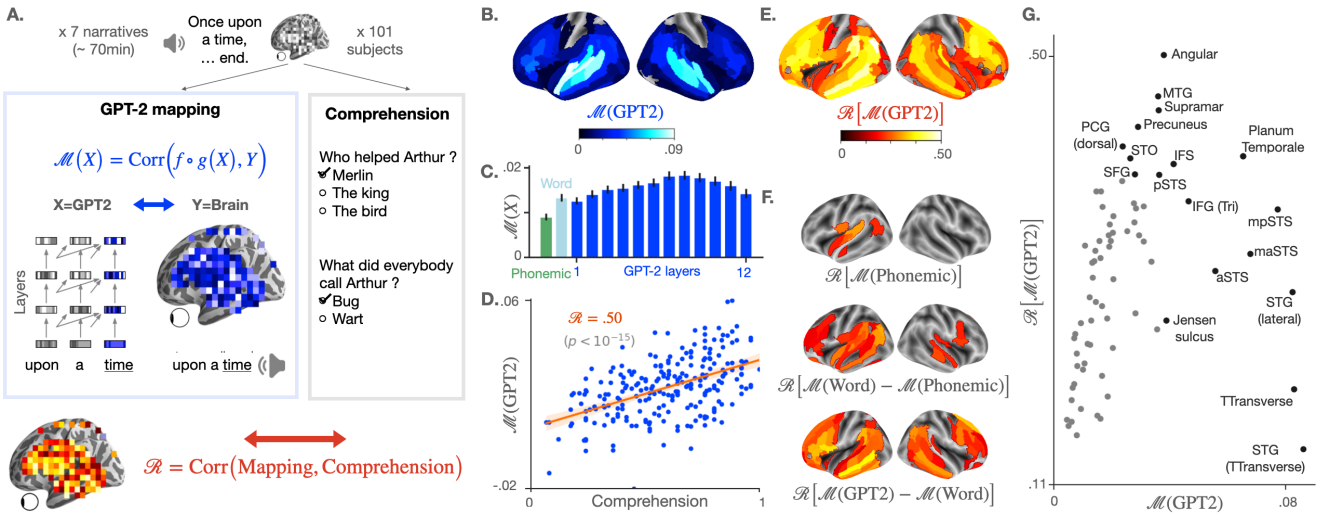


Figure 2.5: Methods and Results. A. 101 subjects listen to narratives (70 min of unique audio stimulus in total) while their brain signal is recorded using functional MRI. At the end of each story, a questionnaire is submitted to each subject to assess their understanding, and the answers are summarized into a *comprehension score* specific to each (narrative, subject) pair (grey box). In parallel (blue box on the left), we measure the mapping between the subject's brain activations and the activations of GPT-2, a deep network trained to predict a word given its past context, both elicited by the same narrative. To this end, a linear spatio-temporal model ($f \circ g$) is fitted to predict the brain activity of one voxel Y, given GPT-2 activations X as input. The degree of mapping, called "*brain score*" is defined for each voxel as the Pearson correlation between predicted and actual brain activity on held-out data (blue equation, cf. Methods). Finally, we test the correlation between the comprehension scores of the subjects and their corresponding brain scores using Pearson's correlation (red equation). A positive correlation means that the representations shared across the brain and GPT-2 are key for the subjects to understand a narrative. **B.** Brain scores (fMRI predictability) of the activations of the eighth layer of GPT-2. Scores are averaged across subjects, narratives, and voxels within brain regions (142 regions in each hemisphere, following a subdivision of Destrieux Atlas (Destrieux et al., 2010), cf. Supplementary Note 6.2.1). Only significant regions are displayed, as assessed with a two-sided Wilcoxon test across (subject, narrative) pairs, testing whether the brain score is significantly different from zero (threshold: .05). **C.** Brain scores, averaged across fMRI voxels, for different activation spaces: phonological features (word rate, phoneme rate, phonemes, tone and stress, in green), the non-contextualized word embedding of GPT-2 ("Word", light blue) and the activations of the contextualized layers of GPT-2 (from layer one to layer twelve, in blue). The error bars refer to the standard error of the mean across (subject, narrative) pairs (n=237). **D.** Comprehension and GPT-2 brain scores, averaged across voxels, for each (subject, narrative) pair. In red, Pearson's correlation between the two (denoted $\mathcal{R}$), the corresponding regression line and the 95% confidence interval of the regression coefficient. **E.** Correlations ($\mathcal{R}$) between comprehension and brain scores over regions of interest. Brain scores are first averaged across voxels within brain regions (similar to B.), then correlated to the subjects' comprehension scores. Only significant correlations are displayed (threshold: .05). **F.** Correlation scores ($\mathcal{R}$) between comprehension and the subjects' brain mapping with phonological features ($\mathcal{M}(\mathrm{Phonemic})$) (i), the share of the word-embedding mapping that is not accounted by phonological features $\mathcal{M}(\mathrm{Word}) - \mathcal{M}(\mathrm{Phonemic})$ (ii) and the share of the GPT-2 eighth layer's mapping not accounted by the word-embedding $\mathcal{M}(\mathrm{GPT2}) - \mathcal{M}(\mathrm{Word})$ (iii). **G.** Relationship between the average GPT-2-to-brain mapping (eighth layer) per region of interest (similar to B.), and the corresponding correlation with comprehension ($\mathcal{R}$, similar to D.). Only regions of the left hemisphere, significant in both B. and E. are displayed. In black, the top ten regions in terms of brain and correlation scores (cf. Supplementary Note 6.2.1 for the acronyms). Significance in D, E and F is assessed with Pearson's p-value provided by SciPy (Virtanen et al., 2020). In B, E and F, p-values are corrected for multiple comparison using a False Discovery Rate (Benjamin/Hochberg) over the 2×142 regions of interest.

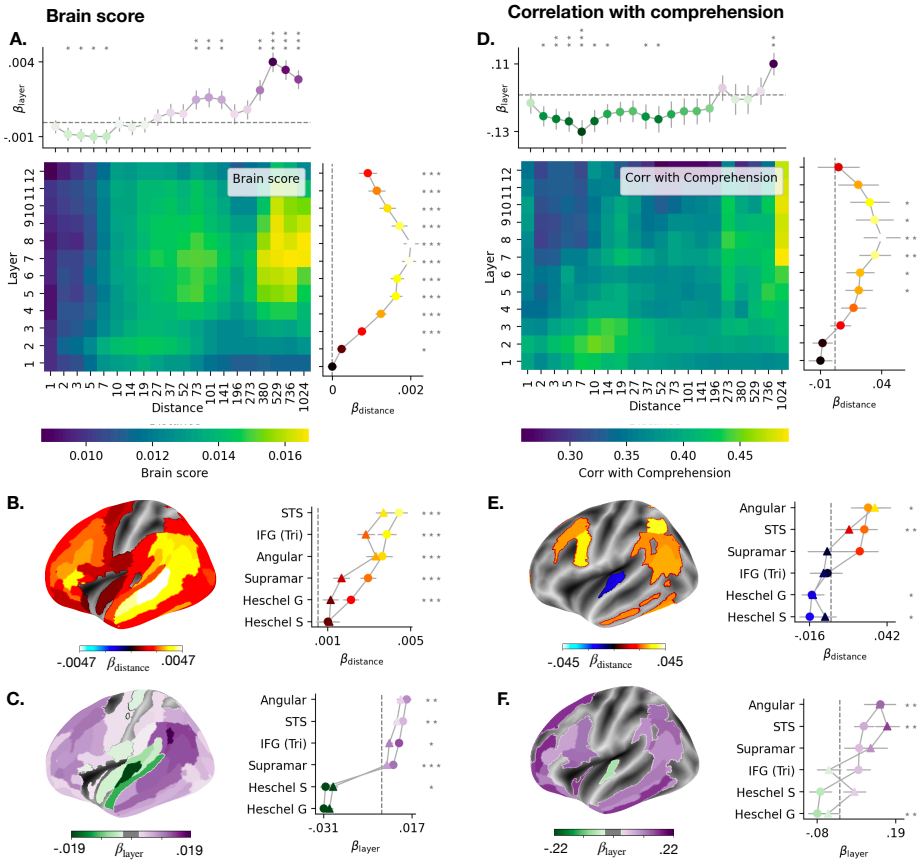


Figure 2.6: Effect of GPT-2's attention span on brain scores and comprehension scores. A. The heatmap displays the average (across subjects, stories and voxels) brain scores as a function of attention span ("distance") and layers. The top line displays the layer coefficients for each attention span (averaged across subjects, stories and voxels). The right line displays the distance coefficient for each layer (averaged across subjects, stories and voxels). The error bars correspond to the Standard Errors of the Mean (SEM) across subject-story pairs. **B.** Distance coefficients for each brain region (averaged across subjects and stories). Statistical significance is assessed with a Wilcoxon test across subject-story pairs. **C.** Layer coefficients for each brain region (averaged across subjects and stories). **D-F**. Similar as A-C, but the layer (and distance, respectively) coefficients now assess the relationship between layer (or distance, respectively) and comprehension scores. Statistical significance is assessed using a bootstrapping procedure with 1,000 subsamples of subject-story pairs. Error bars are standard deviation across subsamples. For all brain plots, only significant values are displayed ($p < 0.05$ after FDR correction across brain regions.)

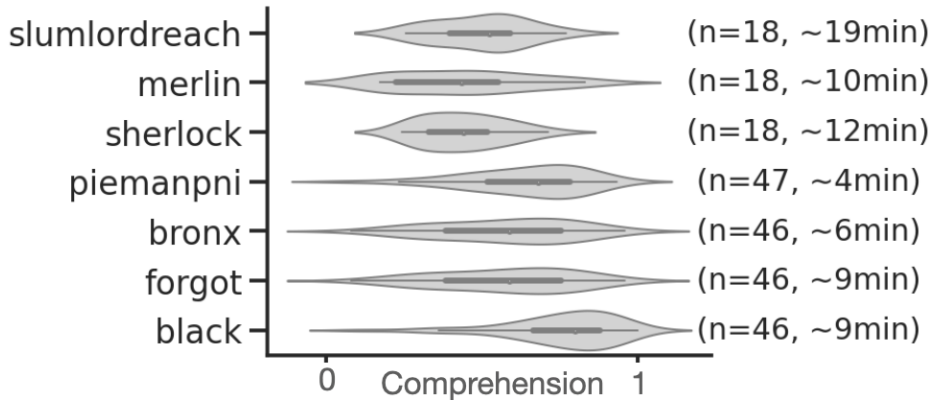


Figure 2.7: Distribution of comprehension scores. For each of the seven narratives: number of subjects (n), distribution of comprehension scores across subjects and length of the narrative.

Chapter 3

Leveraging the similarity to decompose the content, temporal and spatial organization of language representations in the brain

3.1 Disentangling syntax and semantics in the brain with deep networks

3.1.1 Abstract

The activations of language transformers like GPT-2 have been shown to linearly map onto brain activity during speech comprehension. However, the nature of these activations remains largely unknown and presumably conflate distinct linguistic classes. Here, we propose a taxonomy to factorize the high-dimensional activations of language models into four combinatorial classes: lexical, compositional, syntactic, and semantic representations. We then introduce a statistical method to decompose, through the lens of GPT-2's activations, the brain activity of 345 subjects recorded with functional magnetic resonance imaging (fMRI) during the listening of $\sim$4.6 hours of narrated text. The results highlight two findings. First, compositional representations recruit a more widespread cortical network than lexical ones, and encompass the bilateral temporal, parietal and prefrontal cortices. Second, contrary to previous claims, syntax and semantics are not associated with separated modules, but, instead, appear to share a common and distributed neural substrate. Overall, this study introduces a versatile framework to isolate, in the brain activity, the distributed representations of linguistic constructs.

3.1.2 Introduction

Within less than three years, transformers have enabled remarkable progress in natural language processing (Devlin et al., 2019; Radford et al., 2019). Pretraining these architectures on millions of texts to predict words from their context greatly facilitates translation, text synthesis and the retrieval of world-knowledge (Lample & Conneau, 2019; Brown et al., 2020).

Interestingly, the activations of language transformers tend to linearly map onto those of the human brain, when presented with the same sentences (Jain & Huth, 2018; Toneva & Wehbe, 2019; Abnar et al., 2019; Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022). This linear mapping suggests that, in spite of their vast learning[1] and architectural differences[2], the brain and language transformers converge to similar linguistic representations (Caucheteux & King, 2022; Caucheteux et al., 2022).

However, the nature of these shared representations remains largely unknown. Three factors explain this gap-of-knowledge. First, linguistic theories are generally described and interpreted in terms of combinatorial *symbols* (discrete words, syntactic trees, etc). In contrast, brain and language transformers generate high-dimensional *vectors* (a.k.a "distributed" representations). While these formats are formally equivalent (Smolensky, 1990), interpreting vectorial representations in language models and in the brain is particularly challenging.

Second, the representations of deep learning models have been interpreted independently of brain imaging. For example, deep neural networks have been shown to encode lexical analogies in their word embeddings (Mikolov, Sutskever, et al., 2013), as well as singular/plural relationships (Lakretz et al., 2019), long-distance dependency information (Jawahar et al., 2019), and syntactic trees (Manning et al., 2020). Similarly, the brain responses to language have been decomposed into a cascade of representations, which maps speech and reading input into phonetic (or orthographic), morphemic, lexical, and syntactic representations (Hickok & Poeppel, 2007; Dehaene & Cohen, 2011; Pallier et al., 2011; Friederici, 2011; Mesgarani et al., 2014; Huth, de Heer, et al., 2016; Nelson et al., 2017; J. R. Brennan & Hale, 2019; Gwilliams et al., 2020). However, we do not know whether all or any of these representations effectively drive the linear mapping between language models and the brain.

Third, the mapping between language transformers and the brain has been mainly investigated with speech and/or narratives (Schrimpf et al., 2021; Toneva & Wehbe, 2019; Abnar et al.,

[1]The brain learns continuously from a small set of situated sentences, whereas transformers learn from large sets of pure texts.

[2]The brain is a single-stream recurrent architecture, whereas the transformer is a multi-stream feedforward architecture.

2019; Reddy & Wehbe, 2020) (although see (Caucheteux & King, 2022)). The resulting sentences are thus poorly controlled and potentially confound various features such as phonological variations, sentiment contours, semantic contents, and syntactic properties (*e.g.* stressful texts may tend to be read more quickly, and make use of smaller constituency trees). In sum, the linear correspondence observed between language models and the brain may be driven by a wide variety of factors.

Here, we aim to decompose the similarity between the brain and high-performance language transformers like GPT-2 (Radford et al., 2019), in light of four distinct linguistic classes, namely lexical, compositional, syntactic and semantic representations. To that end, we formalize a taxonomy that factorizes them into four distinct vector bases. We then describe a statistical procedure to extract syntactic representations from neural networks, decompose their lexical and compositional components, and separate them from semantic representations. Finally, we assess the linear mapping between i) the factorized activations of GPT-2 and ii) the brain signals of 345 subjects listening to the same narratives (4.6 hours of audio stimulus in total) as recorded with functional magnetic resonance imaging (fMRI) (Nastase et al., 2020).

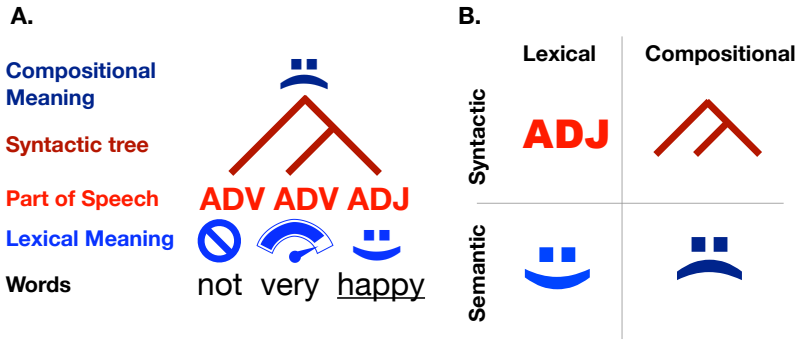


Figure 3.1: Taxonomy A. To understand the meaning of a phrase, one must combine the meaning of each word using the rules of syntax. For example, the meaning of the phrase NOT VERY HAPPY is (roughly) SAD, and can be found by recursively combining the two adverbs and the adjective. **B.** Here, we aim to decompose lexical features (what relates to the word level) from the compositional features (what relates to a combination of words) both for syntactic representations (e.g. part-of-speech versus syntactic tree) and for semantic representations (e.g. the set of word meaning versus the meaning of their combination).

3.1.3 Operational Taxonomy

The notions of lexicon, composition, syntax and semantics are notoriously debated in linguistics. Without pretending to resolve these debates, we propose five definitions that unambiguously decompose the distributed representations of artificial and biological neural networks.

First, we use the standard definition of a *representation* as the information that can be linearly extracted from a vector of activations, with the rationale that a single artificial or biological neuron can read-out this information (Kriegeskorte et al., 2008; King et al., 2018). In this view, a system Ψ_1 is said to share the representation of a system Ψ_2 if there exists a linear mapping from X to Y, where $X = \Psi_1(w)$ and $Y = \Psi_2(w)$ are the activations elicited by the words w in each system.

Second, we define *lexical* representations as the representations that are context-invariant. This definition follows the standard notion of (non-contextualized) word-embeddings, which associate a unique vector to each word of a dictionary.

By contrast, we define *compositional* representations as the "contextualized" representations generated by a system combining multiples words: $\Psi(w_1 \dots w_M)$. For clarity, we restrict the term "compositional" to its strict sense: *i.e.* to the set of representations that cannot be accounted for by lexical representations, and thus by a linear combination of word-embeddings.

Fourth, we define *syntactic* representations as the set of representations associated with the structure of sentences independently of their meaning. Linguistic theories have proposed symbolic representations of such structures (e.g part-of-speech, dependency and constituency trees, see Figure 3.1). Furthermore, deep language models have been shown to linearly encode some of these features (Jawahar et al., 2019; Manning et al., 2020; Lakretz et al., 2019, 2020; Linzen & Baroni, 2021). Here, we introduce a versatile method to extract the distributed representations of syntax in a deep language model. Specifically, we extract these syntactic representations from the average activations elicited by a set of synthetic sentences that share the same syntactic properties (Section 3.1.4).

Finally, even though a variety of meaningful features are captured by both word embeddings (Mikolov, Sutskever, et al., 2013) and contexualized embeddings (Radford et al., 2019), meaning and semantics are notoriously difficult to define formally (Jackendoff, 2002). To decompose syntax and semantics in distributed representations, we thus propose to define *semantic* representations as the lexical or supra-lexical representations of a language system that are not syntactic.

According to these five definitions, lexical and compositional classes fully decompose both syntax and semantics (and *vice versa*). For example, lexico-syntactic representations refer to the functional categories of words (part-of-speech *i.e.* verb, noun, adjective, *etc.*). By contrast, compositional syntax refers to the representations that link words with one another, typically referred to as dependency (or constituency) trees. For example, in the phrase NOT VERY HAPPY (Figure 3.1), the set of lexical meaning can be distinguished from their compositional meaning.

The representation of this composition need not contain syntactic information, because its outcome (≈SAD) can be similar across phrases following distinct syntactic structures (*e.g.* NOT VERY HAPPY = DOWN IN THE DUMPS = SOMEWHAT SAD, *etc.*). Note that, under this definition, the distributed representations of syntax need not have a symbolic counterpart in theoretical linguistics – *e.g.* temporary structures that allow building the syntactic tree of a sentence, represent multiple alternative and their respective probabilities *etc.*

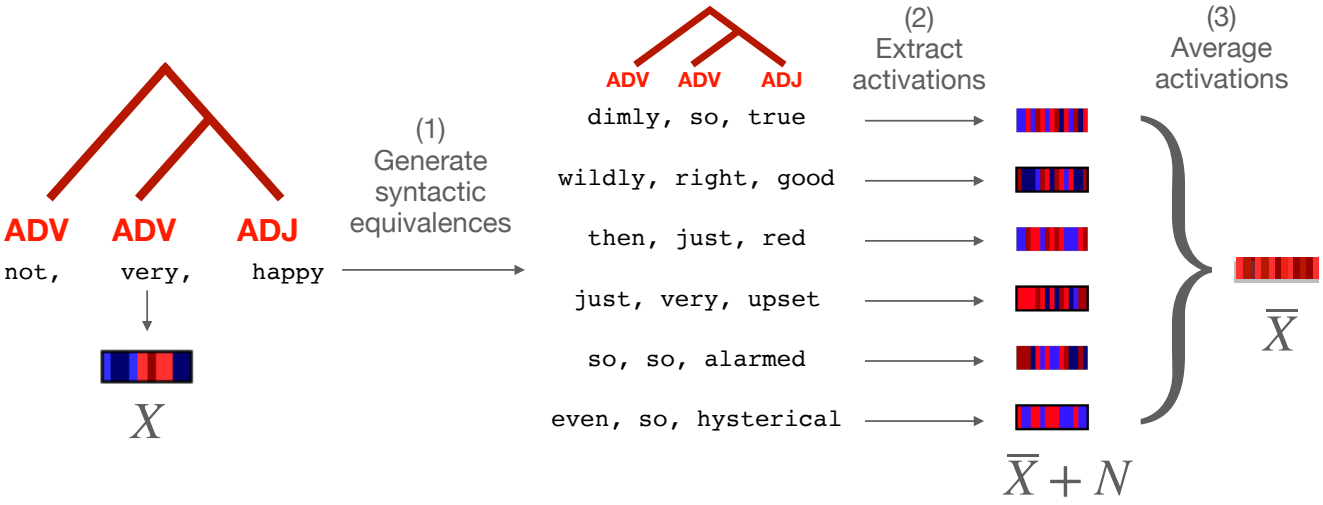


Figure 3.2: Method to isolate syntactic representations in GPT-2's word and compositional embeddings. To isolate the syntactic representations of a sequence of words e.g. w = NOT VERY HAPPY, we (1) synthesize sentences with the same syntactic structure as w (e.g DIMLY SO TRUE, *etc.*), then (2) extract the corresponding GPT-2 activations (from layer 9), and finally (3) average these activation vectors across the synthesized sentences. The resulting vector $\overline{X}$ is an approximation of the syntactic representations of X in GPT-2.

3.1.4 Methods

Isolating Syntactic Representations

We introduce below a method to isolate distributed representations of syntax in neural networks. We assume that a system Ψ ($\Psi : \mathcal{V}^M \rightarrow \mathbb{R}^{d \times M}$, $\mathcal{V}$ a vocabulary of words), takes sequences of M words as inputs and generates activations that encode syntactic properties (among other properties).

Let w be a sentence of M words ($w \in \mathcal{V}^M$, e.g THE CAT IS ON THE MAT), and Ω_w be the set of sentences that have the same syntax as w (*e.g.* A BOY GOES TO A POOL, THIS BOAT FLOATS NEAR THE SHORE, *etc.*). The syntactic representation of w is, by construction, also the syntactic

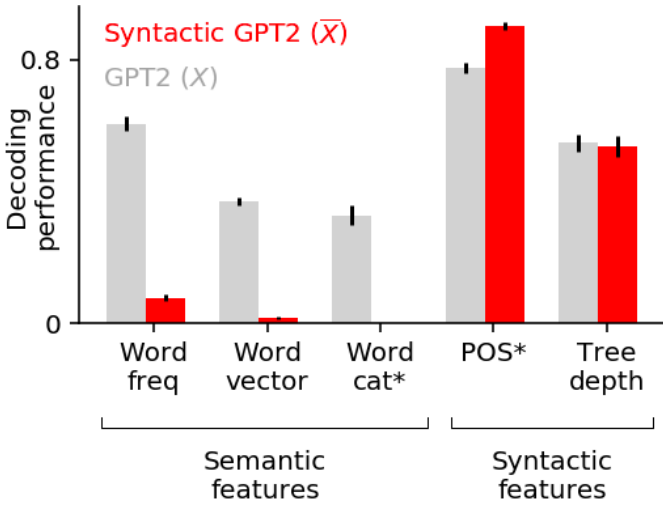


Figure 3.3: Semantic and syntactic information encoded in $\overline{X}$. To check that the syntactic embeddings $\overline{X}$ only contain syntactic information, we train a ℓ_2-regularized linear model to predict three semantic features (frequency, word embeddings and semantic category of content words Binder et al. (2016)) and two syntactic features (part-of-speech and depth of syntactic tree), given the syntactic embedding $\overline{X}$ (red), or the full GPT-2 activations X (grey) (Appendix 6.3.3). On the y-axis, the decoding performance of the model on left-out data (*adjusted* accuracy for the categorical features marked with a star, R^2 for the other continuous features). The chance level is zero. Semantic features (left) can be decoded from X (grey), but not from $\overline{X}$ (red), while syntactic features (right) can be decoded from both.

representations of all sentences $w' \in \Omega_w$. If this common syntactic representation is denoted $\overline{\psi} \in \mathbb{R}^d$, we have:

$$\forall w' \in \Omega_w, \quad \Psi(w') = \overline{\psi} + z_{w'}$$

with $z_{w'}$ a random perturbation of distribution $\mathbb{P}_w$, that corresponds to the non-syntactic part of the randomized activations $\Psi(w')$. If the density of $\mathbb{P}_w$ is well-defined and centered around 0, then:

$$\mathbb{E}\left[\Psi(w')\right] = \overline{\psi} \ ,$$

where w' is sampled uniformly in Ω_w. Thus, $\overline{\psi}$ (the syntactic representation of w) can be approximated through:

$$\overline{\Psi}_k = \frac{1}{k}\sum_{i=1}^{k}\left(\overline{\psi} + z_{w_i}\right) \xrightarrow[k\to\infty]{l.l.n} \overline{\psi}$$

with $(z_{w_1}, \ldots, z_{w_k})$ *i.i.d* samples from $\mathbb{P}_w$.

Overall, the syntactic component of the activations is the average of activations induced by random sentences of the same syntax (Figure 3.2).

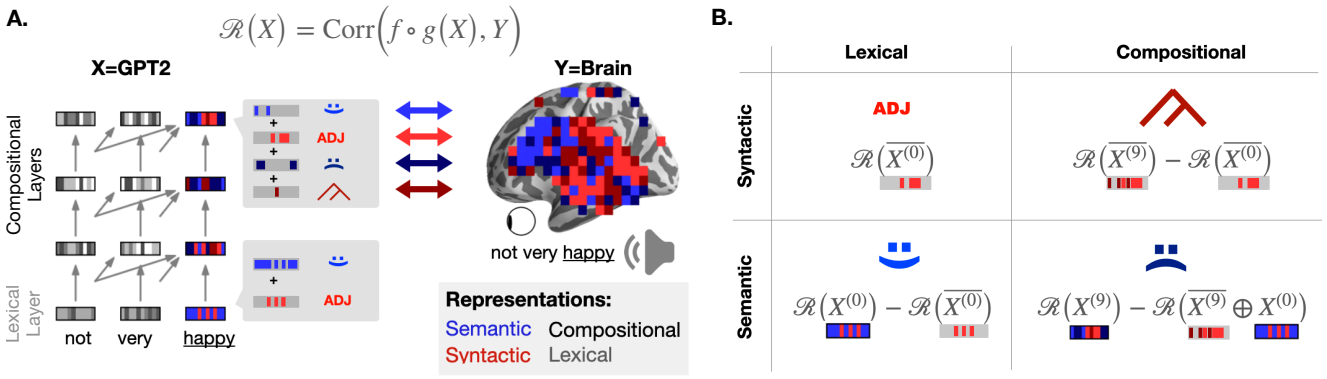


Figure 3.4: Method to decompose the language representations shared between brains and deep language models A. The human brain and modern language models like GPT-2 both generate *distributed* representations, which are thus difficult to link with the *symbolic* properties of linguistic theories. We introduce a method to decompose the representations of GPT-2, and the corresponding activations X onto the brain activations Y, elicited by the same sequence of words (e.g. NOT VERY HAPPY) with a spatio-temporal estimator $f \circ g$. This mapping is evaluated through cross-validation, with a Pearson correlation between the predicted and the actual brain signals $\mathcal{R}(X)$. **B.** Comparison used to decompose the brain score $\mathcal{R}(X)$ into the four linguistic components. $X^{(l)}$ refers to the the l^{th} layer's activations of GPT-2 input with the sentences heard by the subjects; $\overline{X^{(l)}}$ refers to the average l^{th} layer's activations of GPT-2 input with the synthetic sentences with a similar syntax (cf. Figure 3.2); $\oplus$ indicates a feature concatenation, and '$-$' indicates a subtraction between scores.

Mapping Representations onto FMRI Signals

In the present section, we aim to map the activations of two systems Ψ_1, a neural network, and Ψ_2, the brain, input with the same sequence words $w = (w_1, \ldots, w_M)$. Let $X = \Psi_1(w) \in \mathbb{R}^{M \times d}$ be a vector of Ψ_1 activations elicited by w (M vectors of dimension d, one per input word), and $Y = \Psi_2(w) \in \mathbb{R}^N$ the observable brain response at each of the N fMRI recorded time sample (a.k.a TR). For simplicity, we consider the analysis for one particular fMRI voxel, the same analysis can be repeated to map X with every voxel in the brain.

To assess the mapping between X and Y, we use the standard model-based encoding analysis of fMRI signals (Huth, de Heer, et al., 2016; D. L. K. Yamins & DiCarlo, 2016; Naselaris et al., 2011), and evaluate a linear spatio- (f) temporal (g) encoding model trained to predict the i^{th} fMRI volume given the network's activations X, on a given interval $I \subset [1 \ldots N]$:

$$\mathcal{R}(X) : f \mapsto \mathcal{L}\left(f \circ g(X)_{i \in I}, \overline{(Y_i)}_{i \in I}\right) \tag{3.1}$$

Specifically, given a story w of M words ($w = (w_1, \ldots, w_M) = (\text{THE}, \text{CAT}, \text{IS}, \text{ON}, \text{THE}, \text{MAT}, \ldots \text{END})$, we first extract the corresponding brain measurements Y of length N time samples. To max-

imize signal-to-noise ratio, we average the responses across the subjects that listened to that story, and apply the analysis to the average signal $\overline{Y}$.

The sampling frequency of fMRI is typically lower than word rate. Furthermore, fMRI signals are associated with delayed time responses that can span several seconds. Following others (Huth, de Heer, et al., 2016; Deniz et al., 2019; Shain et al., 2020), we align the word-times features X, of length M, to the dynamics of the fMRI signals applying a finite impulse response (FIR) model g (cf. Appendix 6.3.4).

Finally we learn a "spatial" mapping $f \in \mathbb{R}^d$ from the zero-mean unit-variance of X to the zero-mean unit-variance fMRI recordings Y with a ℓ_2-regularized "ridge" regression:

$$\underset{f}{\operatorname{argmin}} \sum_{i \in I_{\text{train}}} \left(\overline{Y_i} - f^T g(X)_i\right)^2 + \lambda ||f||^2$$

with λ the regularization parameter. We summarize the mapping with a Pearson correlation score evaluated on left out data:

$$\mathcal{R} = \operatorname{corr}\left(f \circ g(X), \overline{Y}\right) \ . \tag{3.2}$$

This correlation score measures the linear mapping between the brain and the activation space X. Following others (D. L. K. Yamins & DiCarlo, 2016), we will refer to this score as the *brain score* of the embedding X.

Decomposing Shared Activations between Brains and Neural Language Models

Here, we use the definitions and methods introduced in Section 3.1.3, 3.1.4 and 3.1.4 to decompose the shared representations of two systems: a deep neural network that encode linguistic properties, and the average brain of 345 subjects listening to narratives.

To that end, we (i) compute the activations of the neural language model elicited by the same narratives as the subjects (ii) factorize its activations into linguistic components, (iii) map with supervised learning the factorized components onto brain activity, and finally (iv) decompose the brain activations by evaluating this mapping.

Language transformers are composed of multiple layers ($l \in [1 \dots L]$), stacked over a (non contextualized) word embedding layer ($l = 0$). Each layer can be written as a non-linear system $\Psi^{(l)}$ that transforms a sequence of words w (e.g. NOT, VERY, HAPPY) into a vectorial representation of the same length,

$$\begin{aligned} \Psi^{(l)} : \mathcal{V}^M &\to \mathbb{R}^{M \times d} \\ w &\mapsto \Psi^{(l)}(w) = [\Psi^{(l)}(w)_1, \dots, \Psi^{(l)}(w)_M] \end{aligned}$$

with $\mathcal{V}$ the set of vocabulary words, M the length of the sequence, and d the dimensionality of the output representation taken at each word.

We denote $X^{(l)}$ the activations of $\Psi^{(l)}$elicited by w, and $\overline{X^{(l)}}$ the syntactic representations extracted from $X^{(l)}$ using the method introduced in Section 3.1.4. Following the definitions of Section 3.1.3, we can decompose the activations X of Ψ into their:

- lexical representations: $X^{(0)}$, the word embedding of the network.
- compositional representations: $X^{(l)}, l > 0$.
- syntactic representations: $\overline{X^{(l)}}$, that can be extracted for any layer $l \in [0 \dots L]$. The *lexical* syntactic representations $\overline{X^{(0)}}$ isroughly equivalent to the part-of-speech of the word. *Compositional* syntactic representations can be extracted from any layer $l > 0$ that encode syntactic information.
- semantic representations: $X^{(l)} - \overline{X^{(l)}}$, as the residuals of syntactic representations. They can be defined at both the lexical $X^{(0)} - \overline{X^{(0)}}$ and compositional levels ($l > 0$).

In practice, to verify that our syntactic embedding ($\overline{X}$) only contains syntax, we evaluate its ability to predict three semantic and two syntactic features (Figure 3.3, Appendix 6.3.3). The results confirm that semantic features can be decoded from X but not from $\overline{X}$, whereas syntactic features can be decoded from both.

Finally, following Section 3.1.4, we can compute the brain scores of the network's representations to decompose brain activity into:

- lexical representations: $\mathcal{R}(X^{(0)})$
- compositional representations: $\mathcal{R}(X^{(l)}), l > 0$. *Strictly* compositional representations are defined as the compositional representations that cannot be explained by lexical features: $\mathcal{R}(X^{(l)}) - \mathcal{R}(X^{(0)})$, with $l > 0$. For clarity, and except if stated otherwise, we will refer to strictly compositional representations as "compositional" representations.
- syntactic representations: $\mathcal{R}(\overline{X^{(l)}}), l \in [0 \dots L]$
- semantic representations: $\mathcal{R}(X^{(l)}) - \mathcal{R}(\overline{X^{(l)}})$, i.e. the residual brain scores of syntactic representations, for any layer $l \in [0 \dots L]$

3.1.5 Experiments

Here, we apply the general method described in Section 3.1.4, 3.1.4 and 3.1.4 to decompose the activations of two nonlinear systems, GPT-2 (Ψ_1) and the brain activity of 345 subjects listening to narratives (Ψ_2).

Functional MRI dataset. We analyze the "Narratives" public dataset (Nastase et al., 2020), which contains the fMRI measurements of 345 unique subjects listening to narratives. The narratives consist of 27 English spoken stories, ranging from ≈ 3 minutes to ≈ 56 minutes, for a total of ≈ 4.6 hours of unique stimuli. The original paper included two fMRI preprocessing pipelines, one with spatial smoothing and the other without. All our analyses are tested on the unsmoothed fMRI. As suggested in the original paper, we exclude (story, subject) pairs because of noisy fMRI recordings or missing transcripts, resulting in 617 unique (story, subject) pairs in total and ≈ 4 hours of unique audio stimuli.

Phonological features. To focus on lexical and supra-lexical language processing – as opposed to low-level speech processing, we extract three potential sets of confounds: the phone rate (the number of phones between two fMRI measurements, of dimension 1), the word rate (the number of words between two fMRI measurements) and the concatenation of the phoneme, stress and tone of the words in the stimulus. For each story, a phoneme-level transcript was provided in the Narratives database thanks to Gentle[3], a forced-alignment algorithm. Gentle annotations led to 117 unique categories (with unique phone, stress and tone), resulting in a one-hot encoded feature of the same dimension.

Language model features. GPT-2 is a high-performing causal (i.e. left to right) language model trained to predict a word given its previous context (Radford et al., 2019), and known to generate brain-like representations (Goldstein et al., 2022; Caucheteux & King, 2022; Affolter et al., 2020; Schrimpf et al., 2021; Caucheteux et al., 2022). It is comprised of 12 Transformer (contextual) layers ($l \in [1 \dots 12]$) stacked over a (non-contextual) embedding layer ($l = 0$), each of dimensionality 768, with 1.5 billion parameters in total. We used the pretrained version of GPT-2 from Huggingface (Wolf et al., 2020), trained on a dataset of 8 million web pages. In practice, the 27 stories are pre-processed, tokenized and input to the model (Appendix 6.3.1). The activations of each GPT-2 layer are extracted, resulting in 12 vectors of 768 activations for

[3]https://github.com/lowerquality/gentle

each token of each story transcript. For comparison, we also study five other transformers: BERT (Devlin et al., 2019), XLnet (Yang et al., 2020), Roberta (Liu et al., 2019), AlBert (Lan et al., 2020) and DistilGPT-2 (a smaller version of GPT-2) and recover similar – although lower – brain scores (Appendix 6.3.1).

Extracting syntactic representations from GPT-2 . To isolate the syntactic representations of GPT-2 , we synthesize, for each sentence of each story, $k = 10$ sentences with the same syntactic structures (Figure 3.2). We ensure in supplementary analyses that (i) the k synthetic sentences do *not* include the target sentence and (ii) these syntactic embeddings ($\overline{\Psi_k}$) lead to stable representations of syntax (Appendix 6.3.2). To this end, we proceed as follows:

- The transcript is formatted, split into sentences and tokenized using the large English tokenizer provided by spaCy (Honnibal et al., 2020) (cf. Appendix 6.3.1).
- Then, we use Supar, a state-of-the art dependency parser (Y. Zhang et al., 2020) to extract the dependency structure of each sentence and the part-of-speech.
- For each target word of each sentence of the Narratives dataset, we sample, from a $\approx$ 58,000 word corpus, consisting of Wikipedia combined with Narratives' transcripts, up to to $k' = 1,000$ words that have the same part-of-speech and dependency tags (e.g. CAT: NOUN, SINGULAR, SUBJECT OF). At this stage, k' versions of the target Narratives transcripts are synthesized.
- The synthesized sentences are not always grammatically correct. Thus, we automatically correct the sentences with Gector (Omelianchuk et al., 2020), and filter out the sentences that do not have the same length or part-of speech as the target sentence in the Narratives corpus.
- Some of the generated sentences may end up with a distinct syntactic tree than the original sentence, because semantics can disambiguate syntax (e.g. I SHOT AN ELEPHANT IN MY PYJAMAS). To assess the syntactic similarity between the original and the generated sentences, we compute, from their respective syntactic trees, the Pearson correlation between the words' pairwise distances, following (Manning et al., 2020)'s method. Then, we select the sentences whose syntactic trees are the most similar. 95% of the generated sentences have a syntactic tree that correlates with the tree of the target sentence above R=90%.

Mapping embeddings onto onto the fMRI signals. As described in equation (3.1), we evaluate the mapping between a set of modeling features X and the fMRI signals $Y \in \mathbb{R}^{N \times d_y}$ by fitting a linear spatio- (f) temporal (g) encoding model. $f \circ g$ was fitted on $I_{\text{train}} = 99\%$ of the dataset, and evaluated on $I_{\text{test}} = 1\%$ of the left out-data (2.5 min of audio). We evaluate the quality of this mapping with a Pearson R correlation between predicted and actual brain signals on I_{test}. Specifically, we use the linear ridge regression from scikit-learn (Pedregosa et al., 2011), with penalization parameters chosen among 10 values log-spaced between 10^{-1} and 10^{8} and g was a finite impulse response (FIR) model with 5 delays, following (Huth, de Heer, et al., 2016). X and Y are normalized (mean=0, std=1) across scans for each story, using a robust scaler clipping below and above the 0.01^{st} and 99.99^{th} percentiles, respectively. We repeat the procedure 100 times with a 100-fold cross-validation, using scikit-learn 'KFold' without shuffling (Pedregosa et al., 2011).

Statistical significance. We assess the significance of our results across test folds ($k = 100$). To this end, we first average the brain scores within each brain region, as defined by the Destrieux Atlas parcellation (Destrieux et al., 2010). Then, we apply a Wilcoxon two-sided signed-rank test across folds to evaluate whether this average brain score is significantly different from zero. The p-values of the 75 brain regions were corrected for multiple comparison using a False Discovery Rate, (Benjamini/Hochberg) as implemented in MNE-Python (Gramfort et al., 2013). Non-significant p-values ($p \geq .05$) are masked in Figure 3.5.

3.1.6 Results

Phonological features. To isolate the sublexical speech representations, we compute the brain scores using a concatenation of three sets of features, *i.e.*, word rate, phone rate, and phone categories. These sublexical features lead to significant brain scores across the expected language networks and mainly peak within the bilateral superior temporal lobe, the temporo-parietal junction, the lateral intra-parietal sulcus, the infero-frontal cortex (IFG) as well as in the right motor cortex (Figure 3.5A and 3.6).

To isolate lexical and compositional representations, we focus the next analyses on the *gain* in brain scores obtained over those of sublexical features (*i.e.* to the increase of brain scores obtained with each feature set, as compared to the scores obtained with phonological features). For simplicity, the $\mathcal{R}$ scores reported in Figure 3.5, 3.6 and in the text below refer to this gain.

The brain scores corresponding to the lexical ($\mathcal{R}(X^{(0)})$), compositional ($\mathcal{R}(X^{(9)})$), syntactic ($\mathcal{R}(\overline{X^{(9)}})$ and semantic representations ($\mathcal{R}(X^{(9)}) - \mathcal{R}(\overline{X^{(9)}})$) of the ninth layer of GPT-2 are

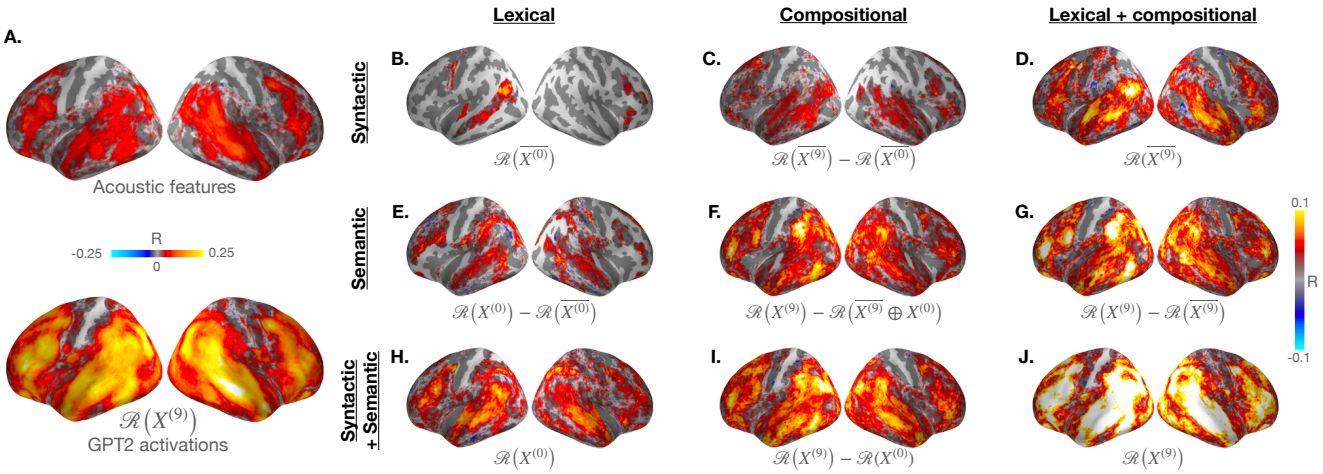


Figure 3.5: Results Decomposition of the brain scores of 345 subjects listening to narratives into their phonological (A) syntactic (B-D), semantic (E-G), lexical (B-H), compositional (C-I) components and their combinations (ten combinations in total). **A** Comparison between the brain scores of three phonological features (word rate, phone rate, and phone categories, on the top) and the brain scores of the activations extracted from the 9^{th} layer of GPT-2, when input with the same narratives (on the bottom). **B-J**. Brain scores decomposed into different sub-processes. To focus on language – and not low-level speech – processing, we display the *gain* in brain scores compared to the phonological features. For simplicity, the $\mathcal{R}$ values reported refers to this gain. Brain scores are computed for each fMRI voxel (averaged across subjects), on 100 splits of ≈ 2.5 min of audio stimulus. Non-significant brain regions are not displayed (.05 threshold), as assessed with a two-sided Wilcoxon test across splits, corrected for multiple comparison across the 75 regions of interest (cf. Section 6.3.5).

displayed in figures 3.5 and 3.6 (non-significant scores after correction for multiple comparisons across regions are masked).

Lexical features. The lexical representations of the brain have been repeatedly investigated through the lens of a word-embedding (Mitchell et al., 2008; Huth, de Heer, et al., 2016; Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022). Here, we replicate these analyses: GPT-2's word embedding $X^{(0)}$ leads to lexical brain scores significantly higher than sublexical features' in most of the language network, *i.e.* in the bilateral superior temporal lobe and the infero-frontal cortex (Figure 3.5H).

Lexical syntax. Do these brain scores result from semantic and/or syntactic representations? To tackle this issue, we compute brain scores from the word embeddings ($\overline{X^{(0)}}$) input with synthesized and syntactically-matched sentences: *i.e.* word sequences sharing the same syntax as the target sentence in the original Narratives corpus (Figure 3.5B). The results reveal sig-

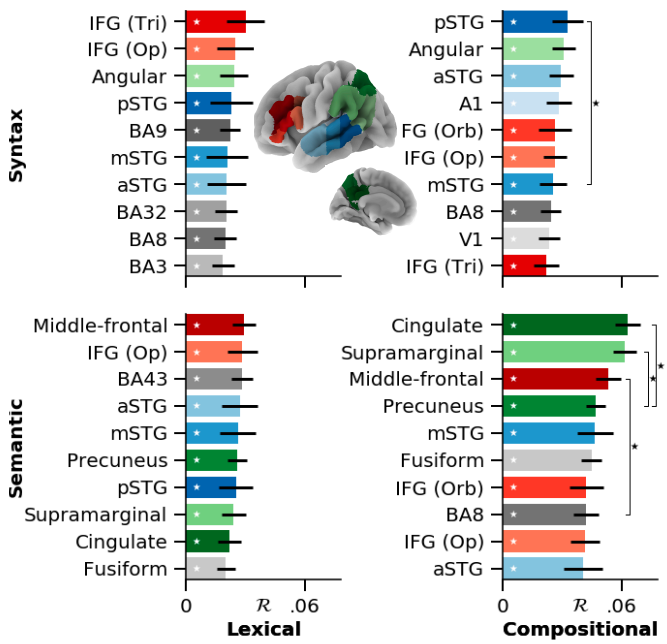


Figure 3.6: Brain scores for ten regions of interest. Same as Figure 3.5.BCEF, with voxel-averaged brain scores (after subtraction of phonological brain scores), for the top ten regions of interest of the left hemisphere (Appendix 6.3.5). Error bars are the standard-errors of the mean across the 100 cross-validation folds. Significance ('*') is assessed with a Wilcoxon test across folds, with $p < .05$ as a threshold.

nificant brain scores (*i.e.* higher than sublexical ones) in a distributed network including the infero-frontal cortex, the angular gyrus and the posterior superior temporal gyrus (Figure 3.6).

Lexical semantics. To identify the representations of lexical semantics, we compare the brain score obtained with the word embedding to those obtained with the embedding of lexical syntax ($\mathcal{R}(X^{(0)}) - \mathcal{R}(\overline{X^{(0)}})$ in Figure 3.5E). The resulting brain scores are significant mainly in the left hemisphere, and peak in the superior temporal gyrus, the infero-frontal cortex as well as in the precuneus and the tranverse temporal gyrus. These results are more modest than we anticipated given past work (Huth, de Heer, et al., 2016).

Compositional representations. Recent studies have shown that the contextual (*i.e.* deep) layers of language models better predict brain activity than word embedding (Jain & Huth, 2018; Jat et al., 2019; Toneva & Wehbe, 2019; Caucheteux & King, 2022). We replicate this result with a representative contextual layer of GPT-2 (layer 9 out of 12, Figure 3.5J): $\mathcal{R}(X^{(9)})$ almost doubles the brain scores obtained with the word embedding $\mathcal{R}(X^{(0)})$ in the bilateral temporal, infero-frontal and infero-parietal cortices.

Compositional syntax. Do these gains in brain score reflect compositional semantics and/or compositional syntax? To tackle this issue, we compare the brain scores obtained with the ninth layer of GPT-2 input with the syntax-matched synthesized sentences $\mathcal{R}(\overline{X^{(9)}})$, to the the brain scores obtained with the first layer of GPT-2, input with those same synthesized sentences $\mathcal{R}(\overline{X^{(0)}})$. The results show that the representations of compositional syntax are distributed over the bilateral temporal and infero-frontal cortices, and actually extend to a relatively large set of brain areas (Figure 3.5C-D). Overall, these results, although correlational, thus favor a distributed (Fedorenko et al., 2012) rather than a modular (Pallier et al., 2011; Friederici et al., 2000) view of syntax: both lexical and compositional syntactic effects do not appear to be confined within a single brain area.

Compositional semantics. Finally, we estimate the brain representations of compositional semantics by comparing the brain scores obtained with the syntactic representations $\mathcal{R}(\overline{X^{(9}})$ to those obtained with the "normal" activations $\mathcal{R}(X^{(9}})$, *i.e.* GPT-2's activations obtained with the same sentences as subjects heard. Again, the resulting effects proved to be remarkably distributed, and peaked in the cingulate, supramarginal, and middle-frontal cortex (Figure 3.5G). These brain scores appear to result from strictly compositional semantics: these effects remain significant even when we subtract away the contribution of lexical semantics (Figure 3.5E and 3.6).

Control 1: low-level linguistic properties. Do the syntactic representations evidenced above simply capture the length of sentences? To address this issue, we input the above analyses with i) random words sequences (*i.e.* non grammatical) and ii) random but well-formed sentences that have the same length as those of the Narratives corpus. The results show that neither of these two embeddings match the brain scores obtained with syntactic and/or semantic representations (Figure 3.7). Similarly, using the GPT-2 activations elicited by the sentences of the Narratives after a random word permutation leads to lower brain scores than our original analyses. Together, these results confirm that our decomposition of syntactic and semantic representations in the brain cannot be reduced to simplistic representations like bags of words and/or sentence length.

Control 2: generalisation to other layers and architectures. The above results are obtained using the ninth layer of GPT-2. We chose to study this model and this layer, because a) GPT-2, like the brain, processes words in a *causal* way, b) it is known to best predict brain responses

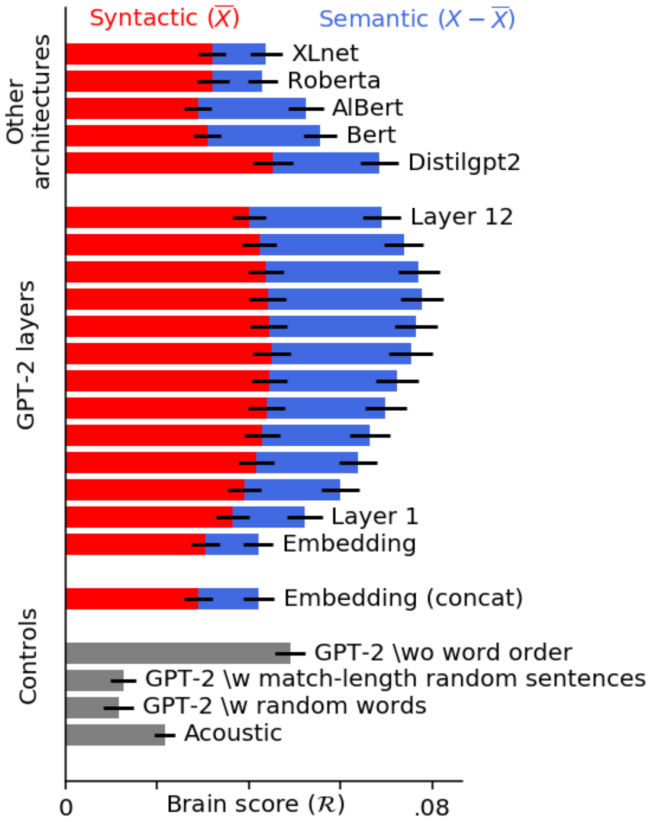


Figure 3.7: Generalisation to other layers and architectures In red, the brain scores of the syntactic embeddings ($\mathcal{R}(\bar{X})$) built out of GPT-2 layers (from the word embedding to layer 12), and the middle layer of five transformer architectures (top, cf. Appendix 6.3.1, $l = 2/3 \times n_{\text{layers}}$). In blue, the residuals of syntax ($\mathcal{R}(X) - \mathcal{R}(\bar{X})$) in the brain. Bottom, the brain scores of i) acoustic features (the concatenation of word rate, phoneme rate, phoneme stress and tone), GPT-2 activations induced ii) by random words sampled in the stimulus, iii) by sentences randomly sampled from Wikipedia, matching in length with the sentences of the stimulus, iv) by the actual sentences of stimulus, but with random word order in each sentence (Appendix 6.3.6.)

(Schrimpf et al., 2021; Caucheteux et al., 2022), c) its middle layers best encode complex semantic and syntactic properties (Jawahar et al., 2019; Manning et al., 2020). To test the generality of our study, we apply the same analyses to five other language transformers as well as to all of the layers of GPT-2 (Figure 3.7). The results generalize to each layer of GPT-2, and peak around layer 9. The five other transformers (for their middle layer $l = 2/3 \times n_{\text{layers}}$) result in similar, although significantly lower brain scores (Appendix 6.3.1).

3.1.7 Discussion

In the present study, we introduce a simple taxonomy and its associated method to decompose the distributed representations of language in brains and deep language models.

Our taxonomy capitalizes on classic linguistic proposals (Lycan, 2018; Givón, 2001; Chomsky, 2014) to offer precise definitions of lexicality, compositionality, syntax and semantics, which operate on *distributed* representations. Our results show that these four sets of linguistic features, typically theorized in terms of discrete symbols, can be, as long predicted (Smolensky, 1990), investigated in artificial and biological neural networks.

The present definitions remain imperfect. First, compositionality is often associated with specific properties that are not presently considered (e.g. systematicity and generalisation (Szabó, 2004; Hupkes et al., 2020; Baroni, 2020)). Furthermore, we here define semantics as the *residual* representations of any text embedding once syntactic representations have been removed. This proposal is very coarse: semantics is generally defined as the study of meaning (which is itself not easy to define). Yet, some language features like emotional value and textual style may arguably not "mean" anything, in that they do not necessarily refer to a state of the world and yet would be categorized as semantics according to our proposed taxonomy. In spite of these limits, the advantage of our framework is that it makes simple, precise and quantifiable predictions to investigate distributed linguistic representations in the human brain. Furthermore, the present framework is particularly versatile in that i) it can, in principle accommodate any natural sentences and ii) its conclusions can be refined with the development of better and/or more biologically-plausible models of language.

The present study follows suit with past research on naturalistic and thus poorly-controlled linguistic stimuli (Mesgarani et al., 2014; Huth, de Heer, et al., 2016; J. Brennan, 2016; J. R. Brennan & Hale, 2019; Stehwien et al., 2020; Gwilliams et al., 2020). While we replicate previous neuroscientific findings regarding lexical semantics (Figure 3.5E) (Huth, de Heer, et al., 2016) and lexical *vs* compositional processing in the brain (Figure 3.5.H,J) (Toneva & Wehbe, 2019; Schrimpf et al., 2021; Goldstein et al., 2022), our systematic decomposition of language representations brings new light on the brain bases of syntax (Figure 3.5.BCDFG). In addition, our approach diverges with and complements previous practices, consisting of carefully designed stimuli, typically matched for word length, word frequency (Kutas & Hillyard, 1980) and/or constituent size (Pallier et al., 2011; Ding et al., 2016), which becomes exponentially difficult when the number of variables to control increases (Hamilton & Huth, 2018). This change of paradigm has been empowered by the rise of high-performing language models: previous

research lacked a method to make single trial/single sentence predictions and could thus only compare the average activations across blocks of similarly constructed sentences. By contrast, modern language models offer the possibility to predict the representations of individual words and sentences (Hale et al., 2018; Toneva & Wehbe, 2019; Caucheteux & King, 2022; Schrimpf et al., 2021; Heilbron et al., 2022). Consequently, carefully-controlled experimental designs can now be relaxed to naturalistic settings, and allow one to refine her tests and hypotheses without having to conduct new (and arguably artificial) experiments.

The main drawback of such an uncontrolled setting is undoubtedly signal-to-noise ratio: like any bias/variance trade-off, relaxing the set of hypotheses that one can test in a given dataset reduces the probability of a successful finding. To accommodate this issue, we here opted to analyze the average brain signal across subjects. Even then, brain scores remain far from 100%. Given that the brain bases of languageare notoriously variable across individuals (Fedorenko et al., 2010) future works remain necessary to better account for the functional and anatomical variability across subjects.

Thanks to machine learning, our method sheds new light on the neural bases of language in general, and of syntactic processes in particular. First, it supplements previous work on the neural basis of lexical (Friederici et al., 2000; Mitchell et al., 2008) and compositional representations of language (Pallier et al., 2011; Nelson et al., 2017; Fedorenko et al., 2012; J. R. Brennan & Pylkkänen, 2017): syntactic processes, in particular, appear to be linked to a remarkably wide-spread *distribution* of activation in the language networks. This result favours a distributed (Fedorenko et al., 2012) as opposed to a modular (Pallier et al., 2011; Friederici et al., 2000) view of syntactic processes. Second, our study highlights the remarkably-large recruitment of compositional semantics – an observation that strengthens and extends what had already been reported at the lexical level (Huth, de Heer, et al., 2016). Overall, these results thus reinforce the idea that speech comprehension results from the coordination of a huge cortical network. While its functional principles remain largely unexplored, the similarity between the human brain and deep language models offers a new and powerful mean to understand the laws of language.

3.2 Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects

3.2.1 Abstract

A popular approach to decompose the neural bases of language consists in correlating, across individuals, the brain responses to different stimuli (e.g. regular speech versus scrambled words, sentences, or paragraphs). Although successful, this 'model-free' approach necessitates the acquisition of a large and costly set of neuroimaging data. Here, we show that a model-based approach can reach equivalent results within subjects exposed to natural stimuli. We capitalize on the recently-discovered similarities between deep language models and the human brain to compute the mapping between i) the brain responses to *regular* speech and ii) the activations of deep language models elicited by *modified* stimuli (e.g. scrambled words, sentences, or paragraphs). Our model-based approach successfully replicates the seminal study of (Lerner et al., 2011), which revealed the hierarchy of language areas by comparing the functional-magnetic resonance imaging (fMRI) of seven subjects listening to 7 min of both regular and scrambled narratives. We further extend and precise these results to the brain signals of 305 individuals listening to 4.1 hours of narrated stories. Overall, this study paves the way for efficient and flexible analyses of the brain bases of language.

3.2.2 Introduction

One of the most successful paradigms to decompose the brain bases of language consists in correlating the brain responses of multiple subjects listening to the same carefully controlled stimuli (J. Brennan et al., 2012; Fedorenko et al., 2016; Blank et al., 2016; Mollica et al., 2019). In particular, (Lerner et al., 2011) recorded subjects with functional magnetic resonance imaging (fMRI) while they listened to a story whose (1) sounds (2) words, (3) sentences or (4) paragraphs were scrambled, as well as (5) to the regular version of the story (Figure 3.8A). The authors then estimated the Inter Subject Correlation (ISC), i.e. the correlation between i) the brain activity of a voxel in response to one scrambling condition and ii) the brain activity of a voxel averaged across all other subjects, in response to the same scrambled stimulus (Figure 3.8B). While successful, this 'model-free' approach is costly: it requires $n_{\text{subjects}} \times n_{\text{conditions}}$ acquisitions of brain activity in response to the same variably scrambled stimuli.

Here, we investigate whether and how a model-based approach can replicate Lerner et al.'s findings, even if we only have access to the recordings elicited by the regular story in a single

subject. We further apply the method to extend Lerner et al's results to a large dataset of 305 individuals.

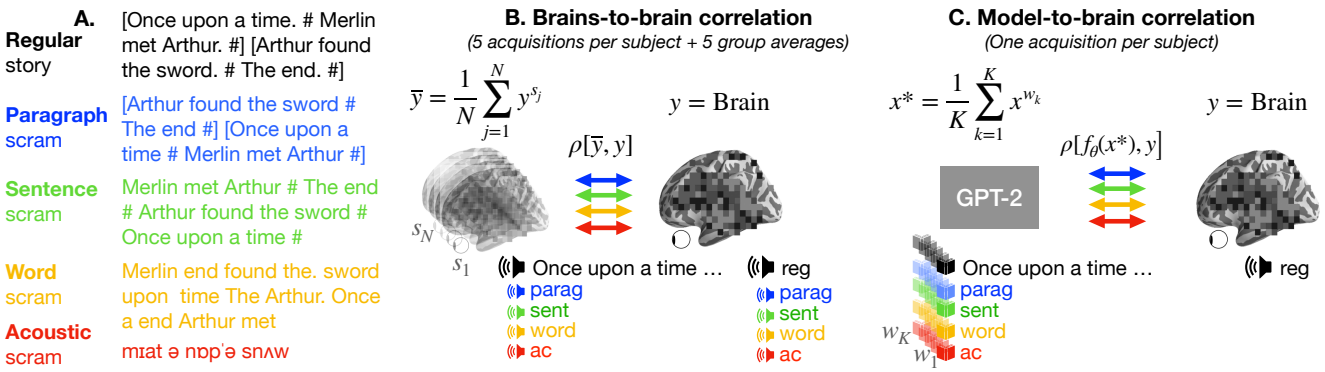


Figure 3.8: Objective and methods A. In Lerner et al.'s seminal study, each subject is presented successively with i) a 7 min long story (black), ii) the same story after its paragraphs (blue) iii) sentences iv) words (orange) or iv) acoustics (red) has been scrambled. **B.** For each condition, subject and voxel, the authors compute the inter-subject correlation (ISC), i.e the correlation ρ between i) the brain of the current subject y and ii) the average brain signals of the *other* subjects $\bar{y}$. This method allows to decompose the hierarchy of language processing in the brain, from the acoustic to the paragraph level. **C.** We aim to replicate the results of Lerner et al. using only the recordings induced by the regular story (black). To this aim, we scramble, not the stimulus of the subject, but the inputs of a deep language model (GPT-2). For each condition (word, sentence or paragraph), we extract the corresponding activations x^* averaged over K random scrambles. We then compare the brain signals of the current subject y with the activations x^* elicited by the scrambled texts, after a linear transformation f_θ that maps x^* onto a brain-like space. Because GPT-2 is not trained to process waveform, we use the phonemes, stresses and tones of the stimulus instead of x^* for the acoustic condition.

3.2.3 Methods

First, we formalize the 'model-free' and 'model-based' approaches in the context of narrative listening, and explicit the link between the two.

Definitions Let's define

- w = ('Once', 'upon', ... , 'The', 'end.') the regular story. Ω the story's vocabulary.
- $w_{|\text{sound}}, w_{|\text{word}}, w_{|\text{sent}}, w_{|\text{parag}}$ the story scrambled at the acoustic, word, sentence and paragraph level, respectively, following the setting of Lerner et al. (cf. Appendix 6.4.2 for the scrambling paradigm).

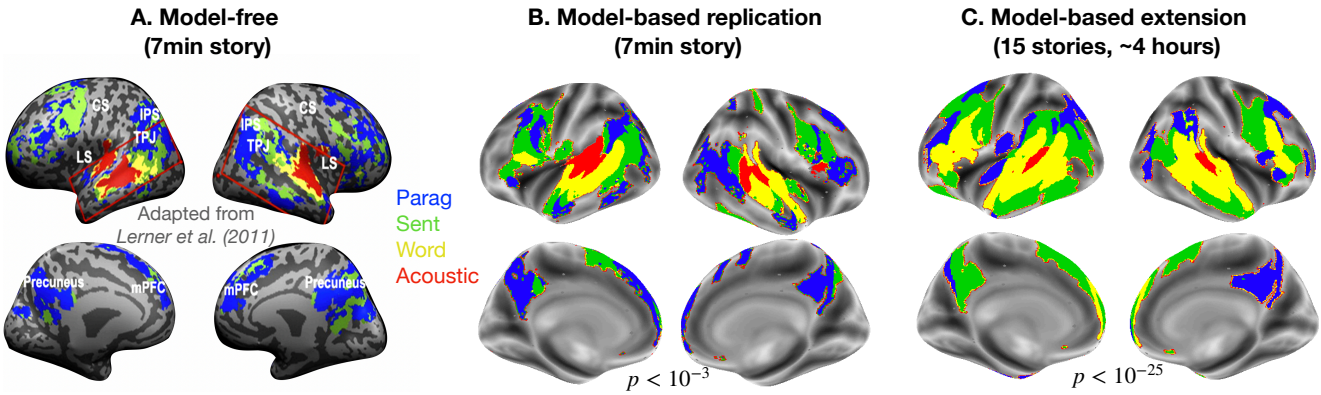


Figure 3.9: Results. Following Lerner et al's, a brain region is considered to process 'acoustic' level information if its acoustic score (either brains-to-brain or model-to-brain correlation) is significant (red). It is considered to process 'word'-level (yellow) if its word score is significant but not its acoustic one – and similarly for 'sentence' (green) and 'paragraph' (blue). **A.** Adapted from (Lerner et al., 2011). Labels are based on the brains-to-brain correlation scores (Figure 3.8B) averaged over seven subjects listening to a 7 min story. **B.** Labels are based on the model-to-brain scores (Figure 3.8C), averaged over 75 subjects listening to the same 7 min story. Significance is inferred using a Wilcoxon test across subjects, corrected with False Discovery Rate (FDR) across the 465 brain regions in each hemisphere (*cf.* Appendix 6.4.4), with a significance threshold of $p < 10^{-3}$ (*cf.* Appendix 6.4.5). **C.** Same as B., but on the brain of 305 subjects listening to 4 hours of 15 audio stories (including the 7 min one). Because of the large number of subjects, the significance threshold is set to $p < 10^{-25}$.

- $\mathcal{B} : \Omega^M \to \mathbb{R}^T$: the function returning the brain recordings of length T time samples (*i.e.*, the number of fMRI pulses) induced by a sequence of M words.
- $\mathcal{A} : \Omega^M \to \mathbb{R}^{M \times D}$ the function returning the activations of a deep language model induced by a sequence of M words.
- $y \in \mathbb{R}^T$ the brain recordings of one subject elicited by w, recorded at one voxel. Here, $\mathcal{B}(w) = y$.
- $y_{|\text{sound}}, y_{|\text{word}}, y_{|\text{sent}}, y_{|\text{parag}}$ the recordings elicited by the scrambled versions of w.
- $\rho : \mathbb{R}^T \times \mathbb{R}^T \to \mathbb{R}$, Pearson's correlation

For clarity, we describe below the model-free and model-based approaches for the *sentence* condition. The same methods can be used for the sound, word and paragraph conditions.

Model-free analysis Lerner et al. do not have a model of how the brain should react to sentences. Instead, they assume that the neural signature of sentence-level processing corresponds

to the brain response shared across all subjects listening to scrambled sentences $w_{|\text{sent}}$. They thus compute the 'ISC score' for each subject, *i.e.,* the correlation between i) the brain response to the scrambled story $w_{|\text{sent}}$ of a given subject ($y_{|\text{sent}}$) and ii) the brain response to the same stimulus averaged across all other subjects

$$R = \rho\left(y_{|\text{sent}}, \overline{y_{|\text{sent}}}\right) \ . \tag{3.3}$$

This approach boils down to a leave-one-subject-out cross-validation, using Pearson correlation as evaluation metric and the average population response as estimator.

Model-based analysis Here, we propose a model-based analysis to circumvent the need for .

To eliminate the need for , we capitalize on the recent findings that deep language models tend to linearly predict brain responses to language (Jain & Huth, 2018; Gauthier & Levy, 2019; Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022). We can thus assume that the average brain response ($\overline{\mathcal{B}}$) can be well approximated by f_θ, a linear function that maps the deep language model to the brain response. *i.e.,*

$$i) \quad \overline{\mathcal{B}} \approx f_\theta \circ \mathcal{A} \ .$$

In practice, the coefficients θ of f_θ are estimated using ridge regression. Finite Impulse Response functions are employed to allow the activations of the deep language model of length M (number of words) to map onto the slow and delayed brain recordings of length T (number of pulses) (cf. Appendix 6.4.3).

First, we separate the representation of the sentence from that of its context. To this end, for each sentence s of w, we note Ω_s the set of sequences ending with s, and whose preceding context is random. The representation of s without context, is, by construction, also the sentence representations of all sequences $w' \in \Omega_s$. Thus, if we denote this common representation, the brain response of one subject to a sequence w' can be modeled as

$$\forall w' \in \Omega_s, \quad \mathcal{B}(w') = y_s^* + \varepsilon_{w'} \ , \tag{3.4}$$

with $\varepsilon_{w'}$ the context-dependent contribution to $\mathcal{B}(w')$. Assuming it is a zero-mean random perturbation we have:

$$\mathbb{E}_{w'}\left[\mathcal{B}(w')\right] = y_s^* \ , \tag{3.5}$$

with w' sampled uniformly in Ω_s. Importantly, we do *not* assume that words are independent of their context but that the *shufflings* defined for each sentence are independent of one another. This statement is true by construction: shuffled contexts are realizations of a uniform sampling of permuted texts. Furthermore, the assumption that activations of shuffled versions of the same context have a zero-mean is not critical: assuming a constant mean would not alter the methods and results, because the final metrics (Pearson correlation) is invariant to such constant.

Similarly, we can retrieve x_s^*, the context-independent representation of a particular sequence s in a deep language model

$$\mathbb{E}_{w'}\left[\mathcal{A}(w')\right] = x_s^* \ . \tag{3.6}$$

In practice, it is approximated with an average over K *i.i.d.* samples:

$$x_s^* \approx \frac{1}{K}\sum_{k=1}^{K}\mathcal{A}(w_k) \ , \tag{3.7}$$

where $w_1, \ldots, w_K$ are sentences uniformly sampled in Ω_s.

3.2.4 Experiment

To test our model-based approach, we first apply it to the fMRI responses of 75 subjects listening to the same 7 min story analysed in Lerner et al (Nastase et al., 2020)[4]. Thus, for each condition (word, sentence and paragraph), subject and voxel, we compute the model-to-brain correlation $R = \rho\big(y, f_\theta(x^*)\big)$.

The extraction of the fMRI signals y, and the estimation of the mapping function f_θ are standard and thus detailed in Appendices 6.4.1 and 6.4.3. To estimate context-free representations, we i) scramble the stimulus at the word, sentence or paragraph level, ii) extract the corresponding activations x from a deep language model, and iii) compute x^*, as detailed below.

Scrambling the stimulus at the word, sentence and paragraph level Words and sentences of the stimulus are delimited using Spacy tokenizer (Honnibal et al., 2020). Note that punctuation marks are not considered as words (*e.g.*, 'time.' forms *one* token, not two). We define paragraphs as contiguous chunks of eight sentences. To 'scramble' a sequence at the word (resp. sentence, paragraph) level, we uniformly shuffle the indices of its words (resp. sentences, paragraphs) and form the new sequence accordingly.

[4] http://datasets.datalad.org/?dir=/labs/hasson/narratives

Extracting deep models' activations For each version of the scrambled stimulus, we extract the activations from GPT-2 ($\mathcal{A}$), a deep neural language model trained to predict a word given its past context. GPT-2 consists of 12 transformer layers of dimensionality 768, 8 heads, and has 1.5 billion parameters in total. We use the model provided by Huggingface (Wolf et al., 2020), trained on a dataset of 8 million web pages.

To extract the activations elicited by a sequence w of M words from layer l, we proceed as follows: we tokenize the sequence into sub-words called "Byte Pair Encoding" (BPE) (Sennrich et al., 2016) using the GPT-2 tokenizer provided by Huggingface. Then, we feed the network with the M' BPE tokens ($M' \geq M$, up to 256 tokens in memory) and extract the corresponding activations from layer l, of shape $(M' \times D)$ with $D = 758$. Then, we sum the activations over the BPEs of each word to obtain a vector of size $(M \times D)$.

All our analyses are based on the eighth layer of GPT-2. We choose GPT-2 because it has been shown to best encode the brain activity elicited by language stimuli (Caucheteux et al., 2021a; Schrimpf et al., 2021). We choose its eighth layer because the intermediate layers of transformers have shown to encode relevant linguistic features (Jawahar et al., 2019; Manning et al., 2020) and to better encode brain activity than input and output layers (Caucheteux & King, 2022; Toneva & Wehbe, 2019). Our results successfully generalize to two other architectures as well as to the other intermediate layers of GPT-2 (Appendix 6.4.6).

Computing x^* for the word, sentence and paragraph conditions For each of the word, sentence and paragraph conditions, we compute x^*: a context-free representation of x. In short, x^* are the activations of GPT-2, averaged over several scrambled contexts. For clarity, we focus on the sentence level to detail the approach.

To build the sentence-level representation x^* of the stimulus, we use the approximation introduced in equation (3.7). For each sentence s of one story w, we i) generate K=10 sequences ending with s, but with scrambled previous context. The scrambled context is uniformly sampled from the other sentences in the same story w. Then, ii) we extract the K corresponding activations from GPT-2 (as described in the previous section) and iii) average the activations across the K samples. GPT-2 activations are extracted for each word. Thus, for each of the M_s words of sentence s, we obtain a vector x_s^* of shape $M_s \times D$. We concatenate these vectors to obtain x^*, a sentence-level representation of the whole story w, of shape $M \times D$. This method is adapted from (Caucheteux et al., 2021a), in which we computed the average over GPT-2's activations to extract syntactic representations from the input sequence.

Acoustic features GPT-2 takes words as input and not sounds. To build x^* at the acoustic level, we simply use non-contextual acoustic features: the word rate ($D = 1$), phoneme rate ($D = 1$) phonemes, stress, and tone (categorical, $D = 117$). For the latter, we use the annotations provided the original Narratives dataset (Nastase et al., 2020).

3.2.5 Results

The results are displayed in Figure 3.9B. The hierarchy of temporal receptive fields (TRFs) typically associated with acoustic, word, sentence and paragraph processing along the temporo-parietal axis is remarkably well replicated in both hemispheres (Figure 3.9B). Notably, both the model-free and model-based methods evidence that the precuneus, the superior frontal gyrus and sulcus are characterized by sentence- and paragraph-level TRFs (Figure 3.9A and B).

Our results differ from Lerner et al.'s in several ways. First, the acoustic TRFs are slightly more inferior with the model-based method. Second, frontal regions are detected to be associated not only with sentences and paragraphs, but also with words (consistent with (Huth, de Heer, et al., 2016; Caucheteux et al., 2021a; Goldstein et al., 2022)). Given that Lerner et al's dataset is not public, it is difficult to quantify these differences and determine whether they reflect an improved sensitivity, or, more simply, inter-individual differences.

Our model-based method can, in principle, be applied to any natural stories. To test this prediction, we extend our analyses to 305 subjects listening to 4.1 hours of fifteen narratives (Figure 3.9C). Our model-based approach recovers the hierarchy of TRFs, and further reveals additional word- and sentence-level representations in the precuneus and prefrontal regions.

3.2.6 Discussion

Here, we leverage the modeling power of deep language models to show that the seminal results of Lerner et al. can be retrieved without having subjects listening to multiple scrambled stimuli. Critically, we formalize the assumptions under which 'model-based' and 'model-free' approaches can be linked (Lerner et al., 2011).

Our model-based method recovers the hierarchy of TRFs evidenced by Lerner et al., in the brain of an unusually large cohort of 305 subjects. Thus, our study complements the recent work of (Jain & Huth, 2018; Toneva & Wehbe, 2019; Toneva, Mitchell, & Wehbe, 2020a) who predict brain responses to speech from language models input with variably-long contexts. Specifically, we show that previous model-based results unravel the same mechanisms that was previously identified with model-free approaches.

The replication is not perfect: the acoustic and word TRFs slightly differ between the two methods. This may be explained by individual subject's variability, which is only captured by the model-based approach. Further research, using the non-public data from Lerner et al. should investigate these remaining differences.

In line with previous work (J. Brennan, 2016; J. R. Brennan & Hale, 2019; Gauthier & Levy, 2019; Schrimpf et al., 2021), our study demonstrates that deep neural networks build constructs that predict brain activity, accurately enough to recover the hierarchy of language processing in the brain. The success of replication thus reinforces the idea that naturalistic stimuli and deep neural networks form a powerful couple to study the neural bases of language (Hamilton & Huth, 2018).

3.3 Toward a realistic model of speech processing in the brain with self-supervised learning

3.3.1 Abstract

Several deep neural networks have recently been shown to generate activations similar to those of the brain in response to the same input. These algorithms, however, remain largely implausible: they require (1) extraordinarily large amounts of data, (2) unobtainable supervised labels, (3) textual rather than raw sensory input, and / or (4) implausibly large memory (e.g. thousands of contextual words). These elements highlight the need to identify algorithms that, under these limitations, would suffice to account for both behavioral and brain responses. Focusing on speech processing, we here hypothesize that self-supervised algorithms trained on the raw waveform constitute a promising candidate. Specifically, we compare a recent self-supervised model, wav2vec 2.0, to the brain activity of 412 English, French, and Mandarin individuals recorded with functional Magnetic Resonance Imaging (fMRI), while they listened to approximately one hour of audio books. First, we show that this algorithm learns brain-like representations with as little as 600 hours of unlabelled speech – a quantity comparable to what infants can be exposed to during language acquisition. Second, its functional hierarchy aligns with the cortical hierarchy of speech processing. Third, different training regimes reveal a functional specialization akin to the cortex: wav2vec 2.0 learns sound-generic, speech-specific and language-specific representations similar to those of the prefrontal and temporal cortices. Fourth, we confirm the similarity of this specialization with the behavior of 386 additional participants. These elements, resulting from the largest neuroimaging benchmark to date, show how self-supervised learning can account for a rich organization of speech processing in the brain, and thus delineate a path to identify the laws of language acquisition which shape the human brain.

3.3.2 Introduction

The performance of deep neural networks has taken off over the past decade. Algorithms trained on object classification, text translation, and speech recognition are starting to reach human-level performance (Xu et al., 2020). Furthermore, the *representations* generated by these algorithms have repeatedly been shown to correlate with those of the brain (Kriegeskorte, 2015; D. L. K. Yamins & DiCarlo, 2016; Kietzmann et al., 2018; A. J. Kell & McDermott, 2019; Cichy &

Kaiser, 2019; Toneva & Wehbe, 2019; Millet & King, 2021; Caucheteux & King, 2022), suggesting that these algorithms converge to brain-like computations.

Such convergence, however, should not obscure the major differences that remain between these deep learning models and the brain. In particular, the above comparisons derive from models trained with (1) extraordinarily large amounts of data (40GB for GPT-2 (Radford et al., 2019), the equivalent of multiple lifetimes of reading), (2) supervised labels, which is rarely the case for humans (e.g. (D. L. K. Yamins & DiCarlo, 2016)), (3) data in a textual rather than a raw sensory format, and/or (4) considerable memory (e.g., language models typically have parallel access to thousands of context words to process text). These differences highlight the pressing necessity to identify architectures and learning objectives which, subject to these four constraints, would be sufficient to account for both behavior and brain responses.

Here, we hypothesize that the latest self-supervised architectures trained on raw sensory data constitute promising candidates (Borgholt et al., 2022; Bardes et al., 2022; Baevski et al., 2020). We focus on wav2vec 2.0 (Baevski et al., 2020), an architecture that stacks convolutional and transformer layers to predict a quantization of the latent representations of speech waveforms. We train wav2vec 2.0 on 600 h of effective speech – a quantity roughly comparable to what infants are exposed to during early language acquisition (speech only makes up a small fraction of infants' daily experience) (Dupoux, 2018; Hart & Risley, 1992; Gilkerson et al., 2017).

We use standard encoding analyses (Naselaris et al., 2011; Huth, de Heer, et al., 2016; D. L. K. Yamins & DiCarlo, 2016; A. J. E. Kell et al., 2018) (Figure 3.10) to compare this model to the brains of 412 healthy volunteers (351 English speakers, 28 French speakers, and 33 Mandarin speakers) recorded with functional magnetic resonance imaging (fMRI) while they passively listened to approximately one hour of audio books in their native language (Nastase et al., 2020; Li et al., 2021) (8.5 hours of distinct audio materials in total).

To better understand the similarities between wav2vec 2.0 and the brain, we compare brain activity to each layer of this model, as well as to several variants, namely (1) a random (untrained) wav2vec 2.0 model, (2) a model trained on 600 h of non-speech sounds, (3) a model trained on 600 h of non-native speech (for example, a model trained on English speech and mapped onto the brain responses to French-speaking participants), (4) a model trained on 600 h of native speech (for example, a model trained on English speech and mapped onto the brain responses to English participants), and (5) a model trained directly on speech-to-text (i.e., a supervised learning scheme) on the native language of the participants.

Our results provide four main contributions. First, self-supervised learning leads wav2vec 2.0 to learn latent representations of the speech waveform similar to those of the human brain.

Second, the functional hierarchy of its transformer layers aligns with the cortical hierarchy of speech in the brain, and reveals the whole-brain organisation of speech processing with an unprecedented clarity. Third, the auditory-, speech-, and language-specific representations learned by the model converge to those of the human brain. Fourth, behavioral comparisons to 386 supplementary participants' results on a speech sound discrimination task confirm this common language specialization.

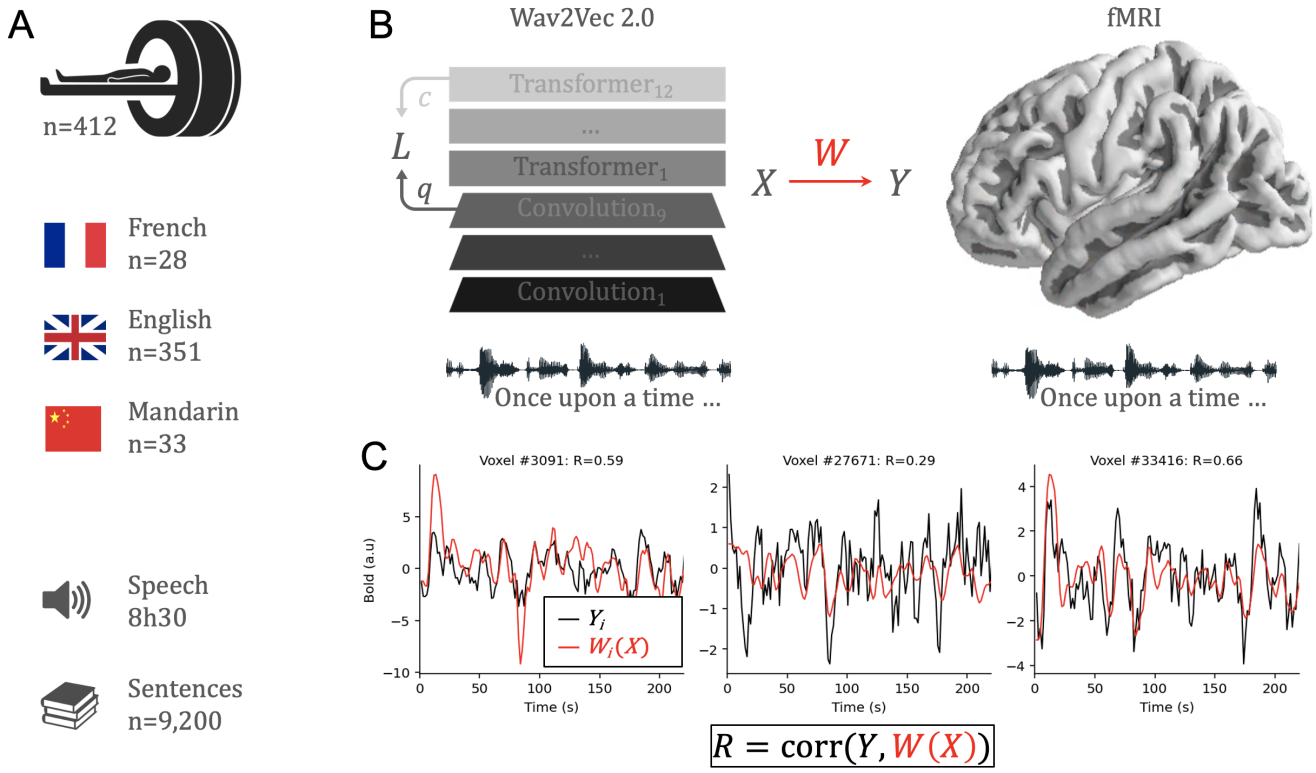


Figure 3.10: Comparing speech representations in brains and deep neural networks. A. We analyze the brain activity of 412 participants recorded with functional Magnetic Resonance Imaging (fMRI) while they passively listened to audio books in their native language (French, English or Mandarin). **B.** After training wav2vec 2.0 (Baevski et al., 2020) with self-supervised learning (L) over 600 h of unlabelled, effective speech, we extract its activations in response to the audio books that were presented to the participants. We assess the similarity between the activations of the model X and brain activity Y with a standard encoding model W (Nastase et al., 2020) evaluated with a cross-validated Pearson correlation R. **C.** Examples of the true BOLD response (black) and the predicted BOLD response (red) estimated from a linear projection of the model's activations in three voxels randomly selected from the 10^{th} percentile of best voxels identified by the noise ceiling analysis for the first 200 s of a representative story in the test set.

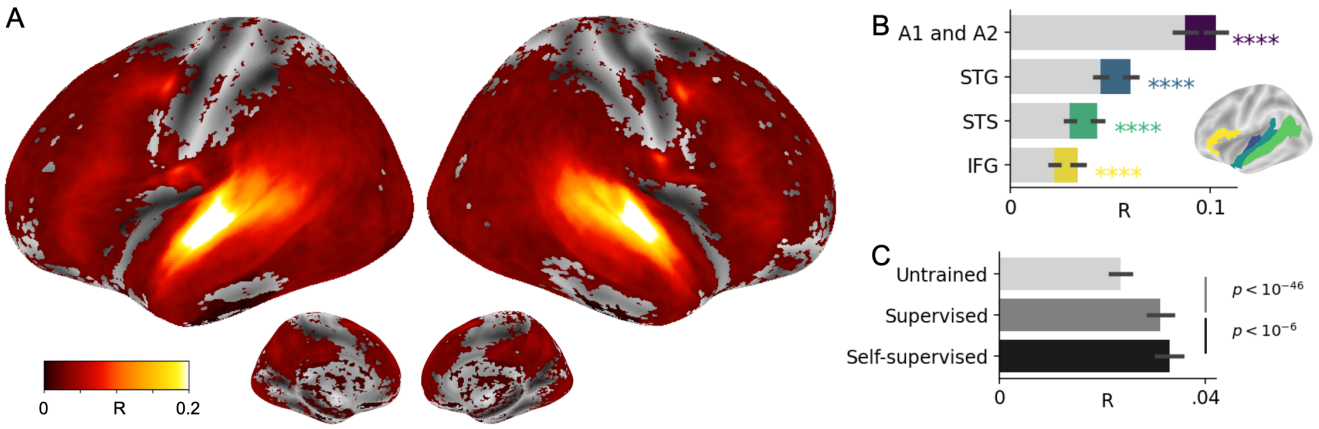


Figure 3.11: Self-supervised learning suffices for wav2vec 2.0 to generate brain-like representations of speech. A. Brain score (R) assessed for each subject and voxel independently, and here averaged across subjects for clarity. Only scores significantly above chance level, as assessed using a two-sided Wilcoxon test across subjects after correction for multiple comparison are color-coded ($p < 10^{-10}$). **B.** R scores for the same wav2vec2 model, averaged across subjects and voxels in four brain areas typically involved during speech processing (the primary and secondary auditory cortices, the superior temporal gyrus, the superior temporal sulcus, and the infero-frontal gyrus). In grey, the brain score obtained with a randomly initialized wav2vec 2.0 architecture. Error bars are the standard errors of the mean (SEM) across subjects. The stars indicate a significant difference between the random and trained model (all $p < 10^{-4}$). **C.** R scores of wav2vec 2.0 without training (top), trained with a supervised (middle) and self-supervised learning rule (bottom), on the same 600 hours of effective speech. Scores are averaged across subjects and voxels and error bars are SEM across subjects.

3.3.3 Methods

Models

We train several variants of wav2vec 2.0 (Baevski et al., 2020) from scratch on different speech datasets using two different learning objectives (a self-supervised and a supervised objective).

Architecture

Wav2vec 2.0 consists of three main modules. First, a feature encoder composed of seven blocks of temporal convolutions (output dimension 512) transforms the speech input S (raw mono waveform at 16 kHz) into a latent representation z (output dimension of 512, frequency 49 Hz, stride of 20 ms between each frame, receptive field of 25 ms). Second, a quantization module discretizes z into q, a dictionary of discrete and latent representations of sounds. Third, z is input to a "context network" consisting of 12 transformer blocks (model dimension 768, inner dimension 3072, and 8 attention heads), which together yield a contextualized embedding c, of

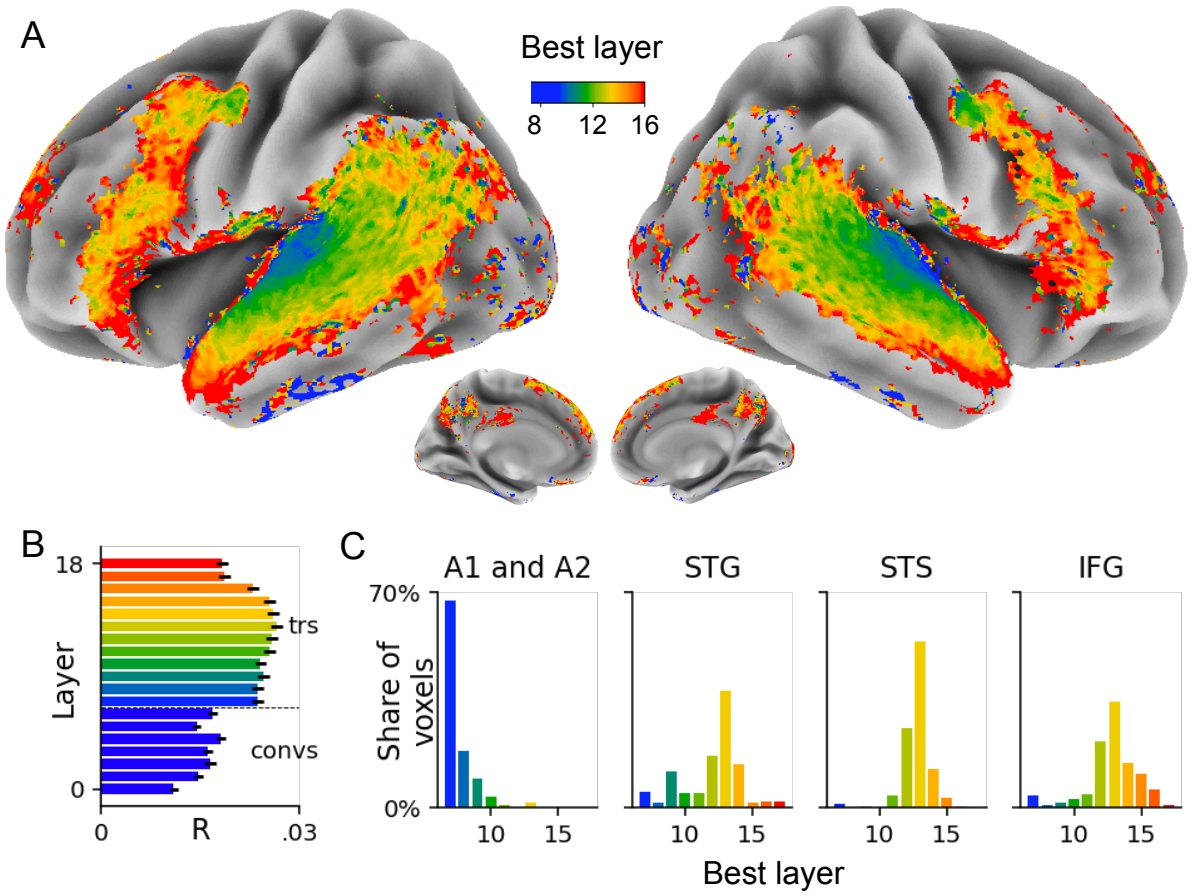


Figure 3.12: The functional hierarchy of wav2vec 2.0 maps onto the speech hierarchy in the brain. A. We compute the R score for each layer of wav2vec 2.0 separately and estimate, for each voxel, the layer with highest brain score on average across subjects. Only the voxels with significant brain scores are displayed ($p < 10^{-18}$). While the first transformer layers (blue) map onto the low-level auditory cortices (A1 and A2), the deeper layers (orange and red) map onto brain regions associated with higher-level processes (e.g. STS and IFG). **B.** Layer-wise R scores averaged across all voxels. Error bars are SEM across subjects. **C.** Proportion of voxels with most predictive layer (x-axis) in four regions typically involved in speech processing. While most voxels in the primary cortex are best predicted by the first layers of the transformer, higher-level brain areas are best predicted by deeper layers.

the same dimensionality of q.

Learning objective

Self-supervised learning. In this training paradigm, the model optimizes two losses. The first loss is contrastive and requires the model to predict the quantized representation q of some masked input using c, from a finite set of quantized representations drawn from the input sample. The second loss ensures that the quantized representations are diverse. See Section

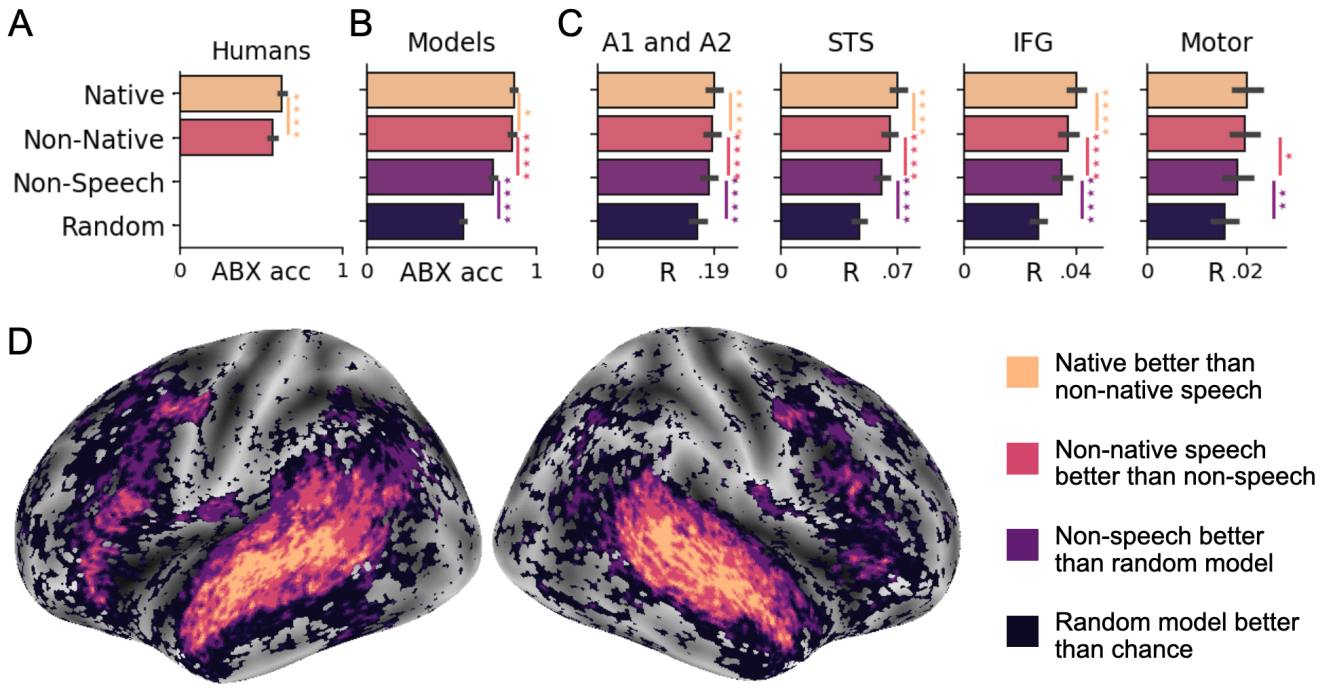


Figure 3.13: The specialization of wav2vec 2.0's representations follows and clarifies the acoustic, speech, and language regions in the brain. **A.** We first evaluate humans' language specificity by quantifying their ability to perceive phonemes of their native or non-native languages (Section 3.3.3) in a ABX matching-to-sample task (Schatz, 2016) (higher is better). As expected, humans are better at matching phonemes of their native language. **B.** Then, we train four wav2vec 2.0 models with self-supervised learning on four datasets – non-speech acoustic scenes, English, and French, and compute their ABX accuracy on the same speech datasets as humans. The 'random' model is wav2vec 2.0 without any training. **C.** Brain score (R) of each model (with an added model trained on Mandarin), averaged across voxels, in four regions of the brain (Section 3.3.3). **D.** Acoustic, speech and language specificity for each voxel. For instance, one voxel is considered specific to the 'native' model if its native R score is higher than its 'non-native' R score ($p < .05$). Only the voxels with significant R scores for the untrained model are displayed ($p < 10^{-18}$). Error bars are the SEM across phone pairs in A and B, and across subjects in C. The stars indicates a significant difference between two conditions (Section 3.3.3).

Appendix 6.5.2 and (Baevski et al., 2020) for details.

Supervised learning. In this training paradigm, the quantization module is discarded and a linear layer mapping c to phonemes is added at the end of the pipeline. The model is randomly initialized and all layers (including the feature encoder) are trained using a Connectionist Temporal Classification (CTC) (A. Graves, 2012) loss to perform phone recognition. For both training paradigms, we extract the activations of each layer from both the feature encoder (outputting z) and the context network (outputting c). We extract the representations of the

convolutional and transformer blocks using an input window of 10 s of raw waveform (stride = 5 s).

Training

Datasets. We successively train different wav2vec 2.0 models using each of four datasets: (i) the French and (ii) English CommonVoice corpora (Ardila et al., 2020), (iii) the MAGICDATA Mandarin Chinese Read Speech Corpus (Co., 2019), and (iv) a non-speech subset of the Audioset dataset (Gemmeke et al., 2017), which contains recordings of various acoustic scenes.

Preprocessing. All the audio datasets were randomly subsampled to have an approximate size of 600 hours, downsampled to 16 kHz and converted to mono with the Sox software[5]. We randomly split the datasets into a training (80%), a validation (10%) and a test set (10%). The audio recordings we use from the Audioset dataset are filtered so that they do not contain speech or any sounds produced by humans, such as laughter or singing. For the speech datasets, we also use their corresponding annotations (in the supervised settings). We phonemize these annotations using eSpeakNG[6]. The number of different phoneme symbols in these annotations is similar for French (32), English (39), and Mandarin Chinese (33).

Implementation. We train all of our models using the fairseq implementation of wav2vec 2.0[7] using default hyperparameters. We also analyze a model whose parameters were randomly initialised ("untrained" model).

We use self-supervised learning to train four models: three on the speech datasets (French, English, and Mandarin) and one on the acoustic scenes dataset. In each case, the training was performed using the same configuration file (namely, the base configuration provided in the fairseq repository for pretraining wav2vec 2.0 on LibriSpeech (Panayotov et al., 2015)). We train the models for 400k updates and select the ones with the best validation loss.

We also use the supervised training paradigm to train three models, on the French, English, and Mandarin datasets, respectively. Each training was performed using the same configuration file, which was identical to the configuration provided in the fairseq repository for fine-tuning wav2vec 2.0 on the 960 hour Voxpopuli corpus (C. Wang et al., 2021), except that parameters were not frozen (`freeze_finetune_updates`= 0) and learning was performed on all parameters

[5] `http://sox.sourceforge.net/`

[6] `https://github.com/espeak-ng/espeak-ng`

[7] `https://github.com/pytorch/fairseq/tree/main/examples/wav2vec`

of the models using the CTC loss (`feature_grad_mult`= 0.1). We train the models for 400k updates and we use the ones with the best word error rate (WER) on the validation set. The French model obtains 13.9 WER, the English model 28.6 WER, and the Mandarin model 4.6 WER, on their respective test sets.

Functional MRI

We analyse a composite set of fMRI recordings aggregated from the *Little Prince* (Li et al., 2021) and the *Narratives* public datasets (Nastase et al., 2020).

Narratives. This dataset[8] contains the fMRI recordings of 345 native English-speaking participants listening to English narratives (4.6 hours of unique audio in total). The participants listened to different stories varying from 7 to 98 min (mean = 26 min). Following (Nastase et al., 2020), we (1) focus on fifteen representative stories and ignore the narratives that have been modified by scrambling and (2) exclude eight participants because of noisy recordings. Overall, this selection results in a curated dataset of 303 participants listening to fifteen stories ranging from 3 min to 56 min, for a total of 4 hours of unique audio (36,018 words from a vocabulary of 4,004 unique words).

The Little Prince. This dataset[9] contains fMRI recordings of 48 English native speakers, 33 Mandarin native speakers, and 28 French native speakers listening to *The Little Prince* in their respective native language. The experiment itself was divided into nine runs of approximately 10 min of passive listening. For each language condition, the story was read by a single native speaker. The English, Mandarin, and French audiobooks last 94, 90 and 97 minutes respectively.

Preprocessing. For Narratives, we did not perform additional preprocessing: we use the public preprocessing of the dataset already projected on the surface space ("fsaverage6") without spatial smoothing (labelled "afni-nosmooth" in the data repository). In contrast, the *Little Prince* dataset is only provided in a volumetric space. Consequently, for each language condition separately, we subselected the cortical voxels by computing a brain mask using the average of all participants' fMRI data realigned onto a common template brain via Freesurfer (Fischl, 2012). These voxels are then projected onto a brain surface using nilearn's `vol_to_surf` function with defaults parameters (Abraham et al., 2014) and a 'fsaverage6' template surface

[8] `https://openneuro.org/datasets/ds002345`

[9] `https://openneuro.org/datasets/ds003643/versions/1.0.4`

(Fischl, 2012). For both *Narratives* and *The Little Prince*, fMRI signals are normalized across the time dimension to have a mean of 0 and a variance of 1, for each participant, surface voxel and session independently.

Brain parcellation. For the purposes of certain analyses, we group the fMRI voxels into regions of interest using the Destrieux Atlas (Destrieux et al., 2010). This parcellation results in 75 brain regions in each hemisphere. For simplicity, we label the regions as follows: A1 and A2 represents Heschl gyrus, which is the anatomical location of the primary and secondary auditory cortices, STG and STS are the superior temporal gyrus and sulcus, and IFG is the inferior frontal gyrus.

Brain score (R)

To quantify the similarity between the network's activations X and the brain recordings Y, we use a standard linear encoding model (Huth, de Heer, et al., 2016; D. L. K. Yamins & DiCarlo, 2016). For each subject, we split the data into train and test sets using a five-fold cross-validation setting. For each train split, a linear mapping W is fitted to predict the brain response Y_{train} given X_{train}. W combines a temporal alignment function with fixed weight, and a trained penalized linear regression.

Temporal alignment. The sampling frequency of the model's activations (between 49 and 200 Hz) differs from the sampling frequency of fMRI BOLD signals (0.5 Hz). Furthermore, the BOLD signals have delayed responses spanning over several seconds. Thus, we first convolve the model activations with a standard hemodynamic response function (HRF) using nistats (Abraham et al., 2014) `compute_regressor` function with the 'glover' model and default parameters. This results in the convolved activations X'_{train} with the same sampling frequency as the fMRI Y_{train} (see Appendix 6.5.3).

Penalised linear regression. Once temporally aligned, we fit an ℓ_2-penalised linear regression that predicts the brain signals Y_{train} given the activations X_{train}. We use the `RidgeCV` function from scikit-learn (Pedregosa et al., 2011), with the penalization hyperparameter λ varying between 10 and 10^8 (20 values scaled logarithmically) chosen independently for each dimension with a nested cross-validation over the training set (see Appendix 6.5.4).

Evaluation. We evaluate the linear mapping W on the held out sets by measuring Pearson's correlation between predicted and actual brain responses: $R = \text{corr}\big(Y_{\text{test}}, W \cdot X_{\text{test}}\big)$. Finally, we average the correlation scores across test splits to obtain the final "brain score". To report the average layer k^* with the highest brain score for each voxel (Figure 3.12), while being robust to regression-to-the-mean biases, we first find the best layer k_s for each participant s and each voxel independently and then compute a circular mean across the $N = 412$ participants and the $K = 19$ layers: $k^* = \text{angle}\left(\frac{1}{N}\sum_{s=1}^{N}\exp\left(\frac{2i\pi k_s}{K+1}\right)\right)$

Statistics. We assess the reliability of brain scores with second-level analyses across participants thanks to a Wilcoxon signed-rank test across participants. Thus, the resulting p-values are not affected by fMRI auto-correlation within participants. We perform statistical correction for multiple comparisons with Benjamini–Hochberg False Discovery Rate (FDR) across voxels (Benjamini, 2010).

Behavioral experiment

To compare the phonetic representations of our models to those of humans, we compare the forced-choice discrimination judgements of online participants[10] to an analogous method applied to wav2vec 2.0 (Schatz, 2016). Specifically, for each triplet of sound "ABX", participants judged whether the stimulus X was more similar to A or B. Analogously, we computed the Euclidean distance in the most discriminative layer of wav2vec 2.0 (here transformer layer 5) to determine whether X was closer to A or B. Additional data, analyses and model-human comparison can be found in (Millet & Dunbar, 2022). We focus on the French and English stimuli, which represent ≈6,000 ABX triplets (testing 508 English and 524 French phone pairs), with 386 participants in total (193 from each language group).

In Figure 3.13-A, we report the ABX accuracy of English- and French-speaking participants in both their native and non-native language (either English or French). We first average results per phone pair, and then average over phone pairs to obtain the ABX discrimination accuracy. Similarly, in Figure 3.13-B, we compute the ABX accuracy of our wav2vec 2.0 models on the same evaluation sets as the participants, using the parameters described in (Millet & Dunbar, 2022). English and French models are evaluated on the same ('native') or different ('non-native') language stimuli as their training. The random and non-speech models are evaluated on both French and English speech stimuli.

[10] https://docs.cognitive-ml.fr/perceptimatic/

3.3.4 Results

Wav2vec 2.0 maps onto brain responses to speech. We estimate whether the activations of wav2vec 2.0 models linearly map onto the human brain activity of 412 individuals listening to audio books in the fMRI scanner. For this, we first independently train three models with 600 h of French, English, or Mandarin, respectively, and compute the brain scores (R) with the corresponding participants. Specifically, we (1) convolve the activations (X) of the model with a hemodynamic response function (HRF), (2) train a ℓ_2-penalized linear regression on a training split to map them to brain activity Y, and (3) compute the Pearson correlation coefficient between (i) the true fMRI activity and (ii) the predicted activations on a test split. The models' activations significantly predict brain activity in nearly all cortical areas, reaching the highest R scores in the primary and secondary auditory cortices (Figure 3.11-A B). These scores are significantly higher than those obtained with a randomly initialised model ($p < 10^{-50}$ on average across voxels), and this comparison is robust across language groups (all $p < 10^{-5}$).

Comparison of self-supervised to supervised models. Does self-supervision reach representations that are as brain-like as those obtained with supervised learning? To address this issue, we trained wav2vec 2.0 with an alternative, supervised objective, namely, predicting phonetic annotations from the same 600 hours of effective speech sounds. We then implemented the R score analyses described above. The results show that self-supervised learning in fact leads to modestly but significantly better R scores than supervised learning (Figure 3.11-C): $\Delta R = 0.002, p < 10^6$.

The hierarchy of wav2vec 2.0 maps onto the hierarchy of the cortex. To compare the speech hierarchy in the brain with the functional hierarchy learned by wav2vec 2.0, we evaluate the R score of each layer of the model (Figure 3.12). First, we observe that convolutional layers are less predictive than transformer layers. Second, within the transformers, the hierarchy of representations aligns with the expected cortical hierarchy (Hickok & Poeppel, 2007): while low-level areas (A1, A2) are best predicted by the first transformer layers, higher level areas (IFG, STS) are best predicted by deeper layers. Remarkably, this hierarchy extends to supplementary motor and motor areas in both hemispheres (Figure 3.12-A).

Language specificity in phone discrimination tasks. The acoustic features underlying speech (fricatives, vowels, and so on) may also characterize non-speech sounds (the sound of tree leaves in the wind, of a stone falling, and so on). Does the model show commonalities merely with

general auditory processing in the brain, or does it capture speech-specific processing? If so, does it show commonalities with brain representations that are specific to the native language of the participants, or merely to general speech processing? We first evaluate the specialization of humans' perception to their native language using an ABX behavioral task (Section 3.3.3). Specifically, we compare 386 French and English participants on their ability to distinguish native and non-native phones. As expected (Bohn, 2017; Kuhl et al., 2005), participants were better at discriminating native sounds than non-native ones (across phone pairs: $p < 10^{-18}$, Figure 3.13-A). Second, applying the same test to our self-supervised French and English models shows that, like humans, models best discriminate sounds from their 'native' language (i.e., the French model better distinguishes French stimuli than English ones, across phone pairs, and vice versa: $p < 0.05$). Interestingly, the ABX accuracy of the model is significantly higher than participants'. This quantitative difference may be partially explained by the fact that participants – and online participants in particular – undergo fluctuating attention, and adopt strategies which can negatively impact performance (Humphreys, 1939). Finally, as expected, the random and acoustic models obtain the worst ABX accuracy. Overall, These results confirm that 600 h of self-supervised learning on effective speech suffices for wav2vec 2.0 to learn language-specific representations (Figure 3.13-B).

Wav2vec 2.0 and the brain learn language specific representations. Next, we compare the brain scores of random, non-speech, non-native and native models (Figure 3.13-C D). First, our results show that the non-speech model attains higher R scores than the random model (on average across voxels, $\Delta R = 0.006$, $p = 10^{-31}$) confirming the importance of learning to generate brain-like representations. Second, non-native models attain higher R scores than the non-speech model ($\Delta R = 0.002$, $p = 10^{-9}$), confirming that wav2vec 2.0 learns speech-specific representations of sounds when trained on speech. Finally, the native model attains higher R scores than non-native models ($\Delta R = 0.002$, $p = 10^{-15}$).

3.3.5 Discussion

Human infants acquire language with little to no supervision: A few hundred hours of speech suffices for their young brain to learn to discretize phonemes, segment morphemes, and assemble words in the language(s) of their social group (Dupoux, 2018; Gilkerson et al., 2017). However, the learning principle that allows this unique feat remains, to date, unknown.

Here, we test whether self-supervised learning applied to a limited amount of speech effectively accounts for the organization of speech processing in the human brain as measured with

fMRI. For this, we train several variants of wav2vec 2.0 (Baevski et al., 2020) with three curated datasets of French, English, and Mandarin, and compare their activations to those of a large group of French, English, and Mandarin speakers recorded with fMRI while passively listening to audio stories. Our results show that this self-supervised model learns (i) representations that linearly map onto a remarkably distributed set of cortical regions (Figure 3.11), (ii) a computational hierarchy that aligns with the cortical hierarchy (Figure 3.12), and (iii) features specific to the language of the participants (Figure 3.13).

Towards a biologically-plausible learning principle. These results extend recent findings on the similarities between the brain and a variety of deep learning models trained with biologically-implausible objectives and data. First, fMRI (A. J. E. Kell et al., 2018; Millet & King, 2021; Thompson et al., 2021), electroencephalography (Huang et al., 2018), and multi- or single-unit responses to sounds (Koumura et al., 2019; Begus et al., 2022) have been shown to be linearly predicted by the activations of deep convolutional networks trained on *supervised* auditory tasks. For example, (Millet & King, 2021) showed that a supervised speech-to-text model better accounted for brain responses to speech in 102 individuals when it was trained on speech recognition rather than auditory scene classification. Similarly, (A. J. E. Kell et al., 2018) showed that eight participants listening to brief speech and non-speech sounds demonstrated fMRI responses in the temporal lobe that aligned with those of a deep convolutional neural network trained on a binary auditory classification task. Our results, based on up to 50 times more fMRI recordings of the entire cortex show that such representational similarities hold with a self-supervised objective (Lerner et al., 2011; Berezutskaya et al., 2017; Caucheteux et al., 2021b, 2023). Second, a growing series of MEG (Toneva & Wehbe, 2019; Caucheteux & King, 2022), fMRI (Mitchell et al., 2008; Qian et al., 2016; Pereira et al., 2018; Schwartz et al., 2019; Antonello et al., 2021; Jain & Huth, 2018) and electro-physiology studies (Schrimpf et al., 2021; Goldstein et al., 2022) showed that text-based language models trained on very large corpora generate brain-like representations too. While these results suggest elements of convergence between language models and the brain (Caucheteux & King, 2022), they also remain biologically implausible: not only are these algorithms pre-equipped with abstract linguistic units such as characters and words, but they are trained on corpora that no one would ever be able to read in their lifetime. In contrast, wav2vec 2.0 is here trained with a reasonable amount of raw speech waveforms (Hart & Risley, 1992; Gilkerson et al., 2017; Dupoux, 2018). The functional similarity between wav2vec 2.0 and the brain thus opens the way to clarify how humans learn to process speech.

The emergence of a brain-like hierarchy of speech processing. The present study reveals the hierarchical organization of speech processing with remarkable clarity. First, the functional hierarchy learnt by wav2vec 2.0 is aligned with the anatomy: *e.g.* the superior temporal sulcus and the temporal pole are known to project to the ventral and dorsal part of the inferofrontal gyrus, respectively (Petkov et al., 2015). Second, the identification of functional gradients within the prefrontal cortex, and down to the motor areas typically associated with larynx and mouth control (Dichter et al., 2018) reinforces the relevance of motor processes to speech perception (Kellis et al., 2010; Mugler et al., 2014; Shamma et al., 2021). Finally, the existence of multiple levels of representations around the inferofrontal cortex is consistent with the idea that Broca's area may be responsible for merging linguistic units (Chomsky, 2000; Friederici, 1999; Hagoort, 2005; Poeppel et al., 2012). It should be noted, however, that our results aggregate a large cohort of individuals which could mask a more modular organization at the individual level.

Interpreting the neural representations of speech. Interpreting neural representations is a notoriously difficult challenge to both AI and neuroscience. Here, we first investigate language specificity and show that the neural representations specific to the native models are primarily represented in the superior temporal sulcus and middle temporal gyrus (Figure 3.13D): areas known to represent phonetic features (Mesgarani et al., 2014). However, these effect are relatively modest (Figure 3.13): the random model and the non-speech model reach, in STS and STG, 67% and 87% of the brain scores obtained by the "native" model, respectively. While this high baseline initially surprised us, this phenomenon could be explained by the fact that the auditory cortex is continuously bombarded by – and should thus be tuned to – non-speech input. Second, our probing analyses show that the models trained with self-supervised learning learn relevant acoustic and linguistic representations (Supplementary Figure S17). This result, consistent with Vaidya et al. (2022) and Stephenson et al. (2019), suggests that the difference of brain scores observed between the random, non-native and native models (Figure 3.13) may be partly driven by the corresponding spectro-temporal, phonetic, word and sentence-level representations, respectively. These elements of interpretation remain, however, scarce, and a systematic interpretation of the representations shared between wav2vec 2.0 and the brain remains necessary.

Scope of the study. It is important to stress that the scope of the present study could be broadened in several ways. First, our study focuses on adult speakers, whose cultural and educational background is not representative of the population (Henrich et al., 2010). Second,

we focus on the passive listening of three languages. Third, we focus on one self-supervised learning architecture (Baevski et al., 2020), and its functional alignment with fMRI, whose temporal resolution is notoriously limited. Generalizing the present approach to more languages (Malik-Moraleda et al., 2022), a larger spectrum of children and adult participants recorded with a variety of electrophysiological and neuroimaging devices will thus be essential to confirm, precise, and/or mitigate the present findings.

The remaining gap between brain and speech models. Several major gaps can be evidenced between wav2vec 2.0 and the brain. First, the transformer layers are not temporally constrained: each layer can access all elements within the contextual window. This differs from the necessarily recurrent nature of processing in the brain. Second, wav2vec 2.0 behaves differently to humans in specific tasks. In particular, it is overly-sensitive to band-pass filtering, non-robustly exploit fine temporal structures (Weerts et al., 2021) and fails to display the expected categorical responses (Millet et al., 2021). Third, recent studies show that wav2vec 2.0 encodes significantly less semantic information than text-based models (Pasad et al., 2021; Vaidya et al., 2022). While our analyses suggest that learning allows wav2vec 2.0 to capture some lexical features in its deep layers (Supplementary Figure S17, Supplementary Table S6), it remains unclear whether these layers also capture complex syntactic structures, such as recursive syntactic trees (Lakretz et al., 2021; Caucheteux et al., 2021a). We speculate that these limitations may be due to the time scales of wav2vec 2.0 which, unlike humans, learns very short-term representations of speech. In any case, these differences likely explain why the brain scores of wav2vec 2.0 remain substantially lower than our noise-ceiling (19% on average, and up to 74% in Heschl's gyrus and sulcus, Supplementary Table S3, Supplementary Figure S18).

Overall, the complexity of the human brain is often thought to be incompatible with a simple theory: "Even if there were enough data available about the contents of each brain area, there probably would not be a ready set of equations to describe them, their relationships, and the ways they change over time" (Gallant, 2013). By showing how the equations of self-supervised learning give rise to brain-like processes, this work contributes to challenge this view.

Chapter 4

Improving the similarity through hierarchical predictions

4.1 Evidence of a predictive coding hierarchy in the human brain listening to speech

4.1.1 Abstract

Considerable progress has recently been made in natural language processing: modern deep learning algorithms are increasingly able to generate, summarize, translate and classify texts. Yet, these language models still fail to match humans' language abilities. Predictive coding theory offers a tentative explanation to this discrepancy: While language models are optimized to predict nearby words, the human brain would continuously predict a hierarchy of representations that spans multiple time scales. To test this hypothesis, we analyze the fMRI brain signals of 304 participants listening to short stories. First, we confirm that the activations of modern language models linearly map onto the brain responses to speech. Second, we show that enhancing these algorithms with predictions that span multiple time scales improves this brain-mapping. Finally, we show that these predictions are organized hierarchically: Fronto-parietal cortices predict higher-level, longer-range and more contextual representations than temporal cortices. Overall, these results strengthen the role of hierarchical predictive coding in language processing, and illustrate how the synergy between neuroscience and A.I. can unravel the computational bases of human cognition.

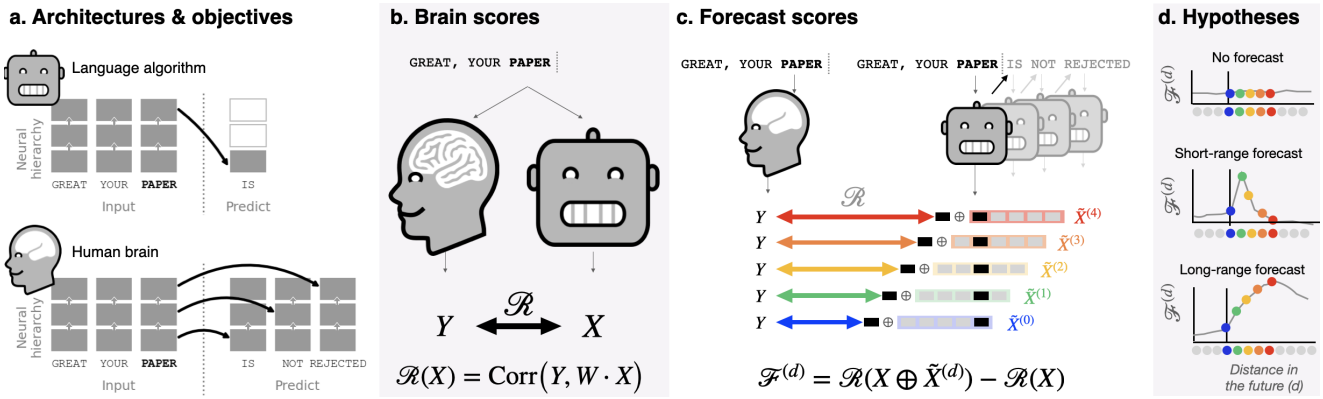


Figure 4.1: Approach. a. Deep language algorithms are typically trained to predict words from their close contexts. Unlike these algorithms, the brain makes, according to predictive coding theory, *(i)* long-range and *(ii)* hierarchical predictions. **b.** To test this hypothesis, we first extract the fMRI signals of 304 subjects each listening to ≈26 min of short stories (Y) as well as the activations of a deep language algorithm (X) input with the same stories. We then quantify the similarity between X and Y with a "brain score": a Pearson correlation $\mathcal{R}$ after an optimal linear projection W (Methods 4.1.5). **c.** To test whether adding representations of future (or predicted, see Supplementary Figure S24) words improves this correlation, we concatenate ($\oplus$) the network's activations (X, depicted here as a black rectangle) to the activations of a "forecast window" ($\tilde{X}$, depicted here as a colored rectangle). We use principal component analysis to reduce the dimensionality of the forecast window down to the dimensionality of X. Finally, $\mathcal{F}$ quantifies the gain of brain score obtained by enhancing the activations of the language algorithm to this forecast window. We repeat this analysis with variably distant windows (d, Methods 4.1.5). **d.** A flat forecast score across distances would indicate that forecast representations do not make the algorithm more similar to the brain (top). By contrast, a forecast score peaking at $d > 1$ (bottom) would indicate that the model lacks brain-like forecast. The peak of $\mathcal{F}^d$ indicates how far off in the future the algorithm would need to forecast representations to be most similar to the brain.

4.1.2 Introduction

In less than three years, deep learning has made considerable progress in text generation, translation and completion (Vaswani et al., 2017; Radford et al., 2019; Brown et al., 2020; Fan et al., 2018) thanks to algorithms trained with a simple objective: predicting words from their nearby context. Remarkably, the activations of these models have been shown to linearly map onto human brain responses to speech and text (Jain & Huth, 2018; Toneva & Wehbe, 2019; Caucheteux & King, 2022; Schrimpf et al., 2021; Toneva, Mitchell, & Wehbe, 2020a; Reddy & Wehbe, 2020; Goldstein et al., 2022; Millet et al., 2022). Besides, this mapping appears to primarily depend on the algorithms' ability to predict future words (Caucheteux & King, 2022; Schrimpf et al., 2021), hence suggesting that this objective suffices to make them converge to brain-like computations.

Yet, a gap persists between humans and these algorithms: in spite of considerable training

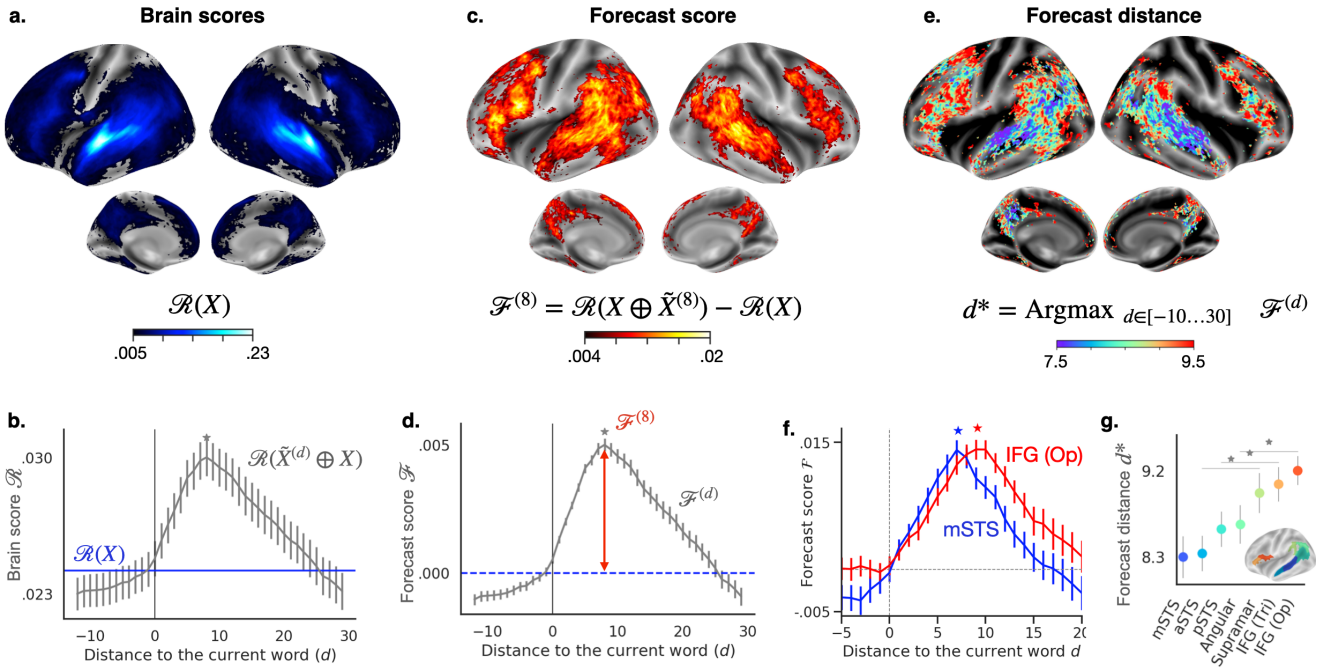


Figure 4.2: Isolating language predictions and their temporal scope in the human brain. a. The "brain score" ($\mathcal{R}$, Figure 4.1**b**, Methods 4.1.5), obtained with GPT-2, for each subject and each voxel, and here averaged across subjects (n=304). Only the voxels with significant brain scores are color-coded. **b.** Average (across voxels) brain scores obtained with GPT-2 with (grey) or without (blue) forecast representations. The average brain score peaks at $d^* = 8$ (grey star). **c.** For each voxel, the average (across subjects) "forecast score" $\mathcal{F}^d$, *i.e.* the gain in brain score when concatenating the activations of GPT-2 with a forecast window $\tilde{X}^{(8)}$. Only the voxels with significant forecast scores are color-coded. **d.** Average (across voxels) forecast scores for different distance d. **e.** Distance that maximizes $\mathcal{F}^d$, computed for each subject and each voxel, and denoted d^*. This "forecast distance" reveals the regions associated with short- and long-range forecasts. Regions in red and blue are associated with long-range and short-range forecasts, respectively. We only display the voxels with a significant average peak ($\mathcal{F}^{d^*} - \mathcal{F}^0, d^* = 8$, cf. Methods 4.1.5). **f.** Forecast score within two regions of interest. For each region, we report the average forecast scores of subjects with a representative peak (subjects whose peak belongs to the [45, 55] percentiles of all peaks, n=30 subjects). **g.** Forecast distance of seven regions of interest, as computed for each voxel of each subject and then averaged within the selected brain regions. For all panels, we report the average effect across subjects, and the error bars are SEM across subjects (n=304). All brain maps are thresholded at $p < .01$, as assessed with a FDR-corrected two-sided Wilcoxon test across subjects (n=304).

data, current language models remain challenged by long story generation, summarization as well as coherent dialogue and information retrieval (Holtzman et al., 2020; Wiseman et al., 2017; Thakur et al., 2021; Raffel et al., 2020; Krishna et al., 2021); they fail to capture several syntactic constructs and semantics properties (Lakretz et al., 2019; Arehalli & Linzen, 2020; Lakretz et al., 2021; Baroni, 2020; B. M. Lake & Murphy, 2021), and their linguistic understanding appears

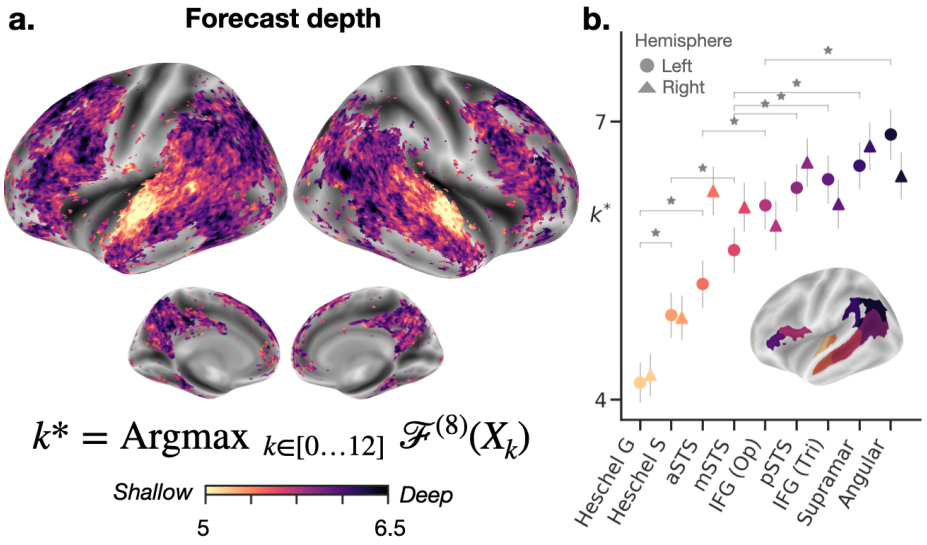


Figure 4.3: Organization of hierarchical predictions in the brain. **a.** Depth of the representation that maximizes the forecast score in the brain, denoted k^*. Forecast scores are computed for each depth, subject and voxel, at a fix distance $d^* = 8$ and averaged across subjects. We compute the optimal depth for each subject and voxel and plot the average forecast depth across subjects. Dark regions are best accounted for by deep forecasts, while light regions are best accounted for by shallow forecasts. Only significant voxels are color-coded, following Figure 4.2**c**). **b.** Same as a., with k^* averaged across the voxels of nine regions of interest, in the left (circle) and right (triangle) hemispheres. Scores are averaged across subjects and error bars are SEM across subjects (n=304). Pairwise significance between regions is assessed using a two-sided Wilcoxon test on the left hemisphere's scores (stars indicate that $p < .05$).

to be superficial (Marcus, 2020a; B. M. Lake & Murphy, 2021; Baroni, 2020; Arehalli & Linzen, 2020; Warstadt & Bowman, 2022). For instance, they tend to incorrectly assign the verb to the subject in nested phrases like 'The keys that the man holds ARE here' (Lakretz et al., 2021). Similarly, when text generation is optimized on next-word prediction only, deep language models generate bland, incoherent sequences or get stuck in repetitive loops (Holtzman et al., 2020).

Predictive coding theory (Rumelhart & McClelland, 1982; Rao & Ballard, 1999; K. Friston & Kiebel, 2009) offers a potential explanation to these shortcomings: while deep language models are mostly tuned to predict the very next word, this framework suggests that the human brain makes predictions over multiple time scales and levels of representations across the cortical hierarchy (Wacongne et al., 2011; Garrido et al., 2009) (Figure 4.1**a**).

Previous work has already evidenced speech predictions in the brain, by correlating word or phonetic surprisal – the extent to which a word or phone is expected – with functional Magnetic Resonance Imaging (fMRI) (Willems et al., 2016; Lopopolo et al., 2017; Okada et al.,

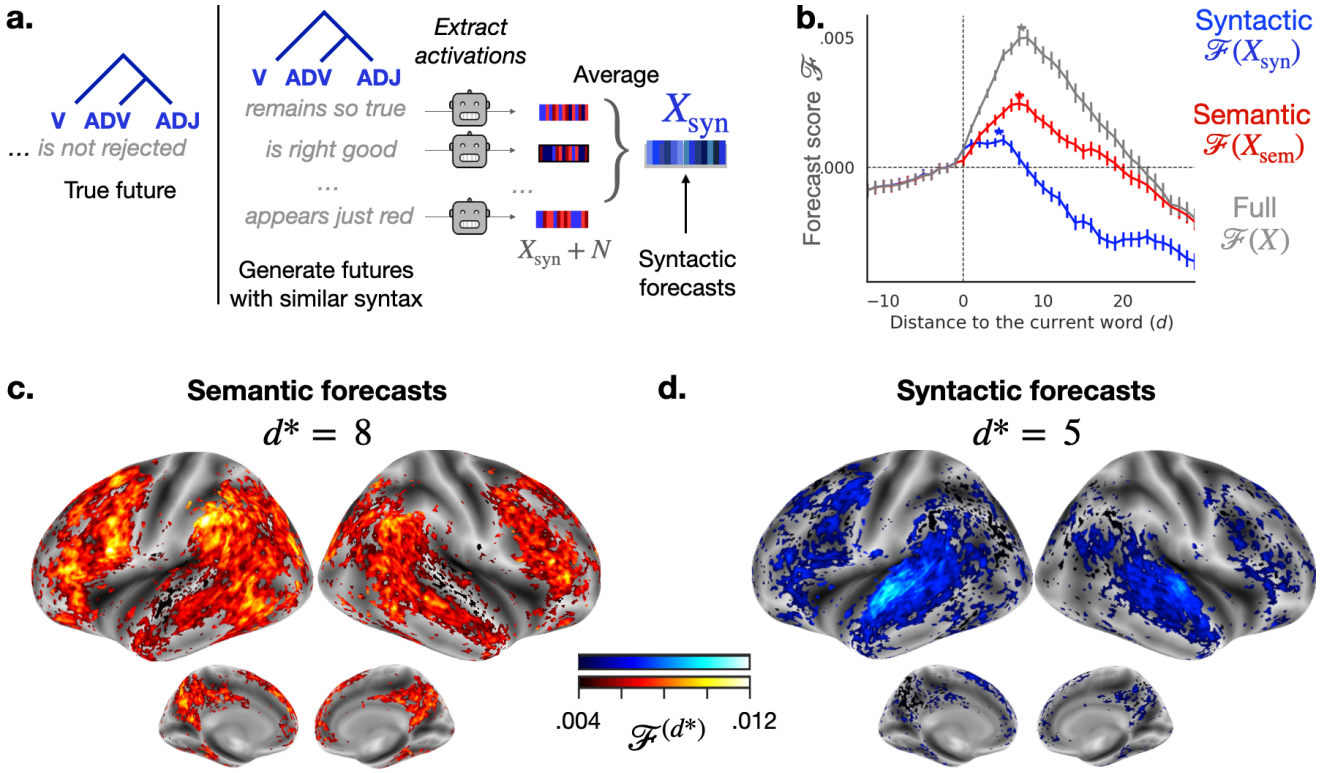


Figure 4.4: Factorizing syntactic and semantic predictions in the brain. a. Method to extract syntactic and semantic forecast representations, adapted from Caucheteux et al. (2021a). For each word and its context (e.g. 'Great, your *paper* ... ', we generate ten possible futures with the same syntax as the original sentence (part-of-speech and dependency tree) but randomly sampled semantics (e.g. '... remains so true', '... appears so small'). Then, we extract the corresponding GPT-2 activations (layer eight). Finally, we average the activations across the ten futures. This method allows to extract the syntactic component common to the ten futures, denoted X_{syn}. The semantic component is defined as the residuals of syntax in the full activations; $X_{\text{sem}} = X - X_{\text{syn}}$. We build the syntactic and semantic forecast windows by concatenating the syntactic and semantic components of seven consecutive future words, respectively (Methods 4.1.5). **b.** Syntactic (blue) and semantic (red) forecast scores, on average across all voxels, following Figure 4.1**c**. Scores are averaged across subjects and error bars are SEM across subjects (n=304). The average peaks across subjects is indicated with a star. **c.** Semantic forecast scores for each voxel, averaged across subjects and at $d^* = 8$, the distance that maximizes the semantic forecast scores in B. Only significant voxels are displayed similarly to Figure 4.2**c**. **d.** Same as c. for syntactic forecast scores and $d^* = 5$.

2018; Shain et al., 2020), electroencephalography (Heilbron et al., 2022; Donhauser & Baillet, 2020), magnetoencephalography (Mousavi et al., 2020) and electrocorticography (Forseth et al., 2020; Goldstein et al., 2022). However, such surprisal estimates derive from models trained to predict the very next word *or* phoneme, and reduce down their output to a single number: the probability of the next token. Consequently, the nature of the predicted representations as well

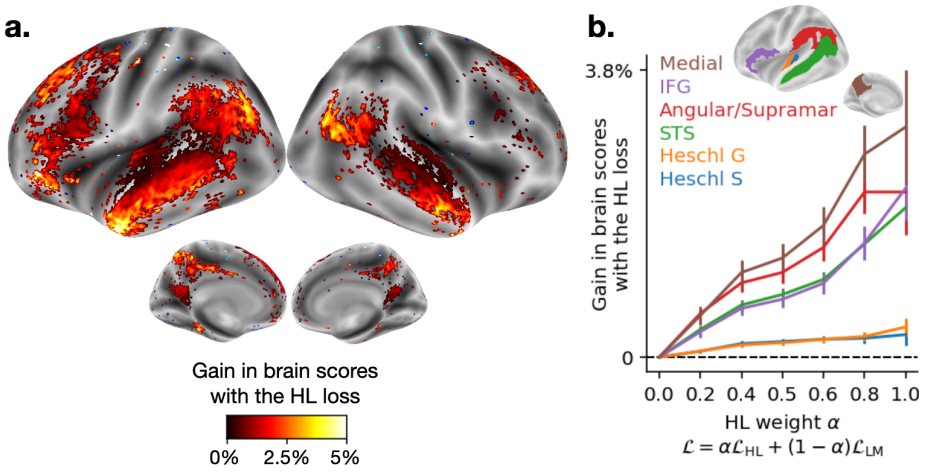


Figure 4.5: Gain in brain score when fine-tuning GPT-2 with a mixture of Language Modeling (LM) and High-Level prediction (HL). A) Gain in brain scores between GPT-2 fine-tuned with LM+HL and LM alone (for $\alpha_{HL} = 0.5$). Only the voxels with a significant gain are displayed ($p < 0.05$ with a two-sided Wilcoxon test after FDR correction for multiple comparison). **B)** Brain scores gain as a function of the HL weight α in the loss ((4.8)), from full LM (left, $\alpha = 0$) to full HL (right, $\alpha = 1$). Gains are averaged across voxels within six regions of interests (cf. Methods 4.1.5 for the parcellation and Supplementary Figure S27 for the other regions in the brain). Scores are averaged across subjects and error bars are SEM across subjects (n=304).

as their temporal scope remain largely unknown.

Here, we address these issues by analyzing the brain signals of 304 subjects listening to short stories while their brain activity is recorded with fMRI (Nastase et al., 2020). After confirming that deep language algorithms linearly map onto brain activity (Schrimpf et al., 2021; Caucheteux et al., 2021a; Toneva & Wehbe, 2019), we show that enhancing these models with long-range and multi-level predictions improves such brain mapping. Critically, and in line with predictive coding theory, our results reveal a hierarchical organization of language predictions in the cortex, in which the highest areas predict the most distant and the highest-level representations.

4.1.3 Results

Deep language models map onto brain activity. First, we quantify the similarity between deep language models and the brain, when these two systems are input with the same stories. For this, we use the Narratives dataset (Nastase et al., 2020), and analyze the fMRI of 304 subjects listening to short stories (27 stories ranging from 7 min to 56 min; 4.6 h of unique stimulus in total, 26 min on average per participant, from 7 min to 99 min). We then fit, for each

voxel and each subject independently, a linear ridge regression to predict the fMRI signals from the activations of a variety of deep language models. Finally, we compute the corresponding "brain scores" using held-out data, *i.e.* the voxel-wise correlation between the fMRI signals and the predictions of the ridge regression input with the activations of a given language model (Figure 4.1**b**, Methods 4.1.5). For clarity, we first focus on the activations of eighth layer of GPT-2, a twelve-layer causal deep neural network, provided by HuggingFace (Radford et al., 2019), as it has been shown to best predict brain activity (Schrimpf et al., 2021; Caucheteux & King, 2022).

In line with previous studies (Caucheteux & King, 2022; Caucheteux et al., 2021a; Wehbe et al., 2014; Jain & Huth, 2018), the activations of GPT-2 accurately map onto a distributed and bilateral set of brain areas. Brain scores peak in the auditory cortex, as well as in the anterior temporal and superior temporal areas (Figure 4.2**a**, Supplementary Note 6.6.1, Supplementary Figure S21, Supplementary Tables S7, S8, S9). The effect sizes of these brain scores are in line with previous work (Huth, de Heer, et al., 2016; Caucheteux & King, 2022; Toneva, Mitchell, & Wehbe, 2020b): for instance, the highest brain scores ($R = 0.23$, in the superior temporal sulcus (Figure 4.2**a**) represent 60 % of the maximum explainable signal, as assessed with a noise ceiling analysis (Methods 4.1.5). Supplementary Figure S22 show that, on average, similar brain scores are achieved with other state-of-the-art language models and Supplementary Figure S23 shows that auditory regions can be further improved with lower-level speech representations. As expected, the brain score of word rate (Supplementary Note 6.6.3), noise ceiling (Methods 4.1.5) and GPT-2 (Figure 4.2**a**) all peak in the language network (Fedorenko et al., 2016). Overall, these results confirm that deep language models linearly map onto brain responses to spoken stories.

Isolating long-range predictions in the brain. Next, we test whether enhancing the activations of language models with long-range predictions leads to higher brain scores (Figure 4.1**d**). Specifically, for each word, we concatenate *(i)* the model activations of the present word (denoted X) and *(ii)* a "forecast window" (denoted $\tilde{X}^{(d)}$), consisting of the embeddings of future words and parameterized by a temporal distance d and width of $w = 7$ words (see Supplementary Note 6.6.4 and Supplementary Figure S24 for the growing window analysis). While the width is the number of concatenated words, d corresponds to the distance between the current word and the last word of the window. For instance, $\tilde{X}^{(10)}$ is the concatenation of words at distance $4, 5$, up to 10 from the current word, and $\tilde{X}^{(8)}$ is the concatenation of words at distance 2, 3, up to 8 from the current word. For each distance d, we compute the "forecast

score" (denoted $\mathcal{F}^d$) by comparing the brain scores obtained with and without the forecast representations (Figure 4.2**b**).

Our results show that $\mathcal{F}$ is maximal for a distance of $d = 8$ words, and peaks in the areas typically associated with language processing (Figure 4.2**b-d**). For comparison, there are 2.54 words per second on average in the stimuli. Thus, 8 words corresponds to 3.15 seconds of audio (the time of two successive fMRI scans). These forecast scores are bilaterally distributed in the brain, at the exception of the infero-frontal and supramarginal gyri ($p < 0.001$ in Pars Opercularis and supramarginal, using a two-sided pairwise Wilcoxon test between the left and right hemispheres, after correcting for multiple comparisons, see Methods 4.1.5).

Supplementary analyses confirm that *(i)* each future word from word zero to ten significantly contributes to the forecast effect, *(ii)* forecast representations are best captured with a window size of 8 words, *(iii)* random forecast representations do not improve the brain scores, and *(iv)* using the words generated by GPT-2 instead of the true future words achieve lower but similar results (Supplementary Note 6.6.4, Supplementary Figure S24, Supplementary Note 6.6.5 and Supplementary Figure S25 and Supplementary Note S26).

Together, these results reveal long-range forecast representations in the brain, which represents a 23% ($\pm$ 9% across subjects) improvement in brain scores (Figure 4.2**a**,**b**).

Predictions' time range varies along the brain hierarchy. Both anatomical and functional studies have shown that the cortex is organized as a hierarchy (Felleman & Van Essen, 1991; Wacongne et al., 2011): for example, low-level acoustics, phonemes, and semantics are known to be primarily encoded in Heschl' gryus, superior temporal gyrus and the associative cortices of the frontal, temporal and parietal lobes, respectively (Lerner et al., 2011; A. J. E. Kell et al., 2018; Mesgarani et al., 2014; Huth, de Heer, et al., 2016; Hickok & Poeppel, 2007).

Do the different levels of this cortical hierarchy predict the same time window? To address this issue, we estimate the peak of the forecast score of each voxel and denote d^* the corresponding distance. The results show that the prefontal areas forecast, on average, further off in the future than temporal areas (Figure 4.2**e**). For instance, d^* in the inferior temporal gyrus (IFG) is higher than in the anterior superior temporal sulcus (aSTS) ($\Delta d^* = 0.9 \pm 0.2$, $p < 0.001$, Figure 4.2**f** and **g**).

The variation of optimal forecast distance along the temporo-parieto-frontal axis is largely symmetric across the two hemispheres (Supplementary Figure S21).

Predictions are increasingly contextual along the hierarchy. What is the *nature* of these predictive representations? To address this issue, we assess whether the forecast score relates to *(i)* low or high as well as *(ii)* syntactic or semantic representations. To this aim, we compute the forecast scores similarly as in Figure 4.1**c**, but now vary the layer used from GPT-2. Then, we identify k^* for each voxel *i.e.* the depth that maximizes the forecast scores (Methods 4.1.5). Here, we consider that the deep layers of language algorithms encode higher-level and more contextualized representations than their first layers (Jawahar et al., 2019; Manning et al., 2020).

Our results show that the optimal forecast depth varies along the expected cortical hierarchy (Figure 4.3**a**). Specifically, associative cortices are best modeled with deeper forecasts ($k^* > 6$) than low-level language areas (*e.g.* $k^* < 6$ in Heschl's gyri/sulci, anterior STS, Figure 4.3**a**-**b**). The difference between regions, while small on average, is highly significant across subjects (e.g. between the angular and Heschl's giri: $\Delta k^* = 2.5 \pm 0.3$, $p < 0.001$) , and observed in both the left and right hemispheres (Figure 4.3**b**).

Together, these results suggest that the long-range predictions of fronto-parietal cortices are more contextualized and higher-level than the short-term predictions of low-level brain regions.

Syntactic and semantic predictions show different time ranges. To factorize forecast representations into syntactic and semantic components, we apply a method introduced in (Caucheteux et al., 2021a) and proceed as follows: for each word and its preceding context, we generate ten possible futures which matches the syntax of the true future words. We choose $k = 10$ possible futures following (Caucheteux et al., 2021a). For each of these possible futures, we extract the corresponding GPT-2 activations, and average them across the ten possible futures (Figure 4.4**a**, Methods 4.1.5). This method allows us to decompose the activations of a given language model X into syntactic (the average vector, denoted X_{syn}) and semantic components (the residuals, $X_{\text{sem}} = X - X_{\text{syn}}$) (Methods 4.1.5). Once the syntactic and semantic forecast windows are built, we compute the corresponding forecast scores (Methods 4.1.5).

The results show that semantic forecasts are long-range ($d^* = 8$) and involve a distributed network peaking in the frontal and parietal lobes. By contrast, syntactic forecasts (Figure 4.4**b**) are relatively short-range ($d^* = 5$) and localized in the superior temporal and left frontal areas (Figure 4.4**c** and **d**). Note that the syntactic model without a forecast window (which has a lower dimensionality) performs better than the syntactic model with a distant forecast window. Such diminished scores can occur when there is no added information in the extra dimension of

the regression, because of the infamous curse-of-dimensionality (Bellman, 1966). This suggests that long-range syntactic forecast is not detectable in the present dataset.

Overall, these results reveal multiple levels of predictions in the brain in which the superior temporal cortex predominantly predicts short-term, shallow and syntactic representations whereas the infero-frontal and parietal areas predominantly predicts long-term, contextual, high-level and semantic representations.

Adapting GPT-2 into a predictive coding architecture. The above results show that concatenating present and future word representations of GPT-2 leads to a better modeling of brain activity, especially in fronto-parietal areas (Figure 4.2). Does fine-tuning GPT-2 to predict longer-range, more contextual and higher-level representations improve the brain-mapping in such regions? To answer this question, we fine-tune GPT-2 on Wikipedia, not only using Language Modelling (*LM*, *i.e.* predicting the next word), but also a High-level and Long-range objective (*HL*, *i.e.* predicting high-level representations of far-off words). Specifically, the *HL* objective is to predict the layer 8 of the pre-trained GPT-2 model, of word t+8 (Methods 4.1.5). The results show that GPT-2 fine-tuned with High-level and Long-range modeling best accounts for fronto-parietal responses (Figure 4.5, above 2% gain in IFG and the Angular/Supramarginal gyri, all $p < 0.001$) . On the contrary, auditory areas and lower-level brain regions do not significantly benefit from such a high-level objective (Figure 4.5 and Supplementary Figure S27). These results further strengthen the role of fronto-parietal areas in predicting long-range, contextual and high-level representations of language.

4.1.4 Discussion

In the present study, we put specific hypotheses of predictive coding theory to the test (Rumelhart & McClelland, 1982; Rao & Ballard, 1999; K. Friston & Kiebel, 2009): while deep language algorithms are typically trained to make nearby and word-level predictions (Vaswani et al., 2017; Radford et al., 2019; Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020; Clark et al., 2020), we assess whether the cortical hierarchy predicts multiple levels of representations, spanning multiple time scales. To this aim, we compare the activations of the brain to those of state-of-the-art deep language models (Huth, de Heer, et al., 2016; Jain & Huth, 2018; Toneva & Wehbe, 2019; Caucheteux & King, 2022; Caucheteux et al., 2022). We successfully validate our hypothesis on a cohort of 304 participants listening to spoken narratives (Nastase et al., 2020). Brain activity is best explained by the activations of deep language algorithms enhanced with long-range and high-level predictions. Our study provides three additional contributions.

First, the lateral, dorso-lateral and infero-frontal cortices as well as the supra-marginal gyrus here exhibit the longest forecast distances. Interestingly, these cortical regions were repeatedly linked to high-level semantics, long-term planning, attentional control, abstract thinking and other high-level executive functions (Gilbert & Burgess, 2008; Shallice & Burgess, 1991). This result echoes with previous studies showing that the integration constant of the fronto-parietal cortices is larger than those of sensory and temporal areas (L. Wang, 2021; Lee et al., 2021; Lerner et al., 2011; Caucheteux et al., 2021b). Specifically, our findings suggest that these regions, located at the top of the language hierarchy, are not limited to passively integrating past stimuli, but actively anticipate future language representations.

Second, we show that the depth of predictive representations varies along a similar anatomical organization: low-level predictions best model the superior temporal sulcus and gyrus, high-level predictions best model the middle temporal, parietal and frontal areas. This finding extends previous studies investigating the multiplicity of predictions underlying complex sound or speech processing (Wacongne et al., 2011; Vidal et al., 2019; Heilbron et al., 2022; Donhauser & Baillet, 2020). While previous studies focused on correlating brain activity with a subset of hand-crafted and unidimensional prediction *errors* (e.g. word or phoneme surprisal), the present analyses explore and decompose high-dimensional predictions. More generally, our results support the idea that, unlike current language algorithms, the brain is not limited to predict word-level representations, but rather predicts multiple levels of representations.

Finally, we decompose these neural activations into syntactic and semantic representations and show that *semantic* features – as opposed to syntactic ones – drive *long-range* forecasts. This finding strengthens the idea that while syntax may be explicitly represented in neural activity (Nelson et al., 2017; Ding et al., 2016; Caucheteux et al., 2021a), predicting high-level semantics may be at the core of long-form language processing (Jackendoff, 2002; Shain et al., 2021).

Together, these results support predictive coding theories, whereby the brain continually predicts sensory inputs, compares these predictions to the truth, and updates its internal model accordingly (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982; Rao & Ballard, 1999). Our study further clarifies this general framework. Not only does the brain predict sensory inputs, but each region of the cortical hierarchy appears to be organized to predict different temporal scopes and different levels of representations (Figure 4.1**a**). However, the link between the hierarchical constructs in syntax and the functional hierarchy in the cortex and in the model remains a major question to explore (Hale et al., 2021; Manning et al., 2020; Caucheteux et al., 2021a).

This computational organization is at odds with current language algorithms which are mostly trained to make adjacent and word-level predictions (Figure 4.1**a**). Some works have investigated alternative learning rules (Jernite et al., 2017; Fan et al., 2018; Devlin et al., 2019; Lewis et al., 2019; Yang et al., 2020; Joshi et al., 2020; Clark et al., 2020), but they do not combine both long-range and high-level predictions. We speculate that the brain architecture evidenced in this study presents at least one major benefit over its current deep learning counter-parts. While future observations rapidly become indeterminate in their original format, their latent representations may remain predictable over long time periods. This issue is already pervasive in speech- and image-based algorithms and has been partially bypassed with losses based on pretrained embedding (Szegedy et al., 2015), contrastive learning and, more generally, joint embedding architectures (T. Chen et al., 2020; He et al., 2020; El-Nouby et al., 2021; Bardes et al., 2022). Here, we highlight that this issue also prevails in language models, where word sequences – but arguably not their meaning – rapidly become unpredictable. Our results suggests that predicting multiple levels of representations over multiple temporal scopes may be critical to address the indeterminate nature of such distant observations, and adjust their relative confidence accordingly (Kepecs et al., 2008).

Three main elements mitigate the above conclusions. First, unlike temporally-resolved techniques (Donhauser & Baillet, 2020; Caucheteux & King, 2022; Goldstein et al., 2022), the temporal resolution of fMRI is around 1.5 s and can thus hardly be used to investigate sublexical predictions. Second, the precise representations and predictions computed in each region of the cortical hierarchy remain to be characterized. This will likely require new probing techniques, as the interpretation of neural representations remains a major challenge to both AI and neuroscience. Finally, the predictive coding architecture presently tested remains rudimentary. A systematic generalization, scaling and evaluation of this approach on natural language processing benchmarks remains necessary to demonstrate the effective utility of making models more similar to the brain.

Beyond clarifying the brain and computational bases of language, our study thus calls for systematically training algorithms to predict multiple times scales and levels of representations.

4.1.5 Methods

Notations

We denote:

- w a sequence of M words (here, several short stories).

- X the activations of a deep language model input with w, of size $M \times U$, with U the dimensionality of the embeddings (for a layer of GPT-2, $U = 768$). Except if stated otherwise, we use the activations extracted from the eighth layer of a 12-layer GPT-2 model (Methods 4.1.5). We will explicitly denote X_k the activations extracted from layer k when using another layer.
- Y the fMRI recordings elicited by w, of size $T \times V$, with T the number of fMRI time samples, and V the number of voxels (Methods 4.1.5).
- $\mathcal{R}(X)$ the brain score of X (Methods 4.1.5).
- $\widetilde{X}^{(d)}$ the forecast window containing information up to d words in the future. In short, the forecast window is the concatenation of the deep net activations of seven successive words, the last word being at a distance d from the current word (Methods 4.1.5).
- $\mathcal{F}^{(d)}(X)$, the forecast score at distance d, i.e. the gain in brain score when concatenating the forecast window $\tilde{X}^{(d)}$ to the network's activations; $\mathcal{F}^{(d)}(X) = \mathcal{R}(X \oplus \tilde{X}^{(d)}) - \mathcal{R}(X)$ (Methods 4.1.5).
- d^*, the distance maximizing the forecast score; $d^* = \text{argmax}_{d \in [-10,\dots,30]}\ \mathcal{F}^{(d)}(X)$ (Methods 4.1.5).
- k^*, the network's depth maximizing the forecast score at a fixed distance $d = 8$; $k^* = \text{argmax}_{k \in [0,\dots,12]}\ \mathcal{F}^{(8)}(X_k)$, with X_k the activations extracted from the k^{th} layer of GPT-2. We use $d = 8$ because it is the distance with the best forecast score on average across subjects and voxels (Methods 4.1.5).

fMRI dataset

We use the brain recordings (denoted Y) of the "Narratives" dataset (Nastase et al., 2020), a publicly available dataset containing the fMRI recordings of 345 subjects listening to 27 spoken stories in English, from 7 min to 56 min (4.6 h of unique stimulus in total). We use the pre-processed fMRI signals from the original dataset, without spatial smoothing (referred to as "afni-nosmooth" in the repository) and sampled with TR=1.5 s: the preprocessing steps were performed using fMRIPrep (Esteban et al., 2019), no temporal filtering was applied. The resulting preprocessing leads to the analysis of cortical voxels projected onto the surface and morphed onto a "fsaverage" template brain, and hereafter referred to as voxels for simplicity.

As suggested in the original paper, some subject-story pairs were excluded because of noise, resulting in 622 subject-story pairs and 4 h of unique audio material in total.

Deep language models' activations

We compare the fMRI recordings with the activations of a variety of pretrained deep language models input with the same sentences presented to the subjects. For clarity, we primarily focus on GPT-2, a high-performing *causal* language model trained to predict words given their previous context. GPT-2 consists of twelve Transformer modules (Vaswani et al., 2017; Radford et al., 2019), each of them referred to as "layer", stacked onto one non-contextual word embedding layer. Here, we use the pre-trained models from Huggingface (Wolf et al., 2020) (1.5 billion parameters, trained on 8 million web pages) (*c.f.* Supplementary Note 6.6.4 for the other deep language models).

In practice, to extract the activations X elicited by a sequence of M words w, from the k^{th} layer of the network, we 1) format the textual transcript of the sequence w (replacing special punctuation marks such as "–" and duplicated marks "?." by dots) 2) tokenize the text using Huggingface tokenizer, 3) input the network with the tokens and 4) extract the corresponding activations from layer k. This results in a vector of size $M \times U$, with M the number of words and U the number of units per layer (here $U = 768$). Given the constrained context size of the network, each word is successively input to the network with at most 1024 previous tokens. For instance, while the third word's vector is computed by inputting the network with (w_1, w_2, w_3), the last word's vectors w_M is computed by inputting the network with $(w_{M-1024}, \ldots, w_M)$. The alignment between the stories' audio recordings and their textual transcripts was provided in the original Narratives database (Nastase et al., 2020).

Brain scores

Following previous works (Huth, de Heer, et al., 2016; Caucheteux & King, 2022; Caucheteux et al., 2022), we evaluate, for each subject s and voxel v, the mapping between 1) the fMRI activations $Y^{(s,v)}$ in response to the audio-stories and 2) the activations X of the deep network input with the textual transcripts of the same stories. To this end, we fit a linear ridge regression W on a train set to predict the fMRI scans given the network's activations. Then, we evaluate this mapping by computing the Pearson correlation between predicted and actual fMRI scans on a held out set:

$$\mathcal{R}^{(s,v)} : X \mapsto \mathrm{Corr}\big(W \cdot X, Y^{(s,v)}\big) \quad , \tag{4.1}$$

with W the fitted linear projection, Corr Pearson's correlation, X the activations of GPT-2 and $Y^{(s,v)}$ the fMRI scans of one subject s at one voxel v, both elicited by the same held out stories.

In practice and following (Huth, de Heer, et al., 2016), we model the slow bold response thanks to a finite impulse response (FIR) model with 6 delays (from 0 to 9 seconds, TR=1.5 seconds). Still following (Huth, de Heer, et al., 2016), we sum the model activations of the words presented within the same TR, in order to match the sampling frequency of the fMRI and the language models (see Supplementary Figure S28 and S29). Then, we estimate the linear mapping W with a ℓ_2-penalized linear regression after standardizing the data, and reducing their dimensionality (computational reasons). We follow scikit-learn implementation (Pedregosa et al., 2011) and use a pipeline with the following steps: standardization of the features (set to 0 mean with a standard deviation of 1 using a 'StandardScaler'), principal component analysis (PCA) with twenty components and ℓ_2-penalized linear regression ('RidgeCV' in scikit-learn). In Figure S23**c**, we replicate the main analyses without PCA (the brain scores and forecast effect are slightly underestimated with PCA). The regularization hyperparameter of the 'RidgeCV' is selected with a nested leave-one-out cross-validation among ten possible values log-spaced between 10^{-1} and 10^{8} for each voxel and each training fold.

The outer cross-validation scheme allowing for an independent performance evaluation, uses five folds obtained by splitting the fMRI time series into five contiguous chunks. The Pearson correlations averaged across the five test folds is called "brain score", denoted $\mathcal{R}^{(s,v)}(X)$. It measures the mapping between the activation space X and the brain of one subject s at one voxel v, in response to the same language stimulus.

In Figure 4.2**a** and B, brain scores are computed for each (subject, voxel) pair. We then average the brain scores across subjects (Figure 4.2**a**) and/or voxels (Figure 4.2**b**) depending on the analysis. For simplicity, we denote $\mathcal{R}(X)$ the brain scores averaged across subjects and/or voxels.

Forecast windows

We test whether adding forecast representations improves our ability to predict brain activity. To this aim, we do not modify the deep network itself, but add forecast representations to the encoding model's input: the forecast window. The forecast window at distance d, denoted $\widetilde{X}^{(d)}$, is the concatenation of the network's activations of seven successive words, the last one being at a distance d from the current word. Precisely, the forecast window of a word w_n, at a distance

d is the concatenation of the network's activations elicited by words $w_{n+d-6}, \ldots, w_{n+d}$. Thus,

$$\widetilde{X}^{(d)} = (X_{w_{n+d-7}} \oplus \cdots \oplus X_{w_{n+d}})_{n \in [1,\ldots,M]} \quad , \tag{4.2}$$

with $\oplus$ the concatenation operator, and M the number of words in the transcript w (Figure S29). Note that d can be negative: in that case, the forecast window only contains past information. Except if stated otherwise, the forecast window is built out of the activations X extracted from the eighth layer of GPT-2. In Figure 4.3, the forecast window is built out of the activations X_k extracted from different layers k of GPT-2. We denote $\widetilde{X}_k^{(d)}$ the corresponding forecast windows. In Figure 4.4, the forecast windows are built out of the syntactic (X_{syn}) and semantic (X_{sem}) activations of GPT-2 (cf. Methods 4.1.5 and 4.1.5).

Forecast scores

For each distance d, subject s and voxel v, we compute the "forecast score" $\mathcal{F}^{(d,s,v)}$, which is the gain in brain score when concatenating the forecast windows to the present GPT-2 activations. Thus,

$$\mathcal{F}^{(d,s,v)} : X \mapsto \mathcal{R}^{(s,v)}(X \oplus \widetilde{X}^{(d)}) - \mathcal{R}(X) \quad , \tag{4.3}$$

To match the dimensionality of X and $\tilde{X}$, the principal component analysis used to compute the mapping (Methods 4.1.5) was trained on X and $\tilde{X}$ separately, before concatenating the two features: i.e. $\mathcal{F}(X) = \mathcal{R}(\text{pca}(X) + \text{pca}(\tilde{X})) - \mathcal{R}(\text{pca}(X))$.

Forecast distance

To test whether the forecast scope varies along the cortical hierarchy, we estimate the distance that maximizes the forecast score. Precisely, the optimal "forecast distance" d^* for each subject s and voxel v is defined as:

$$d^*_{(s,v)} = \text{argmax}_{d \in [-10,\ldots,30]} \, \mathcal{F}^{(d,s,v)}(X) \ , \tag{4.4}$$

with X the activations of the language model, $\mathcal{F}^{(d,s,v)}$ the forecast score at distance d for subject s and voxel v (Equ. (4.3)). The forecast distances d^* are then averaged across subjects and/or voxels depending on the analyses.

The present analysis is only relevant for the brain regions for which forecast scores are not flat. Indeed, computing the distance maximizing a flat curve would be misleading. Thus, in Figure 4.2**e**, we compute the difference $\mathcal{F}^8 - \mathcal{F}^0$ for each subject and voxel, assess the significance with Wilcoxon test across subjects, and ignore the voxels with a non-significant difference ($p > .01$).

Forecast's depth

To test whether the forecast depth varies along the cortical hierarchy, we compute the forecast score for different depth of representation. Precisely, we proceed similarly as in 4.1.5, but replacing X by the activations X_k extracted from layer k of GPT-2 ($k \in [0, \ldots, 12]$) in (4.3) and (4.2). Then, we compute the depth maximizing the forecast score, called "forecast depth", and given by:

$$k^*_{(d,s,v)} = \text{argmax}_{k \in [0,\ldots,12]} \mathcal{F}^{(d,s,v)}(X_k) \quad , \tag{4.5}$$

with $\mathcal{F}^{(d,s,v)}(X_k) = \mathcal{R}^{(s,v)}(X_k \oplus \widetilde{X_k}^{(d)}) - \mathcal{R}(X_k)$ ((4.3)). For simplicity, we focus on the fixed distance $d = 8$ (Figure 4.3**c** and D), which maximizes the forecast score in Figure 4.2.

Decomposing model activationsinto syntactic and semantic components

To extract the syntactic and semantic components of X, a vector of activations in response to a story w, we apply a method introduced in (Caucheteux et al., 2021a) (Figure 4.4**a**). For each word, 1) we generate $k = 10$ futures of the same syntax as the true future (*i.e.* same part-of-speech and dependency tags as the true future), but randomly sampled semantics, 2) we compute the activations for each of the ten possible futures,and 3) we average the activations across the ten futures. We use the same hyper-parameter $k = 10$ as in the original paper. The method actually converges from K 7 (Figure S8 in the paper). This method allows to extract the average vector X_{syn}, that contains syntactic information but is deprived from semantic information. The semantic activations $X_{\text{sem}} = X - X_{\text{syn}}$ are the residuals of syntax in the full activations X. In the original paper (Figure 3), the authors checked with probing analyses that the syntactic embeddings encoded relevant syntactic information (part-of-speech and depth of the syntactic tree), and no longer encoded semantic information (word frequency, word embedding, semantic category).

Syntactic and semantic forecast windows

To investigate syntactic and semantic forecasts in the brain, we build forecast windows out of the syntactic and semantic activations of GPT-2, respectively. To this aim, we first build the forecast windows out of GPT-2 activations $\widetilde{X}^{(d)}$, similarly as 4.1.5. Then, we extract the syntactic $\widetilde{X}^{(d)}_{\text{syn}}$ and semantic $\widetilde{X}^{(d)}_{\text{sem}}$ components of the concatenated activations, as introduced in (Caucheteux et al., 2021a) and described in 4.1.5. Finally, the syntactic forecast score is the increase in brain score when concatenating the syntactic window:

$$\mathcal{F}^{(d)}_{\text{syn}} = \mathcal{R}(X \oplus \widetilde{X}^{(d)}_{\text{syn}}) - \mathcal{R}(X) \tag{4.6}$$

Similarly, the semantic forecast score is given by:

$$\mathcal{F}^{(d)}_{\text{sem}} = \mathcal{R}(X \oplus \widetilde{X}^{(d)}_{\text{sem}}) - \mathcal{R}(X) \tag{4.7}$$

Brain parcellation

We systematically implement whole brain analyses and compute scores for each voxel in the brain. Yet, for simplicity, we report the scores averaged across selected regions of interest in Figure 4.2**f,g** and 4.3**c**. To this aim, we use a subdivision of the Destrieux Atlas (Destrieux et al., 2010). Regions with more than 500 vertices are split into smaller parts. This results in 142 regions per hemisphere, each containing less than 500 vertices. In Figure 4.2**g** and 4.3**c**, we use the following acronyms:

Acronym	Definition
STG / STS	Superior temporal gyrus / sulcus
aSTS	Anterior STS
mSTS	Mid STS
pSTS	Posterior STS
Angular / Supramar	Angular / Supramarginal inferior parietal gyrus
IFG / IFS	Inferior frontal gyrus / sulcus
Tri / Op	Pars triangularis / opercularis (IFG)
Heschl G / Heschl S	Heschl gyrus / sulcus

Statistical significance

We systematically implement single-subject and whole brain analyses: all metrics (brain score, forecast score, forecast distance and depth) are computed for each subject, voxel pair. We report the metrics averaged across subjects and/or voxels depending on the analysis. Statistics are computed across subjects, using the two-sided Wilcoxon test from Scipy (Virtanen et al., 2020) assessing whether the metric (or the difference between two metrics) is significantly different from zero, and then corrected for multiple comparisons using False Discovery Rate. We report an effect as significant if its p-value is lower than 0.01. Error bars systematically refer to the Standard Errors of the Means (SEM) across subjects, following Scipy implementation.

Noise ceiling

FMRI recordings are inherently noisy. To assess the amount of explainable signal, we use a "noise ceiling" analysis, *i.e.* we predict the brain responses $Y^{(s)}$ of each subject s given the other subjects' responses to the same story $\overline{Y}$. We proceed similarly as the brain score computation and apply the same setting (4.1), but use the average brain signals of other subjects' brain $\overline{Y}^{(s)} = \frac{1}{|\mathcal{S}|}\sum_{s' \neq s} Y^{(s')}$ (of size $T \times V$) instead of the network's activations X. Precisely:

- For the brain score computation, $Y^{(s)}$ is the fMRI recordings of subject s, corresponding to all the stories subject s listened to while being scanned. X consists of the contextual embeddings of the corresponding words, summed within each TRs and transformed with FIR. Thus,
$$R_{\text{brainscore}}(s) = \text{Corr}[W^{(s)} \cdot X, Y^{(s)},]$$
with X the GPT-2 embeddings, temporally aligned with Y using FIR.

- For the noise ceiling computation, $Y^{(s)}$ is the same as for the brain score computation. X consists of the average fMRI recordings of the other subjects that listened to the same stories as subject s. X and Y have the same dimensionality here, and the bold delay is assumed to be comparable across subjects, so we do not apply a FIR to X. Thus,
$$R_{\text{noiseceil}}(s) = \text{Corr}[W^{(s)} \cdot \overline{Y}^{(s)}, Y^{(s)}],$$
with $Y^{(s)}$ the average fMRI of the other subjects having listened to the same story as subject s.

For both the brain score and noise ceiling computation, we fit a ridge regression $W^{(s)}$ for each subject s, predicting $Y^{(s)}$ given X, using the same five-folds cross-validation setting. We evaluate the prediction successively on the 5 test folds using Pearson correlation and average the correlation scores across folds. This results in one brain score and one noise ceiling estimate per subject (and voxel). Results averaged across subjects are displayed in Supplementary Figure S30. This score is one possible upper bound for the best brain score that can be obtained given the level of noise in the dataset.

Fine-tuning GPT-2 with a Long-Range and High-level objective

Does fine-tuning GPT-2 to predict long-term, high-level and more contextualized representations increase its similarity with the brain?

To test this question, we fine-tune GPT-2 using a mixture of language modeling loss (*LM*) and a high-level and long-term loss (*HL*). We then evaluate brain scores and test whether the *HL* objective leads to significantly higher brain scores than the *LM* objective.

Architecture and Losses We fine-tune the pre-trained GPT-2 model provided by Huggingface with a mixture of Language Modeling (*LM*) and High-level forecast (*HL*). The mixture loss is parametrized by a hyper-parameter $\alpha \in [0, 1]$. The total loss minimized is given by:

$$\mathcal{L} = \alpha' \mathcal{L}_{HL} + (1 - \alpha') \mathcal{L}_{LM} \, . \tag{4.8}$$

with the constraint that $\alpha' \mathcal{L}_{HL} = \alpha(1 - \alpha') \mathcal{L}_{LM}$. Doing so, setting α to 0.5 means that each term of the loss contributes to 50% of the total loss. The *LM* objective is to predict the next word and it is given by:

$$\mathcal{L}_{LM} = \mathrm{CE}\big[h_{LM} \circ f(x_t), x_{t+1}\big] \, ,$$

with:

- CE the cross entropy loss.
- f is the learned fine-tuned model. f is initialized with the weights of pretrained GPT-2. Thus, f is a twelve-layers transformer network stacked onto a word-embedding, each layer having a dimensionality of 768.
- h_{LM} is the language modeling linear head on top of the last layer of f, from 768 to n_{vocab}, that predicts the next word.
- x_t the input tokens.
- x_{t+1} the input tokens shifted from one time step (the succeeding words).

The *HL* objective is to predict layer k of word at distance d from the current word and it is given by:

$$\mathcal{L}_{HL}^{k,d} = \mathrm{CPC}[h_{HL} \circ f(x_t), N^k(x_{t+d})] \, ,$$

where:

- N^k is a separate and fixed network, here the pretrained version of GPT-2 provided by Huggingface, taken at layer k. Its weights do not vary with training.

- h_{HL} is a linear head on top of the last layer of f, from 768 to 768, that predicts the activations of the k^{th} layer of the fixed network N^k, corresponding to the word at distance d from the current word.
- x the inputs, x_t marks the current words, x_{t+d} marks the words at distance d from the current word.
- CPC is the contrastive predicting coding loss (Hénaff et al., 2019).

$$\text{CPC} = -\text{Log}\frac{\text{Exp}\left[S\left(y_{\text{pred}}, y_{\text{true,pos}}\right)/\tau\right]}{\sum_{neg}\text{Exp}\left[S\left(y_{\text{pred}}, y_{\text{true,neg}}\right)/\tau\right]},$$

  with S a similarity metric, $y_{\text{true,neg}}$ a set of negative samples, and $y_{\text{true,pos}}$ a set of positive samples.

In practice, we choose to predict the hidden states at layer $k = 8$, of the future word at distance $d = 8$. We choose layer $k = 8$ and $d = 8$ because it leads to the best results (Figure 4.2**d**). To compute the CPC loss, we take $\tau = 0.1$ and use the cosine similarity as similarity metric S. We use 2,000 negatives randomly sampled from a negative queue (of size 2,500). The negative queue is updated at each batch by adding the hidden states to the non-target words from the current batch. Such hidden states are extracted from the pretrained network at layer k (N^k). In order for the HL and LM losses to have a fixed contribution α and $1 - \alpha$ over training, we update the parameter α' in (4.8) every 100 gradient steps.

Dataset and training We fine tune GPT-2 on the already pre-processed English Wikipedia dataset (`https://huggingface.co/datasets/wikipedia`) comprised of 6M documents (30 GB), for three days on 2 GPUs. We use the 'Trainer' implementation from Huggingface with the default training arguments (Adam optimizer, learning rate = 0.00005, see `https://huggingface.co/docs/transformers/main_classes/trainer` for the other default parameters). Because of memory constraints, we restrict the context size of GPT-2 to 256 tokens, and use a batch size of 4 per device (thus, $2 \times 4 \times 256 = 1024$ tokens per batch and gradient updates). For stability, we fine-tune the top-tier layers of the network (from layer 8 to layer 12), while the bottom layers are kept frozen. Fine-tuning the whole network with language modelling led to a significant drop in brain scores (with fixed training parameters). Losses were monitored on a separate evaluation set of 1,000 Wikipedia documents.

Evaluation We fine-tune seven GPT-2 models with different HL weight α, from a loss being full LM ($\alpha = 0$), half LM and HL ($\alpha = 0.5$) to full HL ($\alpha = 1$). During the training, we save ≈ 15 model checkpoints (regularly log-spaced between 0 and 10^6 gradient updates). For each model and step, we compute the brain scores of its concatenated layers [0,4,8,12] on the same Narratives dataset (Nastase et al., 2020), as explained in Methods, Section 4.1.5. We here choose to span all layers from 0 to 12 because representations could "move" across layers during the fine-tuning which could bias the results. We then average the brain scores across steps and assess the gain of one network over another. In Figure 4.5, we report the gain averaged across subjects when adding increasingly more HL in the loss.

Chapter 5

Discussion

5.1 Main findings

Do artificial neural networks and the human brain build similar intermediate representations to process language?

In this thesis, we employ a correlational metric, the "brain score", to identify high-level similarities and remaining differences between the language representations of the brain and those of artificial neural networks.

In Chapter 2, we demonstrated that (i) deep networks' activations significantly predict brain activity in response to isolated words, sentences, and narratives, recorded with MEG/fMRI, across large cohorts of more than 500 participants, (iii) in language-related brain areas, (iv) for the most accurate algorithms, i.e., those that best predict a word based on its context, and (v) for participants with higher story comprehension levels (measured by a post-story questionnaire).

In Chapter 3, we decomposed the activations of deep neural networks to better interpret the *nature* of the shared representations and neural responses to natural language stimuli. By combining encoding models with artificial neural networks, we demonstrated a finer-grained decomposition of the spatial and temporal hierarchy of natural language, language specificity, as well as syntactic and semantic processes in the brain.

In Chapter 4, we investigated how to build algorithms more similar to the brain. We showed that artificial neural networks predict short-term, word-level representations, while the brain

predicts long-term, hierarchical representations. We found that enhancing the Generative Pre-trained Transformer 2 (GPT-2) with the ability to predict longer-term and more abstract representations increased its similarity with the brain.

Overall, our findings highlight the potential of artificial neural networks to elucidate the human brain's language processing mechanisms, while also emphasizing the need for further improvements to bridge the gap between the two.

5.2 A thriving field of study

During my three years of doctoral studies, several other teams conducted similar research comparing deep language models and brain recordings. Overall, these complementary results converge toward similar conclusions to the one outline above. Below, we will briefly review the major similarities and differences of these parallel works.

Linear mapping between transformers and the brain. Multiple studies have demonstrated a linear mapping between brain responses and artificial neural network activations, as seen in significant predictions in fMRI, MEG, and EEG scans (Jat et al., 2019; Hollenstein et al., 2019; Schrimpf et al., 2021; Toneva, Stretcu, et al., 2020; Toneva, Mitchell, & Wehbe, 2020a,b; Toneva & Wehbe, 2019; Reddy & Wehbe, 2020; Caucheteux & King, 2022; Sun et al., 2021; Anderson et al., 2021; S. Wang, Zhang, Wang, et al., 2020; Vaidya et al., 2022; Jain et al., 2023). Multiple transformer models have been tested, mostly derived from causal, masked, and permutation language modeling (Devlin et al., 2019; Radford et al., 2019; Yang et al., 2020), or a combination of language modeling and classification tasks (Raffel et al., 2020), with causal language models exhibiting the best brain scores (Schrimpf et al., 2021). Consistently with our results, most studies found that deep neural networks correlate with a distributed and bilateral network in the brain, peaking in regions historically associated with language (Hickok & Poeppel, 2007; Fedorenko et al., 2016), and in the middle layers of deep neural networks (Jain & Huth, 2018; Toneva, Mitchell, & Wehbe, 2020a; Manning et al., 2020). Interestingly, multiple studies reported that randomly initialized models modestly yet significantly predicted brain responses, and showed an increase in brain score with training (Schrimpf et al., 2021; Caucheteux & King, 2022; Pasquiou et al., 2022).

Factors modulating the brain score. Together with our work (Caucheteux & King, 2022), a few studies have investigated the factors that modulate this linear mapping, particularly the network's training task. This approach may be important for gaining insights into *why* the brain work the way it does (Kanwisher et al., 2023). In line with our results, Schrimpf et al. (2021) showed that the ability to predict the next word based on past context is a stronger predictor of brain mapping than other standard natural language processing tasks (GLUE tasks). Compared to (Caucheteux & King, 2022), the authors studied multiple architectures (both LSTMs and Transformers), tasks (including multi-lingual modeling (Lample & Conneau, 2019) and T-5 (Raffel et al., 2020)) yet studied only pre-trained or randomly initialized models. They tested deep nets' ability to predict both fMRI and ECoG recordings (as opposed to fMRI and MEG in (Caucheteux & King, 2022)). They studied a smaller number of participants yet also analysed the participants' reading time. Overall, both the brain and behavioural score strongly correlate with the models' ability to predict the next word (0.44 and 0.67 on average), which strengthens the role of next-word prediction in the brain. Goldstein et al. (2022) reinforced the importance of the next-word prediction task by using artificial neural networks to directly track next-word representations in ECoG recordings. On the other hand, Pasquiou et al. (2022) challenged these conclusions and argued that perplexity is not a reliable predictor of the brain score. Addressing a slightly different question, Gauthier & Levy (2019) explored transfer learning tasks and found that fine-tuning networks on semantic tasks such as natural language inference (MNLI in GLUE) and sentiment analysis (SST-2 in GLUE) increased fMRI predictability. Michelmann et al. (2023) recently showed that training models on complex narrative understanding also improves the brain score, suggesting that the brain may perform higher-level objectives. Antonello & Huth (2022) reached mixed conclusions, by showing that brain scores do correlate with next-word prediction, but also with other transfer learning tasks and translation abilities. Precising our results (Caucheteux & King, 2022), Antonello & Huth (2022) also found that the best performing networks in terms of next-word prediction are not necessarily the most brain-like: by the end of training, the brain score decreases while perplexity continues to improve.

Overall, our results, along with the aforementioned studies, reinforce the importance of next-word prediction in the emergence of similarities between deep language models and the brain. However, they also suggest that the brain may be guided by a higher-level objective, as proposed in Section 4.1.

Decomposing the shared representations. Recent studies have decomposed language representations in artificial neural networks to study the content of language representations in

the brain, disentangling the effects of context (Jain & Huth, 2018; Jain et al., 2023), layer depth (Toneva & Wehbe, 2019; Jain & Huth, 2018; Vaidya et al., 2022), syntax and semantics (Reddy & Wehbe, 2020; S. Wang, Zhang, Lin, & Zong, 2020). Consistently with our results (Caucheteux et al., 2021a,b; Millet et al., 2022), they demonstrate a parallel organization of language representations in both the cortex and the model, where the layer and contextual hierarchy of the model correspond with the anatomical structure (Jain & Huth, 2018; Vaidya et al., 2022), as well as a distributed network of both syntactic and semantic processes in the brain (Reddy & Wehbe, 2020; S. Wang, Zhang, Lin, & Zong, 2020; Pasquiou et al., 2023). The authors employed varying approaches to decomposing syntactic and semantic representations. For instance, Pasquiou et al. (2023) trained a neural network on corpora that were deprived of either syntactic or semantic information, and then measured the mapping between these information-restricted algorithms and brain activity. Consistent with our own findings, Pasquiou et al. (2023) observed that most brain regions were sensitive to both syntactic and semantic processes, but with differing magnitudes of responses. Specifically, frontal, parietal and medial regions exhibited a greater sensitivity to semantics, as opposed to temporal regions. However, contrary to (Caucheteux et al., 2021a), they also found that gains in brain score for *compositional* semantic representations were relatively modest compared to *lexical* semantic representations, and mostly localized in medial regions.

Overall, the simultaneity of these works is not a coincidence, but results from collective efforts, including the development of open-source deep learning models (Wolf et al., 2020) and the creation of publicly available brain recording datasets like Nastase et al. (2020); Schoffelen et al. (2019); Allen et al. (2022). Our study builds on this foundation by conducting a large-scale investigation into the similarities and differences between AI systems and the human brain. By retraining deep learning models from scratch, focusing on the differences between the two systems, and studying multiple datasets of large cohorts of participants, mostly in fMRI, the present manuscript complements the aforementioned studies, and contributes to a rapidly growing field at interface between AI and neuroscience.

5.3 Limitations and future work

5.3.1 Building more accurate encoding models

Despite advances in encoding models based on deep networks, the accuracy of predicting brain activity remains imperfect, with brain scores falling short of 1. Although this can be partly

attributed to the inherent noise in fMRI and MEG data, as well as the analysis of single-trial responses from single sensors, a significant portion of the variability in brain activity remains unexplained. In this section, we present various approaches to address this gap in brain predictability.

Learning better encoding representations. One possible explanation for this gap is the use of suboptimal encoding representations.

- Leveraging pre-trained models from the NLP community
  One approach is to leverage pre-trained models from the natural language processing community. Several research teams have investigated the brain score of recent artificial models (Toneva & Wehbe, 2019; Schrimpf et al., 2021; Michelmann et al., 2023), and launched a dedicated benchmark to quantify language models' brain similarity[1]. Going forward, it will be exciting to explore the brain similarity of less standard models trained to incorporate long-term dependencies (Beltagy et al., 2020; Raikote, 2021), multi-step planning (Hu et al., 2022), information retrieval (Borgeaud et al., 2022; Izacard et al., 2022), models trained on code (Nijkamp et al., 2022), mathematics (Jiang et al., 2022), other languages (Lample & Conneau, 2019), and music (Dhariwal et al., 2020; Agostinelli et al., 2023), in order to elucidate the specific brain representations underlying such higher-level cognitive processes.

- Learning better encoding representations through neuro-scientific intuitions
  By building models that incorporate principles of neural processing, researchers may be able to develop more brain-like models. In Chapter 4, we leverage insights from predictive coding theory to enhance deep language models with long-range and hierarchical learning rules, leading to improved brain predictability. While difficult, this approach may provide a more direct way of addressing questions in neuroscience, rather than simply testing the latest NLP models. An exciting direction for future work would be to draw inspiration from cognitive processes widely studied in neuroscience but currently lacking from the best algorithms, like episodic and semantic memory, continual learning, one shot generalisation, inductive reasoning, imagination and multi-step planning (Hassabis et al., 2017; Mahowald et al., 2023; Bang et al., 2023).

[1] https://github.com/brain-score/language

- Learning better encoding representations through deep encoding models
  Another approach is to train or fine-tune deep models directly to predict brain responses. This approach has already been explored in vision research, which has led to the development of specific benchmarks to compare deep encoding models of brain responses to natural scenes images (Gifford et al., 2023). Such an approach could be an exciting avenue for future work to enable better encoding models of brain responses to language.

Leveraging data from multiple subjects and larger datasets. In the present manuscript, we build separate encoding models for each subject, which limits the amount of data available for training. This approach can be improved by leveraging data from other participants and multiple datasets, but combining heterogeneous recordings is challenging. Previous studies have addressed this challenge by using non-linear methods in fMRI, MEG and EEG (Mohsenvand et al., 2020; Chehab et al., 2022; Thomas et al., 2022; Défossez et al., 2022). Inspired by these approaches, future work may explore linear encoding models that incorporate participant-specific parameters to leverage data from more participants and multiple datasets.

Inter-subject variability. In our analyses, we typically report average scores and statistical significance across individuals, after projecting brain responses onto a common surface map (fsaverage). However, this approach only partially takes into account the inter-subject variability in functional responses. This limitation is particularly problematic in high-level brain regions where responses are known to vary significantly across individuals (Mahowald & Fedorenko, 2016). To address this issue, future studies should explore more powerful anatomical and functional alignment techniques that can better capture individuals' specificity and may improve predictive accuracy (P.-H. C. Chen et al., 2015; Richard et al., 2019; Haxby et al., 2020; Bazeille et al., 2021; Thual et al., 2022).

Improving the estimation of noise levels in neuro-imaging recordings. Some of the unexplained variability in neuro-imaging recordings may be attributed to inherent noise rather than limited encoding models. Defining noise and signal depends on the scientific question at hand. Here, our aim is to comprehend brain responses to *language*, independent of participants' movement, recording conditions, and stimulus modality. In this study, we utilized a shared-response model as an upper bound for the best possible achievable scores. Precisely, it assumes that the signal is brain activity shared across all participants in response to the same stimuli. As language representations vary across individuals, the shared response model is an imperfect

solution (Mahowald & Fedorenko, 2016; Seghier & Price, 2018). Exploring novel methods to quantify the signal in single-trial settings while accounting for inter-individual variability will be crucial to quantify the remaining gap between deep net algorithms and the human brain.

A debate on non-linearity. In our study, we choose to use *linear* encoding models to focus on neural *representations*, which refer to linearly readable information from brain activity (DiCarlo & Cox, 2007; Kriegeskorte et al., 2008). The use of linear models allows us to capture the intended focus of our investigation, without capturing additional information beyond our intended scope. To illustrate the importance of linearity, consider the classic example of information read from the retina. When a participant views an image of a dog, the pixel-level information of the image is represented in the retina. A non-linear model could reconstruct the concept of "dog" from the pixel-level representations of the retina, while the retina itself does not represent the concept of "dog". Thus, while non-linear encoding models may offer increased predictive power, they may also capture information beyond our intended focus.

5.3.2 Improving encoding models' evaluation

Generalizing to out-of-distribution conditions and super stimuli. In this study, we validated our encoding models on 20% of the stimuli presented to each subject. While this provides a strong initial assessment of the models' performance, further evaluation on new recording conditions (e.g. across different laboratories) and novel stimuli (e.g. across novel stories) is necessary to fully test their robustness. Furthermore, it would be worthwhile to investigate the models' ability to perform in out-of-distribution scenarios, including the presence of adversarial examples and "super stimuli" - stimuli that go beyond the normal distribution of natural stimuli, such as caricatures in visual processing (Leopold et al., 2006). In the field of vision research, for instance, Bashivan et al. (2018) used artificial neural networks to generate images that maximally activate selected single neurons in monkeys' visual cortex. Such a methodology could be applied in the language domain by generating sentences that maximize activation in specific brain regions and validating predictions with new acquisitions.

Using sensor-wise similarity metrics. The brain score is a valuable metric predicting *each sensor independently*, but it has limitations. Specifically, while it provides information about individual sensors and can be used with continuous stimuli, it does not explicitly learn a brain representation space that can be interpreted geometrically. Thus, it limits our ability to analyze and visualize how words and sentences are represented in terms of distances. Other

methods, such as representational similarity (Kriegeskorte et al., 2008), have been proposed to characterize the pairwise stimulus correlation matrix of two representations (model and real neurons) for a given set of stimuli. This approach provides a population-level metric while capturing the similarity of the representations of the two populations. However, it does not provide information about individual sensors, and it is highly dependant on averaging brain responses across trials to subtract the effect of uninformative features. Thus, it is not compatible with single-trial studies, where the distances in the raw BOLD/MEG space are dominated by non-linguistic features such as motion, recording conditions, and stimulus modality. Future works could explore alternative linear methods compatible with population-level representations and single trial studies using continuous stimuli (Wegelin et al., 2006)[2]. These approaches may help to preserve a population-level representation space and potentially offer a geometric interpretation of how words and sentences are combined in the brain.

5.3.3 Building more interpretable encoding models

The interpretability of encoding models applied to naturalistic stimuli, compared to factorial designs, remains a major challenge. While encoding models offer high predictive power, reusability and can be used to analyse *natural* stimuli, they lack interpretability.

Clarifying the notion of interpretability. Interpreting language representations assumes that representations are generated from several underlying factors that account for specific linguistic features, such as tense, pronouns, and syntactic tree structure. A space will be considered as *interpretable* when it is *disentangled* into subcomponents, each subcomponent being generated from independent factors. Thus, in a disentangled space, each subcomponent is sensitive to modifications in specific factors (e.g. tense of the verb) and invariant to changes in others (e.g. plural/singular) (Bengio et al., 2014; Higgins et al., 2017). Brain representations exhibited by factorial designs are interpretable by construction, as the experimenter selects pre-existing generative factors (e.g. the constituent size and sentence length), generates corresponding stimuli, and splits the brain response into subcomponents varying with one selected factor but not the others. In contrast, the distributed representations of language models are not directly interpretable: their individual units are correlated with each other and related to multiple entangled linguistic features (Jawahar et al., 2019). Consequently, the encoding models based on deep language models that we utilize in our analyses pose challenges for interpretability.

[2] https://scikit-learn.org/stable/modules/cross_decomposition.html

Using disentangled encoding features. One promising approach to gaining interpretability in encoding models is to disentangle deep networks' activations into independent factors that are each sensitive to specific linguistic features. In Chapter 3, we demonstrated how this approach can be applied to lexical and compositional syntax and semantics, language specificity and the processing of multiple timescales. Specifically, we disentangled syntax and semantics by averaging the activations of language models over syntactically equivalent sentences, and we separated short-term and long-term dependencies by feeding language models with increasingly more context (Caucheteux et al., 2021a). While these methods do not require extra training and can be applied at inference time, their applicability is limited to a relatively small number of factors. More computationally expensive methods can be used, such as training the network on different datasets, each generated by a specific factor (Pasquiou et al., 2023; Millet et al., 2022), or using specific models designed to learn disentangled representations (S. Wang, Zhang, Lin, & Zong, 2020). For example, in Chapter 3, we separate acoustic, speech-specific, and language-specific representations by training the network on acoustic, non-native speech, and native speech datasets, respectively. In the field of computer vision, specific architectures have been developed that directly learn to disentangle the factors of variation in a training set (Higgins et al., 2017; Rolfe, 2017; Oord et al., 2018; Xiao et al., 2018). In natural language processing, a few models have been proposed to build disentangled representations, including via style transfer (Shen et al., 2017), controllable generation (Hu et al., 2018) (controlling for sentiment and tense), conditional generation based on prompts (Ouyang et al., 2022). Exploring the brain predictivity of the disentangled components is an exciting direction to better interpret brain responses to language.

Using deep language models as "in silico" models of the brain. The previous approach depends on our ability to disentangle representations in deep networks, which is not an easy task. Antonello & Huth (2022) proposed a slightly different approach to gain interpretability. In our work, we first disentangled deep networks activations and then predicted brain activity. In contrast, Antonello & Huth (2022) first train the encoding model and then use the encoding model as a simulator to investigate the synthetic brain responses elicited by new controlled stimuli, leveraging factorial designs without new acquisitions (see also (Eickenberg et al., 2017) for a similar approach in vision). While such "in silico" experimentation would not replace in vivo experiments, it can reduce the cost of hypothesis testing and generalizability, while leveraging the interpretability of factorial designs.

5.3.4 Generalizing hierarchical predictions to multiple layers and distances

In Chapter 4, we demonstrated that improving GPT-2's objective to predict a distant latent space increased its similarity with the brain. Our approach had limitations. First, we *fine-tuned* GPT-2 to predict *fixed* distant and latent representations. By using fixed targets, we avoided the problem of the model collapsing all inputs to the same point in the feature space. However, it is not compatible with training the model from scratch. Adapting this approach to be compatible with predicting variable and learnable targets will require methods to prevent model collapse, such as those proposed in (Baevski et al., 2020; T. Chen et al., 2020; He et al., 2020; Grill et al., 2020; El-Nouby et al., 2021; Bardes et al., 2022). Second, our approach was non-systematic (the model predicted layer 8, distance 5 from the current word), and we believe that algorithms should be enhanced with multiple levels of forecast. To achieve this, future work could draw inspiration from XLNet (Yang et al., 2020) and Data2Vec (Baevski et al., 2022). XLNet predicts randomly picked words in a sentence given its distant context while Data2Vec is trained to predict latent representations. By leveraging the permutation modeling approach of XLNet and the abstract objective of Data2Vec, we could build flexible models capable of predicting multiple time steps and levels of abstraction.

5.3.5 Improving NLP benchmarks

Although enhancing GPT-2 with high-level objectives did not result in improved performance on standard NLP tasks such as GLUE (A. Wang et al., 2018, 2020), this outcome may be partially attributable to the limitations of existing NLP benchmarks. Specifically, we contend that assessing hierarchical learning rules on downstream tasks that demand multi-step planning and long-term dependencies may yield more promising results. Downstream tasks in NLP such as summarization (Narayan et al., 2018; Hermann et al., 2015), question answering (Rajpurkar et al., 2016; Fan et al., 2019; Sinha et al., 2019), dialogue (Dinan et al., 2019), and multi-hop reasoning (Yang et al., 2018) may involve the need to predict abstract and distant representations of the future. Nonetheless, they are limited in two ways. First, while word-level metrics such as BLEU and ROUGE are commonly used to evaluate model performance on generation tasks, they only partially capture the semantic and syntactic aspects of the generated text and have limited correlation with human judgment (Papineni et al., 2002; Lin, 2004; T. Zhang et al., 2020). Second, these tasks often rely on a mixture of skills, such as fluent generation, commonsense, and knowledge, more than planning itself. Instead, more direct tests of planning abilities exist in the reinforcement learning community, such as navigation and continuous control tasks that

require the agent to plan a sequence of actions to control a robot in a simulated environment (Tassa et al., 2018; Savva et al., 2019). Adapting these tasks to target the planning abilities of NLP systems is an exciting direction for future work.

Overall, developing high-level tasks and evaluation metrics that more accurately reflect human judgments is a crucial challenge in NLP. I believe that such efforts will strengthen the relevance of hierarchical learning rules, and brain-inspired approaches in general.

5.4 Bridging neuro-linguistics and AI: advancements and challenges

One of the central challenges of my PhD research has been to compare two black boxes: the human brain and deep neural networks (Abnar et al., 2019). While some may question the value of such a comparison in advancing our understanding of neuro-linguistics and AI, I believe that our comparative analysis has the potential to offer valuable insights in both fields. In the following sections, I will discuss the objectives of neuro-linguistics and AI, and describe how our work has sought to contribute towards these goals.

5.4.1 Advancing neuro-linguistics with artificial neural networks

Goal of neuro-linguistics. Neuro-linguistics aims to understand and explain how the brain processes language. In their work, Jain et al. (2023) characterize this goal as identifying a scientific model that can effectively account for brain responses to language stimuli ($R = f_\theta(S)$, where R represents brain activity, f_θ the scientific model, and S represents the stimulus). The aim is then to find a model that exhibits high predictive accuracy, scope, and explainability.

- **Predictive accuracy** refers to the ability of the model f_θ to accurately predict brain activity R based on the experimental conditions and stimuli. The model should provide accurate and reliable predictions that are consistent with the observed data.
- **Scope** refers to the range of brain responses and stimuli that the model f_θ can account for. A good model should be able to explain a wide range of brain responses recorded with different modalities, in response to a variety of stimuli (e.g. audio or visual words, sentences and narratives).

- **Explainability** refers to the ability of the model f_θ to be decomposed into subcomponents that are useful for the experimenter. The subcomponents of interest depend on the scientific question and research goals. In the field of neuro-linguistics, a common approach to modeling brain responses involves describing them in terms of ***interpretable features*** that underlie specific processes, such as syntax and semantics. Complementary to this approach is another possibly inspired by artificial intelligence research, which favors describing the system in terms of simple principles, such as its ***architecture, objective function, and learning rule*** (Richards et al., 2019). These two approaches are not mutually exclusive; advancements in finding relevant features may lead to insights in the computational principles underlying language.

In the pursuit of developing accurate, explainable, and large-scoped models, neuroscience has explored various approaches. In the following, we review two commonly used strategies and demonstrate how the present study fits within this context.

Strengths and weaknesses of combining natural stimuli with deep encoding models. Two common methods to understand brain responses are contrast-based methods and encoding models (Jain et al., 2023). Contrast-based methods are usually combined with factorial designs and involve comparing brain responses to different controlled stimuli, while encoding models assume that f_θ belongs to a certain set of functions and estimate the learnable parameters using true brain responses to either controlled or natural stimuli. In our research, we use linear encoding models based on artificial neural network features to explain brain responses to natural language ($f_\theta = W_\theta \cdot X_S$, with X_S the deep nets activations in response to the natural story S heard by the participants). This combines advantages of using natural stimuli, encoding models and the rich distributed features of artificial neural networks.

- Natural stimuli
  In contrast to factorial designs, which often employ carefully selected stimuli that are matched for various linguistic features like word length, word frequency (Kutas & Federmeier, 2011), and/or constituent size (Pallier et al., 2011; Ding et al., 2016), natural stimuli possess **ecological generalizability** (Hamilton & Huth, 2018). This characteristic allows experimental setups to more closely resemble real-world settings, potentially increasing the generalizability of findings to out-of-lab conditions. Furthermore, the use of natural stimuli allows for the inclusion of **larger cohorts** of participants and facilitates the re-use

of experiments to study a wide range of linguistic phenomena, thereby increasing the statistical power of analysis and promoting collaboration and **replication** across laboratories (Jain et al., 2023).

- Encoding models
  Second, linear encoding models offer several advantages over contrast-based methods like **scale** and **re-usability** (Jain et al., 2023). Contrast-based methods compare brain activity of participants in response to different conditions, and are partially compatible with natural stimuli. For instance, Lerner et al. (2011) compared brain responses to natural stories and stories with shuffled words, sentences, and paragraphs to study the timescale of brain representations. However, this approach requires a large number of recordings (one per participant per condition) and still relies on unnatural stimuli. In contrast, linear encoding models assume that f_θ is a linear transform of a feature space X_S describing the stimulus: $f_\theta = W_\theta \cdot X_S$ (Jain et al., 2023). W_θ is estimated using brain response to the same natural stimulus, and a **wide range of phenomena** can be studied by varying the feature space X_S, without new acquisition.

- Implicit encoding features
  The feature space X_S can be simple linguistic features like an indicator of the part-of-speech of a word, or more complex features. Explicit linguistic features allow greater interpretability but only account for specific phenomena identified *a priori* by the experimenter. In contrast, we here use the implicit features of artificial nets; they are **rich semantic features** but are distributed, correlated, and thus **hard to interpret**.

Overall, while factorial designs combined with controlled stimuli have great interpretability, they suffer from reduced predictivity and scope. On the contrary, encoding models based on deep language algorithms have high predictive power and a wide scope, but low interpretability. Below, we show how this manuscript highlights the strengths of such method and attempts to overcome its limitations.

Improved brain predictivity and scope. In Chapter 2, we demonstrate that linear encoding methods combined with deep language models allows to significantly predict brain activity in wide range of experimental settings, namely in response to isolated words, sentences, and narratives, recorded with MEG/fMRI, across large cohorts of more than 500 participants, for more than thirty artificial neural networks architectures. The robustness of the effect is

reinforced by an increasing number of studies evidencing a significant prediction for various stimuli and brain recordings modalities (Jat et al., 2019; Hollenstein et al., 2019; Schrimpf et al., 2021; Toneva, Stretcu, et al., 2020; Toneva, Mitchell, & Wehbe, 2020a,b; Toneva & Wehbe, 2019; Reddy & Wehbe, 2020; Sun et al., 2021; Anderson et al., 2021; S. Wang, Zhang, Wang, et al., 2020; Vaidya et al., 2022; Jain et al., 2023), and showing that artificial nets better predicts brain activity than discrete linguistic features such as node count, part-of-speech and dependency tags (Reddy & Wehbe, 2020).

Decomposing syntax, semantics, lexical, contextual representations. The representations of deep networks are difficult to interpret due to their distributed nature (Abnar et al., 2019). In Chapter 3, we attempt to address this issue by decomposing the activations of deep neural networks into subcomponents and quantifying the brain predictivity of each disentangled subcomponent. Our approach reveals a finer-grained decomposition of the spatial and temporal hierarchy of natural language, language specificity, as well as syntactic and semantic processes in the brain. In particular, we find a distributed (Fedorenko et al., 2020), rather than modular (Friederici et al., 2000; Friederici, 2011), view of syntactic processes, highlighting a large recruitment of compositional semantics, as well as a hierarchical organization of speech processing along the temporo-parietal axis with high granularity (Lerner et al., 2011).

Computational principle underlying language. In Chapter 2, we investigate the properties that contribute to the brain predictivity of deep neural networks, focusing on architecture and objective function. To this end, we compared thousands of models trained on the same dataset, but varying in performance, architecture, and training duration, and examined their brain score. Our findings revealed that the **models' ability to predict the next word primarily affect the brain score**, with architectural parameters having a lesser impact. Notably, we observed a non-monotonic relationship between next-word prediction performance and the brain score, wherein the very best models slightly deviate from the brain while still improving in their next-word prediction task. These findings suggest that although next-word prediction enables the emergence of brain-like representations, it may also be too specific compared to the brain. In line with this concept, our Chapter 4 research showcases that fine-tuning GPT-2 with a **long-range and hierarchical objective enhances its resemblance with the brain**. This notion stems from predictive coding theories, where the brain continually forecasts sensory inputs, matches them against actual inputs, and updates its internal model (Rumelhart & McClelland, 1982; Rao & Ballard, 1999; K. Friston & Kiebel, 2009). Our study in Chapter 4 further refines

this hypothesis by revealing that individual regions of the cortical hierarchy specialize in predicting various temporal scopes and levels of representations. Thus, our findings reinforce the significance of next-word prediction in the brain (Heilbron et al., 2022; Schrimpf et al., 2021; Goldstein et al., 2022) while also proposing that the brain predicts more distant and abstract representations.

Overall, our works highlight the strong predictive power and wide scope of encoding models based on artificial neural networks, while acknowledging their limited yet improvable interpretability. We believe that manipulating the activations of language models can enhance our understanding of the *features* encoded in brain responses, and carefully comparing different neural networks may help clarify the *computational principle* that govern language processing in the brain.

5.4.2 Advancing AI through brain-inspired models

Goal of artificial intelligence in the context of language. The ultimate objective of Artificial Intelligence (AI) remains a widely debated topic. There are various proposed concepts such as Artificial General Intelligence, Human-Level Intelligence, Strong AI, and Universal AI, among others (Legg & Hutter, 2007; McCarthy, 2007; Searle, 2009; Goertzel, 2014). These concepts differ in their end-goals, ranging from thinking like humans to simply perform well at predefined tasks. Nonetheless, they share the common objective of developing intelligent systems capable of performing diverse tasks in various environments. In the context of natural language processing, this goal translates into **building systems that achieve human-level performance in all language tasks humans could possibly do**, such as dialogue, machine translation, question answering, story generation and text synthesis.

Language models still fall short of human-level performances. Recent large language models have made significant progress in NLP, approaching the goal of human-level performance. However, they still struggle with certain aspects of language understanding that come naturally to humans. In a recent study, Mahowald et al. (2023) demonstrated that even the most advanced language models, such as ChatGPT, still struggle in understanding and using language in a real-world context. As detailed in the Introduction, large language models fail at some tasks such as handling long-span memory, retrieving information, multi-step planning, inductive and commonsense reasoning (Mahowald et al., 2023; Bang et al., 2023). Additionally, these models lack consistency in their responses and are highly sensitive to the quality and quantity

of data used for training (Elazar et al., 2021; Brown et al., 2020; Mahowald et al., 2023). Thus, while the progress made in NLP is undoubtedly impressive, there is still work to be done to bridge the gap between artificial and human language processing.

Neuroscience as a source of inspiration and validation for human-level intelligence. The question of whether neuroscience can help close the gap between humans and artificial neural networks is a long-standing debate, extending beyond the domain of language. It has fueled a history of advances in AI inspired by or developed in conjunction with cognitive neuroscience, including early artificial neural networks, back-propagation, convolution, and, more generally, deep learning frameworks (McCulloch & Pitts, 1943; Hebb, 1949; Turing, 1950; Hubel & Wiesel, 1959; Rumelhart et al., 1986; LeCun et al., 1989, 2015; Schmidhuber, 2015). Given the vast space of possible solutions to human-level AI, investigating how the brain works may guide the search. As highlighted by Hassabis et al. (2017), there may be two primary benefits of developing AI based on our understanding of how the brain works:

> *"First, neuroscience provides a rich source of inspiration for new types of algorithms and architectures, independent of and complementary to the mathematical and logic-based methods and ideas that have largely dominated traditional approaches to AI. [...] Second, neuroscience can provide validation of AI techniques that already exist. If a known algorithm is subsequently found to be implemented in the brain, then that is strong support for its plausibility as an integral component of an overall general intelligence system."*
> *(Hassabis et al., 2017).*

Recent papers, including the work by Zador et al. (2022), also advocate for a more neuroscience-driven approach to AI. However, some critics have questioned the ambiguous link between neuroscience and advances in deep learning. For instance, Sam Gershman argued that

> *"new engineering ideas come from thinking about the structure of problems, not reading the tea leaves of biology"*[3].

To clarify the ongoing debate, it may be worthwhile to consider the **appropriate level of analysis for drawing inspiration from neuroscience**. Should we do so at the computational level (i.e., the goal the two systems optimize), the algorithmic level (i.e., the computations underlying such a goal), or the implementation level (i.e., the physical substrates of the systems) (Marr & Poggio, 1976)? For instance, when the objective is to design flying machines, such

[3] https://twitter.com/gershbrain/status/1583785657767366656?s=20

as planes, building a biological system with feathers may not be necessary. Nevertheless, comprehending the algorithmic properties that enable the balance of lift, weight, and thrust and the capacity to make minor adjustments to the direction and speed of flight can prove beneficial.

In the following paragraphs, we show how the present thesis attempts to leverage brain activity to probe the generality of current AI systems, as well as propose brain-inspired objective functions.

The brain score as a generalization test. Evaluating the ability of machines to generalize across all human language tasks is challenging. While *behavioral* benchmarks have been utilized to assess machine performance across multiple tasks, this approach faces significant limitations (A. Wang et al., 2018, 2020; Srivastava et al., 2022). Formalizing and testing all human tasks is impossible, and designing tasks to target specific high-level properties is often difficult, leading to biases in the evaluation process (Bowman & Dahl, 2021). As an alternative, evaluating algorithms based on their *internal representations* may offer a promising approach. If two systems build the same underlying constructs, they may respond similarly to new stimuli. Building on this idea, we assess the similarity between human and machine internal representations, and use such metric as a complementary way to evaluate the generality of deep neural networks' internal constructs. By evidencing high-level similarities with brain activity, the present study strengthens the generality of the representations learnt by recent language models.

Drawing inspiration from hierarchical objective functions. Next-word prediction has enabled significant advancements in natural language processing (Radford et al., 2019; Brown et al., 2020). However, our research suggests that such task may be limited. Specifically, Chapter 2 of our manuscript shows that the best-performing models slightly diverge from brain-like representations at the end of training, while improving on their next-word prediction task. In Chapter 4, we elaborate on this idea and provide evidence that augmenting GPT-2 with the ability to predict long-range and hierarchical representations improves its similarity to the brain. Our proposal roots in predictive coding theories (Rumelhart & McClelland, 1982; Rao & Ballard, 1999; K. Friston & Kiebel, 2009), and aligns with recent research in vision and speech domains that have explored high-level training tasks (Baevski et al., 2020; T. Chen et al., 2020; He et al., 2020; Grill et al., 2020; El-Nouby et al., 2021; Bardes et al., 2022; LeCun, 2022). These studies share a common objective of predicting latent representations while addressing the issue of model collapse in various ways. For instance, SimCLR and MoCo leverage contrastive

learning along with either a large batch size or a queue to store a large number of negative examples (H.-H. Chen & Cherkassky, 2020; He et al., 2020). VicReg introduces regularization terms to control for sufficient variance in activations while favoring uncorrelated representations (Bardes et al., 2022; LeCun, 2022). BYOL predicts representations of a separate encoder model consisting of an exponential moving average of the model weights (Grill et al., 2020). Alternatively, Baevski et al. (2020) propose predicting latent and quantized representations of masked frames using a contrastive loss combined with a diversity loss. In the field of natural language processing, several studies have investigated alternative learning rules to causal and masked language modelling (Jernite et al., 2017; Fan et al., 2018; Devlin et al., 2019; Lewis et al., 2019; Yang et al., 2020; Joshi et al., 2020; Clark et al., 2020; Baevski et al., 2022). However, none of these methods have combined both long-range and high-level predictions. The present research is a first step towards filling this gap and contributes to the growing body of research exploring high-level training tasks.

Overall, we argue that drawing *inspiration* and *validation* from neuroscience can provide a significant advantage in the development of human-level AI, even if the optimal level of analysis to achieve this remains uncertain.

5.4.3 Closing the gap: what's missing to current AI systems?

Large language models have made significant strides in natural language processing, but they still fall short in several key areas when compared to the human brain. First, they still underperform humans in language tasks involving long-term memory, planning, inductive and commonsense reasoning, logic and math, meta-learning, as well as handling continual learning, uncertainty and data heterogeneity (Elazar et al., 2021; Bang et al., 2023; Mahowald et al., 2023). Second, the brain score remains low, and the representations of deep language models are only partially aligned with the brain. Third, they are trained on unrealistic amounts of data, much more than what a child is exposed to (Warstadt & Bowman, 2022), and are highly sensitive to the quality of the training datasets. Fourth, the brain remains a much more efficient system than artificial neural networks. The question of **how to close these gaps** is a topic of ongoing research, and I believe that seeking inspiration from the brain may provide viable solutions. This may entail exploring alternative computational principles, such as objective functions, learning rules, and network architectures that more closely mirror the organization and function of the human brain. However, some researchers may disagree with this view, particularly in light of the impressive performance of large language models trained on larger

and higher-quality datasets (Kaplan et al., 2020). In the following paragraphs, I will discuss some of the proposed solutions for addressing these gaps.

Training on larger and higher quality dataset is inefficient and biologically implausible. Recent works have achieved remarkable performance by training models on larger amount of higher quality datasets (Hoffmann et al., 2022; Brown et al., 2020; Touvron et al., 2023). However, this approach has limitations. First, the amount of available data is finite, and the data curation process is expensive and time-consuming. Second, it is biologically implausible. Current models are heavily reliant on the curation process, and a few carefully curated datasets for fine-tuning may yield better results than large amounts of non-curated data (Solaiman & Dennison, 2021). In contrast, the brain is adept at handling heterogeneous datasets in a variable and unpredictable environment, and children learn language from less and noisier data than current models (Warstadt & Bowman, 2022). Therefore, I speculate that the performance gains brought by training on larger and higher quality datasets will eventually plateau, and it will not be sufficient to achieve human-level AI.

Directly optimizing for human behavior won't be enough. Optimizing deep networks to match human behavior has gained popularity with the recent success of ChatGPT (Ouyang et al., 2022). ChatGPT relies on pre-training and a second step that includes both a fine-tuning task on instructional data and reinforcement learning from human feedback. The second step is critical for ChatGPT's dialogue performance and requires the annotation of thousands of texts. While ChatGPT yields impressive results, increasing the quality and quantity of annotated data is limited by its cost and is fundamentally different from how humans learn language. Therefore, in my view, relying more on human feedback suffers from both practical and theoretical limitations and is unlikely to be sufficient to achieve human-level AI.

Directly optimizing for brain-like representations is limited by the amount and quality of brain recordings. While it may seem more direct to optimize deep models to match brain recordings in order to achieve brain-like representations and gains in NLP performances, this approach is currently limited by the scarcity and noise of brain recordings. A study by Schwartz et al. (2019) explored this path by fine-tuning a BERT model to predict MEG and fMRI data, which improved the brain score but did not yield significant improvements in NLP tasks. One potential alternative would be to use a mixed loss that combines a brain-oriented objective with a language modelling objective. However, such a method still requires a sufficient amount of

high-quality brain recordings, which are currently limited. Therefore, it may be more beneficial to draw inspiration from the brain to identify relevant computational principles such as the architecture and objective function, and then optimize the model weights on textual data.

The role of multi-modality and grounding. The need for multimodality has gained popularity in recent months, with significant advancements in multi-modal generative models (Ramesh et al., 2022; Rombach et al., 2022), and the recent release of GPT-4[4]. This approach recognizes the importance of integrating rich multimodal inputs to achieve a deeper understanding of situations, rather than treating language as a standalone entity (Hasson et al., 2018; McClelland et al., 2020; Bisk et al., 2020). Human cognition involves constructing situation representations, and relating information to familiar contexts enhances our understanding and memory recall (McClelland et al., 2020; Bransford & Johnson, 1972). However, whether these elements are essential for achieving human-level intelligence remains unclear. For example, blindness does not appear to impair language or reasoning abilities (Knauff & May, 2006), and it is uncertain whether multimodal models are better suited for highly unpredictable and variable environments. While I recognize the practical benefits of integrating multiple modalities, additional factors may be necessary for achieving human-level intelligence.

The ambivalent role of the model size. The exact role of the model size remains ambiguous. On the one hand, training larger models has resulted in improved performance in NLP (Brown et al., 2020; Chowdhery et al., 2022; Hoffmann et al., 2022; Touvron et al., 2023), and even the largest models of today contain significantly fewer weights than the brain has synapses[5], which suggests a path for improvement. On the other hand, scaling the model size has limitations imposed by hardware constraints (Chowdhery et al., 2022; Hoffmann et al., 2022). Drawing inspiration from the brain may provide a path towards computational efficiency. For instance, training GPT-3 would require 1000 megawatt-hour, a consumption rate that vastly exceeds that of the human brain at 20 watt-hour (Patterson et al., 2007; Zador et al., 2022). Some differences may be noted, as analysed in (Zador et al., 2022). For instance, the energy cost of transmitting information depends on the amplitude of activation in the brain, while it is the same for a ”0” or a ”1” in digital signal processing. Similarly, the energy cost of transmitting information from one part of the brain to the other is different from transmitting information locally, which is not

[4] https://openai.com/product/gpt-4

[5] We consider that each of the 90 billion neurons have around 10,000 synapses, and take a 90 billion parameter model as reference. For comparison, GPT-2, Chinchilla, GPT-3 have 1.5, 70 and 175 billion parameters, respectively.

the case in a single layer of Transformer. These insights suggest that new approaches to model design, that take inspiration from biological processes, may provide a way to achieve better efficiency. In summary, I acknowledge the potential power of high-dimensionality models, yet we also recognize the limitations imposed by hardware constraints. As such, simply increasing the size of models is currently not a feasible solution for achieving human-level AI.

Brain-inspired objective functions and architectures. We provided several elements that show how studying the computational principles underlying natural language processing in the brain will likely benefit AI. Firstly, our work calls for higher-level **objective functions** based on evidence of hierarchical predictions in the brain, human planning over multiple timescales, and the limitations of next-word prediction in handling uncertainty (LeCun, 2022; Caucheteux et al., 2023). Predicting low-level representations of stimuli can be a noisy objective, as words become highly indeterminate over a lifetime. Secondly, although **architectural modifications** have fallen out of popularity, the current architecture of Transformers is limited by a fixed context window, hindering its ability to retain –and forget– information on the scale of a lifetime. Thus, there may still be room for architectural innovations, particularly in flexible memory management (Beltagy et al., 2020; Raikote, 2021). More generally, I believe that drawing inspiration from cognitive processes widely studied in neuroscience but currently lacking from the best algorithms, such as episodic and semantic memory, continual learning, one-shot generalization, inductive reasoning, imagination, and multi-step planning, is an exciting path to identify the computational principles underlying intelligence and build better algorithms (Hassabis et al., 2017; Mahowald et al., 2023; Bang et al., 2023).

Overall, achieving human-level AI may require multiple avenues of investigation. In my opinion, simply training models on larger, multi-modal, and higher-quality datasets will not suffice. Rather, I believe that drawing inspiration from the computational principles of the brain – including its architecture and objective functions – may offer valuable insights for advancing AI towards its goal.

5.5 Conclusion

To conclude, the present manuscript evidences similarities and differences between language representations in brains and algorithms. The internal representations of the two systems are partially aligned and share corresponding hierarchies. By leveraging such similarities, we show

how language algorithms may help build more accurate encoding models of brain activity and provide a finer grained decomposition of several linguistic processes in the brain. By identifying their differences, namely the ability to make hierarchical predictions, our approach paves the way for building more brain-like language algorithms. The questions of whether AI will drive progress in neuroscience, and whether neuroscience will contribute to the success of AI models in the future, remain open. However, by developing approaches to gain interpretability in AI systems and refine our understanding of brains' adequate levels of analysis, we believe that a fruitful collaborations between AI and neuroscience will pave the way for significant progress in both fields.

Chapter 6

Appendix

6.1 Brains and algorithms partially converge in natural language processing

6.1.1 Average brain responses to reading

When and where do textual sentences elicit brain activity? As expected (Fedorenko et al., 2020; Dehaene & Cohen, 2011; Hagoort & Indefrey, 2014; Hickok & Poeppel, 2007), average fMRI and MEG responses to written words peak in a distributed and bilateral cortical network, including the primary visual cortex, the left fusiform gyrus, the supra-marginal, and the superior temporal cortices, as well as the motor, premotor and infero-frontal areas (Figure 2.2a). MEG source reconstruction, based on structural MRI and minimum norm estimates, further clarifies the dynamics of this cortical network: on average, word onset elicits a series of brain responses originating in V1 around ≈100 ms and continuing within the left posterior fusiform gyrus around 200 ms, the superior and middle temporal gyri, as well as the pre-motor and infero-frontal cortices between 150 and 500 ms after word onset (Figure 2.2a).

6.1.2 Shared-response model (or noise ceilings)

Shared-response model (SRM) comparison (often referred to as "noise ceiling"), allows us to evaluate the extent to which individual subjects' brain responses can be explained with a model-free approach (Caucheteux et al., 2021b) and can serve as a proxy for a signal-to-noise ratio analysis. For this, we fit, for each subject separately, an SRM model (or noise-ceiling): for each recording of each subject and each sentence Y_{train}, we fit a linear model W from the recordings of all other subjects who read the same sentence X_{train} to predict each voxel and

each MEG sensor at each time sample, separately. Using a cross-validation scheme across sentences, we then evaluate the Pearson correlation R between (1) the true brain responses of subject Y_{test} and (2) the predicted brain responses $\hat{Y}_{\text{test}} = W \cdot X_{\text{test}}$ for each voxel and each MEG sensor separately. This procedure can be thought of as approximating an optimal black box: i.e. evaluating a one-hot encoder of brain responses is trained and evaluated on each element of a unique sentence. Noise ceiling peaks within the expected language network (Fedorenko et al., 2016) (Figure 2.1f-h). These estimates are relatively low: for example, fMRI noise ceilings reach, on average, $R = 0.129$ (± 0.004 SEM across subjects) in the superior temporal gyrus, whereas MEG noise ceilings peak at $R = 0.069 \pm 0.001$ (Supplementary Table 1).

6.1.3 Probe analysis of the language transformer

Middle layers better map onto brain responses than input and output layers. Why is there such a difference between layers? To tackle the question, we measure the level to which the 32,400 transformer embeddings linearly predict two types of linguistic features: part-of-speech (i.e a lexical feature), and the number of open and pending nodes (i.e compositional syntactic features (Nelson et al., 2017)). More precisely, we fit and evaluate an ℓ_2-penalized linear model to predict each of these features given the transformer's embedding and plot this decoding performance as a function of the language performance of the model (Figure S2). While the word embedding and middle layers similarly predict word-level features (word length and part-of-speech of the word), the two high-level syntactic features (number of open and pending nodes) are better predicted by the middle layers of transformers. Finally, the decoding performance of the two syntactic features varies with the layer and the performance, in a manner strikingly similar to the brain score. These analyses suggest that middle layers are more brain-like than extremity layers because they learn to encode abstract linguistic properties like syntax.

6.1.4 Definition of compositionality

Following a recently proposed taxonomy (Caucheteux et al., 2021a), we formally define "compositional" as the language representations that cannot be explained by the linear combination of lexical representations.

This definition may not be fully aligned with the many definitions of compositionality proposed over the years (Szabó, 2004). Specifically, some linguists restrict compositionality to the limited, generally invertible, combinations of words that follow the laws of syntax, and would consequently thus prefer the term "contextual". We believe, however, that the latter

term does not clearly point to the representations that are more than the sum of their parts (Pelletier, 1994) which is critical to the present analyses (Figure 2.3).

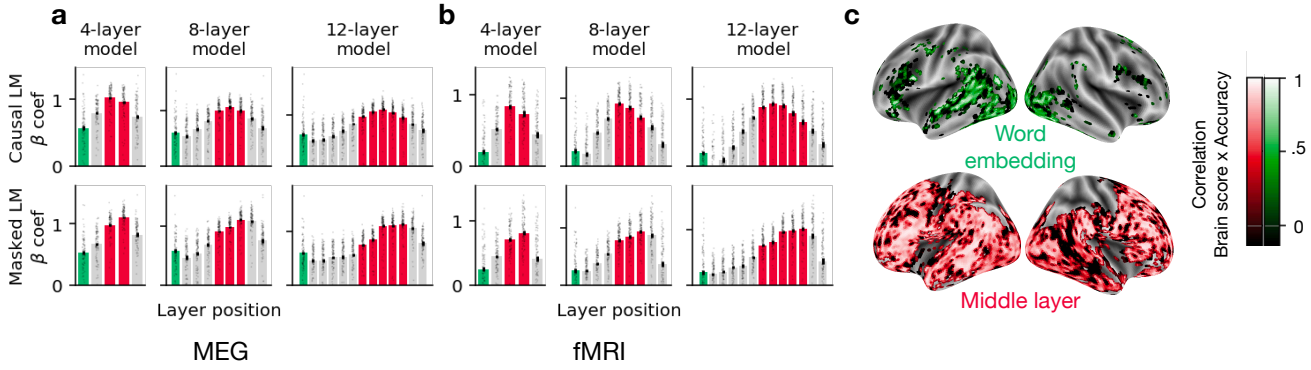


Figure S1: Correlation between the network's performance and brain score. a-b. Standardized beta coefficients between the language modeling performance of the network and its MEG (a) or fMRI (b) scores. For each subject, the brain scores are first scaled (0-mean, 1-std). Then, a linear regression is fit to predict the brain score (averaged across channels and time for MEG, across voxels for fMRI) of each layer of 100 networks (all 512-dimensional, with 12 layers and 8 heads) given their language performance (top-1 accuracy). The beta coefficients of the language performance are reported (y-axis). Results are consistent across 4-, 8-, and 12-layer transformers, trained on a causal (top) or masked (bottom) language modeling task. Error bars are the standard error of the mean beta coefficients across subjects. **c.** Pearson correlation between the performance of the 100 transformers (all 512-dimensional, with 12 layers and 8 heads) and the brain score of their word embedding (top) and ninth layer (bottom), for each voxel. Correlation scores are computed for each (subject, voxel) pair, then averaged across subjects. Only significant voxels are displayed, as assessed with a two-sided Wilcoxon test across subjects and corrected for multiple comparison using false discovery rate across voxels (threshold: .001).

Fronto-polar cortex:	0.054 ± 0.003	$p < 10^{-8}$
Fusiform:	0.120 ± 0.004	$p < 10^{-8}$
Infero-frontal:	0.139 ± 0.005	$p < 10^{-8}$
M1:	0.042 ± 0.003	$p < 10^{-8}$
STG:	0.129 ± 0.004	$p < 10^{-8}$
Supramarginal:	0.078 ± 0.003	$p < 10^{-8}$
V1:	0.150 ± 0.006	$p < 10^{-8}$

Table S1: Average noise ceiling within each region-of-interest. Mean, standard error of the mean and p-values across subjects.

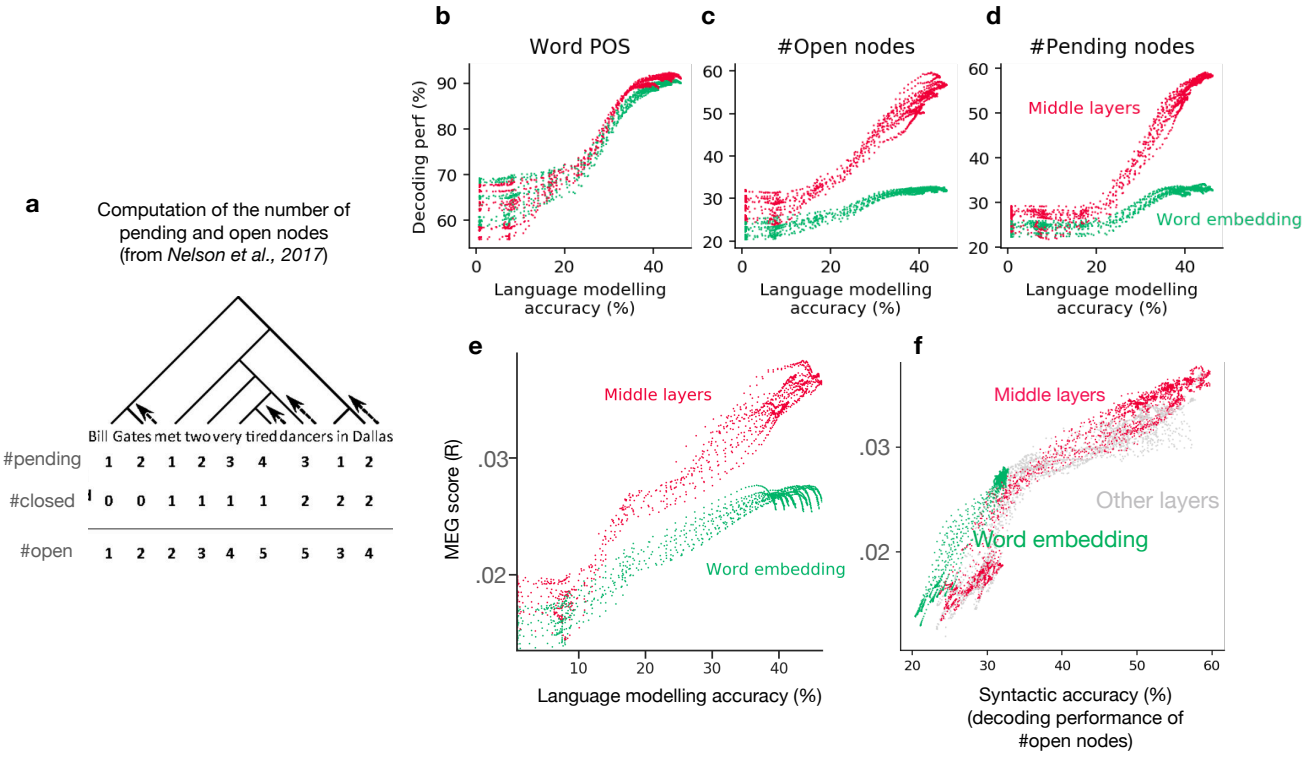


Figure S2: What linguistic information drives the brain score? a. From the stimulus, we compute three linguistic features: the part-of-speech of the words (i) (as given by Spacy), and two higher-level syntactic features: the number of pending nodes (ii) and open nodes (iii). These two syntactic features are derived from the constituency trees of the sentences, following (Nelson et al., 2017). **b-d**. A $\ell 2$-penalized linear regression is fit to predict the three linguistic features from the word embeddings (green), and middle layers (red) of the causal models studied in Figure 2.4b. The decoding performance is reported on the y-axis (accuracy at predicting the part-of-speech for b, r-squared for c, d and e). **e.** MEG scores (averaged across sensors and time) of the embeddings given their language modeling performance (top-1 accuracy at predicting the next word, Figure 2.4b). **f.** MEG scores of the embeddings given their ability to predict the number of open nodes.

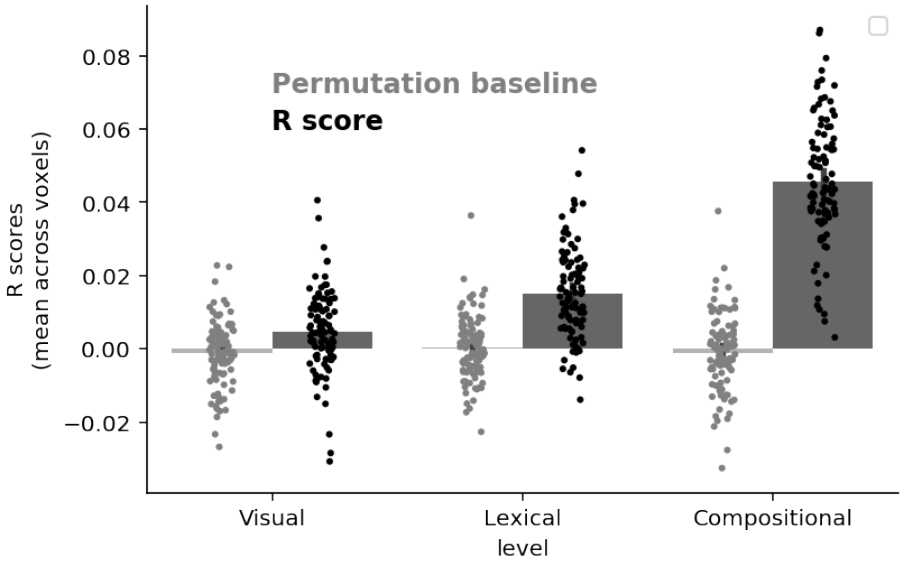


Figure S3: Permutation distribution. As a baseline, we compare the normal R scores (dark colors) to those of a permutation distribution (light colors) for each of the visual, lexical and compositional embeddings introduced in Figure 2.3. For each (subject, voxel) pair, we compute the mapping between the embeddings X and the fMRI of the subject, either (i) shuffled across time samples or (ii) without shuffling. Above, we report scores averaged across subjects and voxels. Error bars are standard-error of the mean across subjects (n=100).

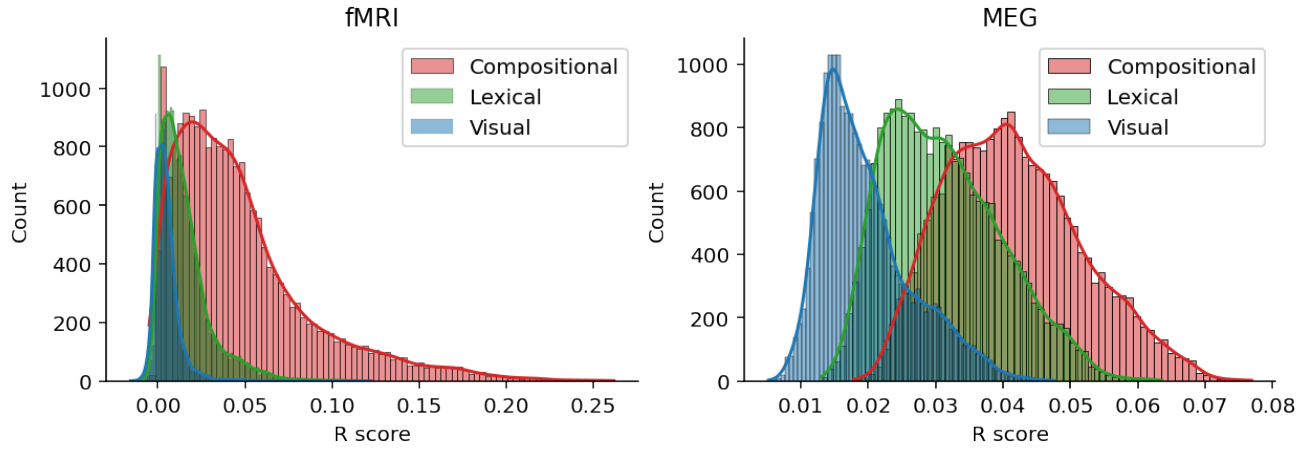


Figure S4: Distribution of R scores across fMRI voxels (left) and MEG sources (right). We compute the brain scores for the visual (blue), lexical (green) and compositional (red) embeddings introduced in Figure 3. We average scores across voxels (resp. sources) and subjects, to obtain one single score per voxel (resp. source). Above, the corresponding distribution of the R scores across voxels and sources.

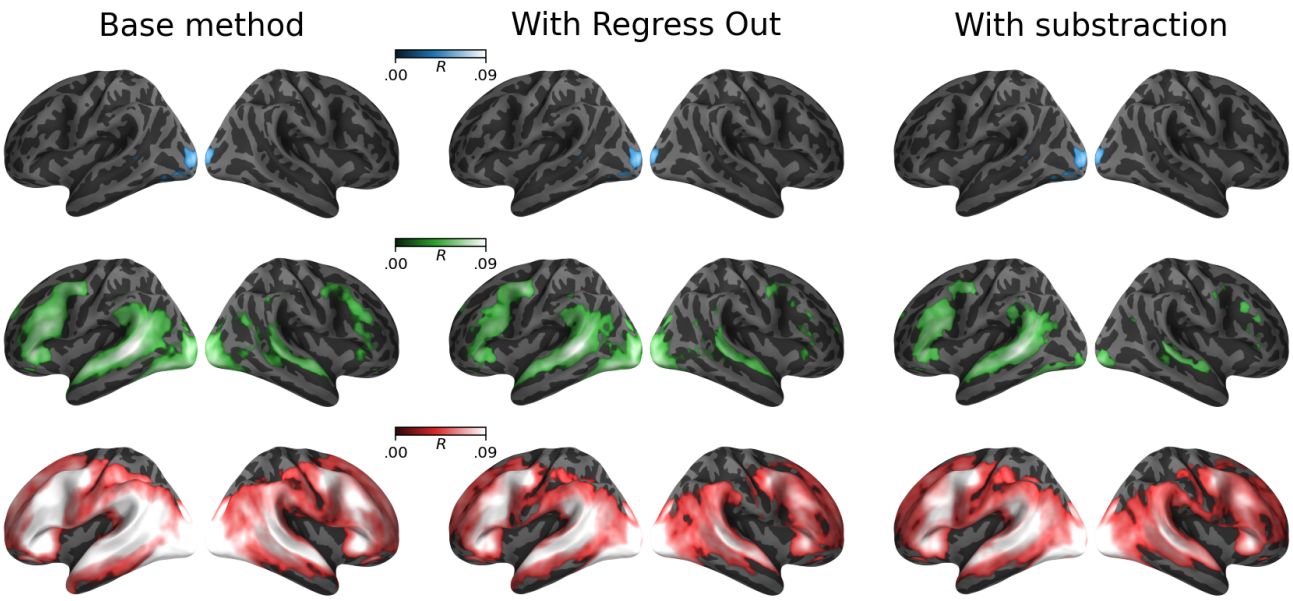


Figure S5: Comparison between two orthogonalization methods. In Figure 3, we report the raw brain scores (without subtraction) for the visual (blue, X_V), lexical (green, X_W) and compositional (red, X_C) embeddings ("base method" on the left). On the right, for each level, we subtract the scores of the level below (e.g. red scores $R_C = \mathcal{R}(X_C) - \mathcal{R}(X_W)$). In the middle, we orthogonalize the predictors before computing the brain scores, by"regressing ou" the effect of the lower level onto the current level. For the compositional score R_C, we fit a ridge regression model f (we use the RidgeCV implementation from scikit-learn, with 10 possible penalization values log spaced between 10^{-3} and 10^{8}) to predict X_C given the concatenation of the visual and word embeddings $X_V \oplus X_W$. Then, we compute the brain scores of the residuals $\tilde{X_C} = X_C - \hat{f}(X_V \oplus X_W)$. We proceed similarly for the lexical residuals $\tilde{X}_W = X_W \; X_V$. As we see, the subtraction method (right) is more conservative than the method with regress out (middle).

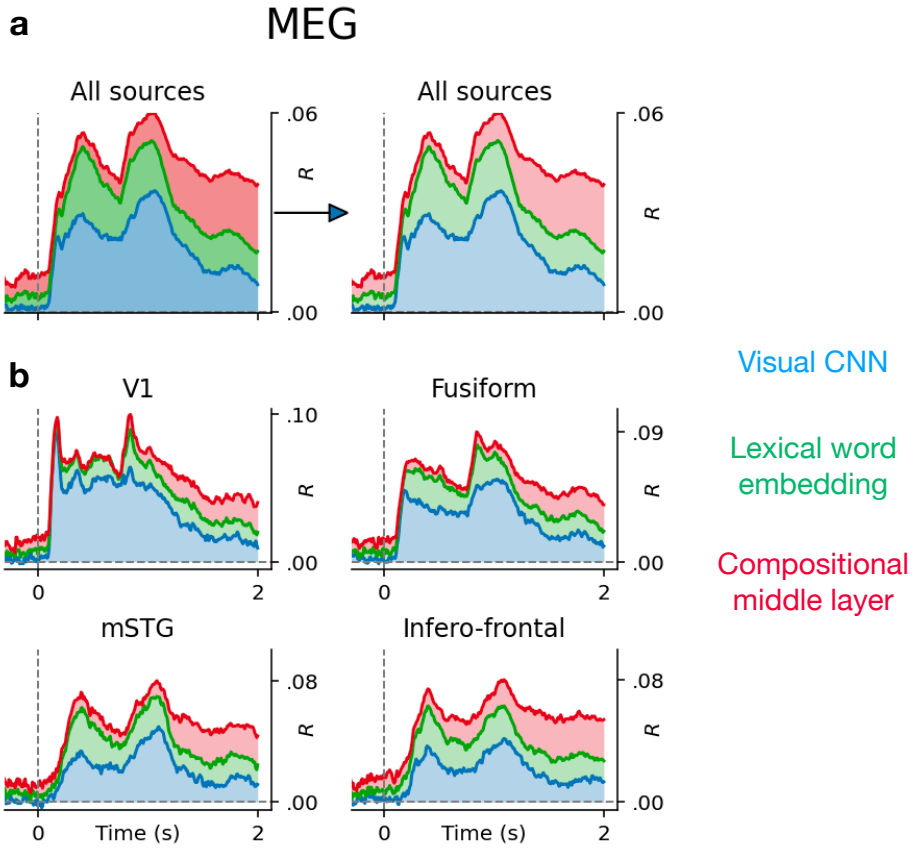


Figure S6: Brain scores over time. a) Same as Figure 2.3c, but without subtracting the scores of the level below. **b)** Same as Figure 2.3c without subtracting the scores.

Task	Dim	Layers	Heads	Best perplexity	Best accuracy
mlm	512	12	8	4.70	67.51
mlm	512	12	4	4.70	67.36
mlm	512	8	4	4.90	66.72
mlm	512	8	8	4.99	66.33
mlm	512	4	8	5.55	64.40
mlm	512	4	4	5.90	63.61
mlm	256	12	8	6.08	63.48
mlm	256	12	4	6.12	63.36
mlm	256	8	8	6.62	62.12
mlm	256	8	4	6.69	61.71
mlm	256	4	8	7.75	59.73
mlm	256	4	4	7.97	59.15
mlm	128	12	8	8.99	57.65
mlm	128	12	4	9.26	57.46
mlm	128	8	8	10.01	56.35
mlm	128	8	4	10.11	56.16
mlm	128	4	8	12.06	53.70
mlm	128	4	4	12.60	53.08
clm	512	12	8	15.00	46.47
clm	512	12	4	15.06	46.38
clm	512	8	4	15.49	46.01
clm	512	8	8	15.49	45.97
clm	512	4	8	16.75	44.93
clm	512	4	4	16.90	44.82
clm	256	12	4	17.85	44.28
clm	256	12	8	17.80	44.26
clm	256	8	8	18.69	43.68
clm	256	8	4	18.83	43.59
clm	256	4	4	20.67	42.53
clm	256	4	8	20.64	42.49
clm	128	12	4	23.26	41.47
clm	128	12	8	23.31	41.38
clm	128	8	4	24.45	40.83
clm	128	8	8	24.36	40.80
clm	128	4	4	27.11	39.61
clm	128	4	8	27.06	39.57

Table S2: Performance of the 36 transformer architectures. Best perplexity (the lower the better) and top-1 accuracy (the higher the better) of 36 transformer architectures, evaluated on a test test of ≈180K words from Wikipedia. Transformers are trained with a masked ('mlm') or causal ('clm') language modeling objective. They vary in their dimensionality ('Dim'), number of layers ('Layers') and number of attention heads ('Heads'). The models are trained on a set of ≈280K words from Wikipedia (in Dutch). The training is stopped when the perplexity on a validation set does not decrease for 5 epochs.

6.2 Deep language algorithms predict semantic comprehension from brain activity

6.2.1 Brain parcellation

In Figure 2.5B, E, and F, we used a subdivision of the parcellation from Destrieux Atlas (Destrieux et al., 2010). Regions with more than 400 vertices were split into smaller regions (so that each regions contains less than 400 vertices). The original parcellation consists of 75 regions per hemisphere. Our custom parcellation consists in 142 regions per hemisphere. In Figure 2.5G, we use the original parcellation for simplicity, and the following acronyms:

Acronym	Definition
STG / STS	Superior temporal gyrus / sulcus
aSTS	Anterior STS
maSTS	Mid-anterior STS
mpSTS	Mid-posterior STS
pSTS	Posterior STS
Angular / Supramar	Angular / Supramarginal inferior parietal gyrus
MTG / MTS	Medial temporal gyrus / sulcus
SFG / SFS	Superior frontal gyrus / sulcus
IFG / IFS	Inferior frontal gyrus / sulcus
Tri / Op	Pars triangularis / opercularis (IFG)
TTransverse	Temporal transverse sulcus
PCG	Posterior cingulate gyrus
STO	Temporo-occipital lateral sulcus

6.2.2 Mixed-effect model

Not all subjects listened to the same stories. To check that the $\mathcal{R}$ scores (correlation between comprehension and brain mapping) were not driven by the narratives and questionnaires' variability, a linear mixed-effect model was fit to predict the comprehension of a subject given its brain mapping scores, specifying the narrative as a random effect. More precisely, if $\mathcal{M}_{w_i} \in \mathbb{R}$ corresponds to the mapping scores of the i^{th} subject that listened to the story w, and $C_{w_i} \in \mathbb{R}$ refers to the comprehension scores, we estimate the fixed effect parameters $\widetilde{\beta} \in \mathbb{R}$ and $\widetilde{\eta} \in \mathbb{R}$ (shared across narratives), and the random effect parameter $\beta_w \in \mathbb{R}$ and $\eta_w \in \mathbb{R}$ (specific to the narrative w) such that:

$$C_{w_i} = (\widetilde{\beta} + \beta_w) \times \mathcal{M}_{w_i} + (\widetilde{\eta} + \eta_w) + \epsilon_{w_i}$$

with ϵ_{w_i} a vector of i.i.d normal errors with mean 0 and variance σ^2. In practice, we use the statsmodels (Seabold & Perktold, 2010) implementation of linear mixed-effect models. Significance of the coefficients were assessed with a t-test, as implemented in statsmodels.

6.2.3 Replication across single narratives

To further support that the $\mathcal{R}$ were not driven by the narratives' variability, we replicate the analysis of Figure 2.5D within single narratives. In Figure S7, we show that correlation scores between brain scores and comprehension scores are positive for each of the seven narratives.

6.2.4 Noise Ceiling Estimates

fMRI recordings are inherently noisy. Thus, we estimate an upper bound of the best brain score that can be obtained given the level of noise in the Narrative dataset. To this end, for each (subject, narrative) pair, we linearly map the fMRI recordings, not with the GPT-2 activations, but with the average fMRI recordings of the other subjects who listened to that narrative. More precisely, we use the exact same setting as in (1.1), but we predict $Y^{(s)}$, not from $g(X)$ (GPT-2's features after temporal alignment, of size $n_{\text{times}} \times n_{\text{dim}}$), but from the mean of the other subject's brains $\overline{Y} = \frac{1}{|S|} \sum_{s' \neq s} Y^{(s')}$ (of size $n_{\text{times}} \times n_{\text{voxels}}$). This score is called the noise ceiling for the (subject, narrative) pair. The noise ceilings for each brain region are displayed in Figure S8, and correspond to upper bounds of the brain scores displayed in Figure 2.5B.

6.2.5 Replication across the contextual layers of GPT-2

Previous analyses mostly focus on the eight layer of GPT-2. In Figure S10, we compute the brain scores of each layer of GPT-2, and report their correlation with the subject's comprehension scores. While the correlation with comprehension is the highest in layers 6-to-12 (and thus best explain comprehension's variability), our results do generalize to other contextual layers of GPT-2.

6.2.6 Distribution of regularization parameters

To quantify the mapping between the brain signals and GPT-2 activations, we use a ℓ_2-penalized linear regression (cf. Methods). To further investigate how penalization affected the brain score, we compute the optimal regularization parameter alpha for each (subject, narrative, voxel, fold), we average the alphas across (subject, narrative, fold) triplets, and report the corresponding

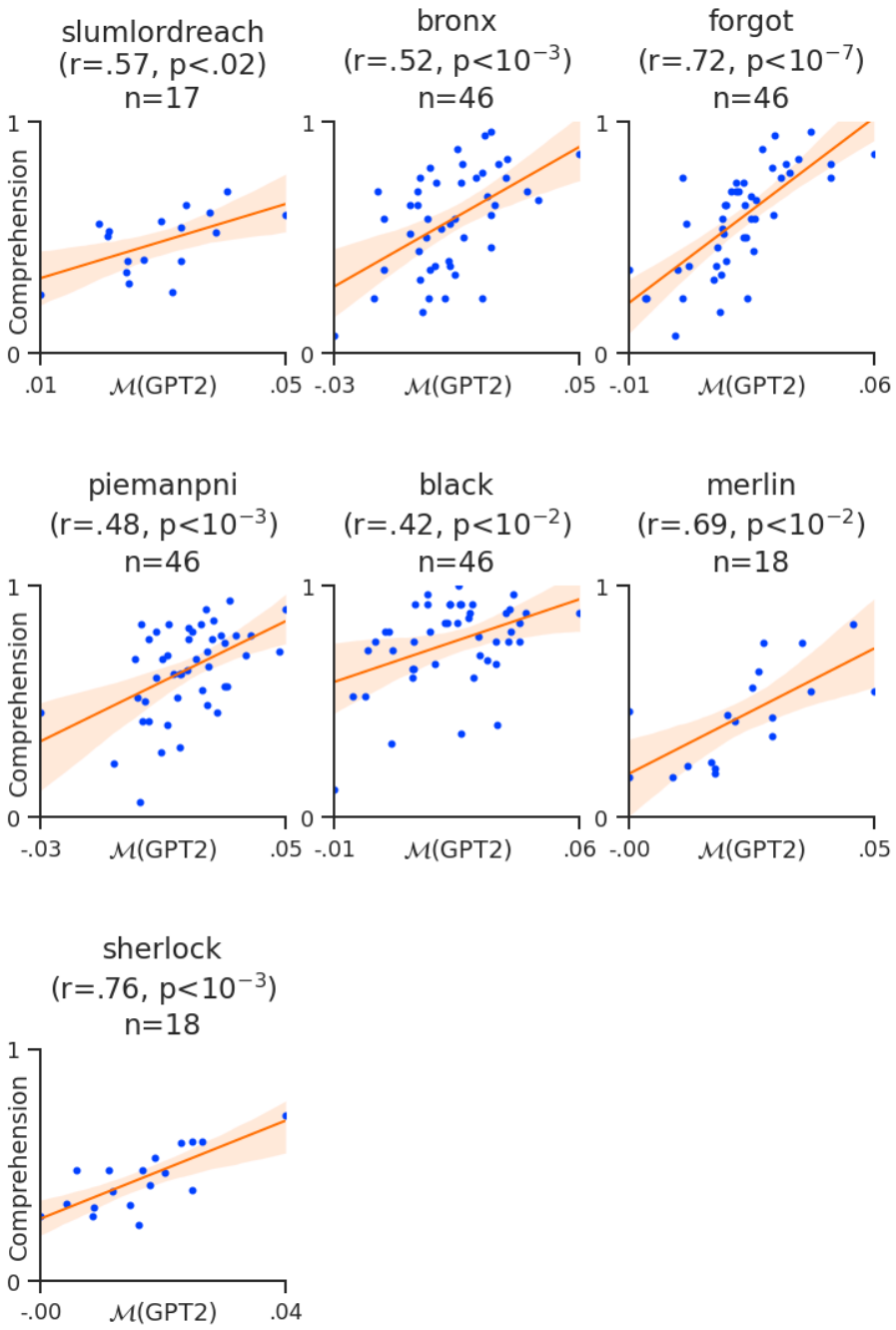


Figure S7: Replication within single narratives. Same as Figure 2.5D for each single narrative.

alphas across voxels (left) as well as the relationship between alphas and brain scores (on the right). As shown in Figure S11, regularization parameters are lower in regions commonly associated with language (auditory cortex, supramarginal, inferior-frontal areas) while higher

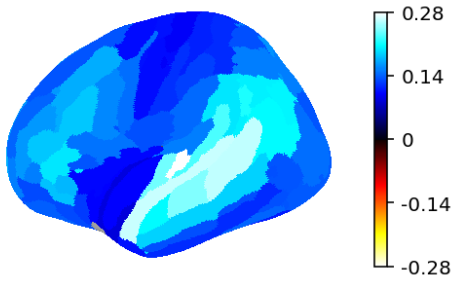


Figure S8: Noise ceiling estimates. Noise ceilings averaged across subjects, narratives and voxels within each region of interest. They are upper bounds of the brain scores in Figure 2.5B.

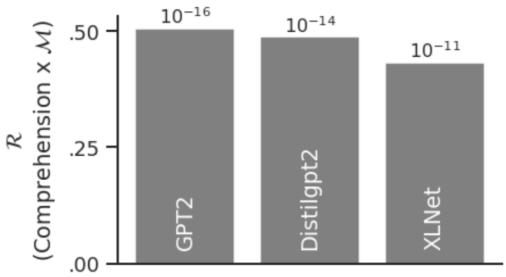


Figure S9: Replication to two other causal transformer architectures from Huggingface (XLNet base and Distilgpt2). The vertical axis shows the average correlation between (i) comprehension scores and (ii) brain scores. The top text displays the p-values of the corresponding correlation. The mapping scores were averaged across all voxels and the correlation with comprehension was computed, similarly to Figure 2.5D.

alphas (yellow) are associated with noisier regions.

6.2.7 Replication using partial correlation analyses

In Figure 2.5F, we compute the specific contribution of phonological, lexical and compositional features, respectively. To do so, we favor the simplest and most conservative method by using hierarchical modeling, which consists of computing the brain score of the two sets of features (e.g. Word Embedding vs. Layer 8) and then subtracting the scores. This approach is particularly conservative: the explainable variance shared by two sets of features is by definition fully attributed to the lower-level feature set (i.e. Word Embedding). Thus, our method tends to underestimate the variance specific to deeper layers. The fact that these effects remain largely above chance is thus good evidence that this layer captures representations specifically predictive of comprehension.

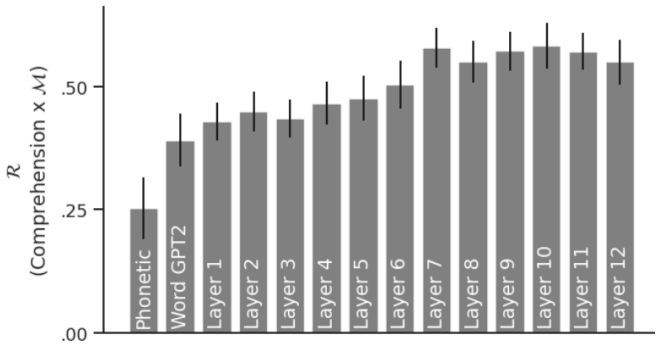


Figure S10: Correlation between comprehension scores and brain scores, for each layer of GPT-2 as well as phonetic features. Error bars are the standard errors of the means across subjects.

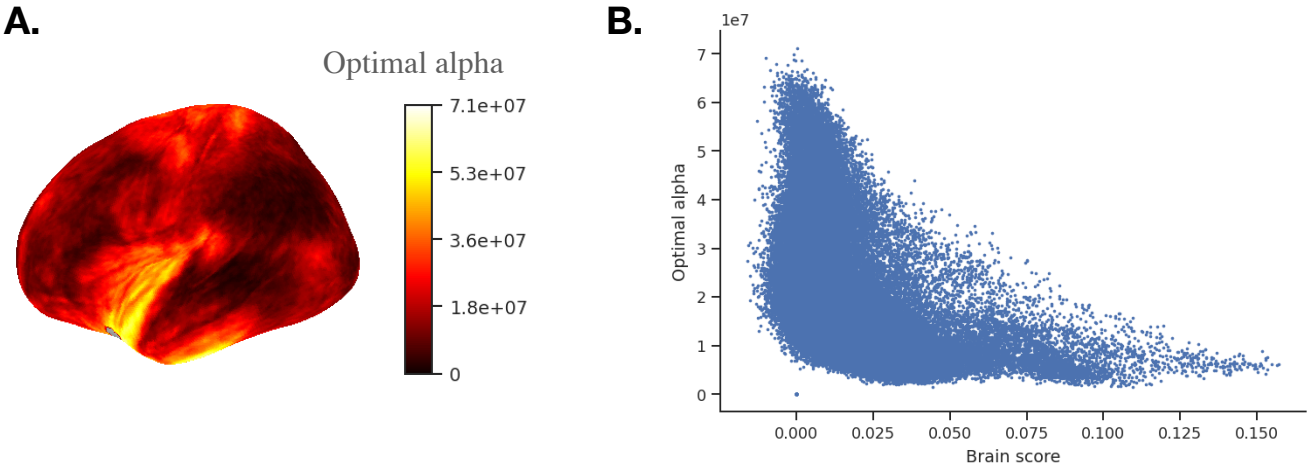


Figure S11: A) Optimal regularization parameters alpha (log-scaled) across voxels. A penalized regression is fitted for each (subject, narrative, voxel, fold) and the corresponding optimal regularization parameters alphas are extracted. Alphas are averaged across (subject, narrative, fold) to obtain one score per voxel. **B)** the same alphas on the y-axis. On the x-axis, the corresponding brain scores for each (subject, narrative, voxel, fold) averaged across (subject, narrative, fold).

In Figure S12, we replicate our results with a partial correlation method, i.e. a method that separates two sets of features (Word Embedding and GPT-2) during the fitting of the linear model. Specifically, we fit both Word and GPT-2 models simultaneously with a banded ridge regression (Nunez-Elizalde et al., 2019), and then evaluate the unique variance accounted for by each sub-model. For simplicity, we follow the setup of the original paper (Nunez-Elizalde et al., 2019) and replicate our results for one pair of features (here, Word Embedding and GPT-2). We use the same modeling and cross-validation setting as in Figure 2.5F. Figure S12 shows the specific brains scores attributed to Word and GPT-2 embeddings, and their specific correlation with comprehension. We obtain similar results as in Figure 2.5F, but the correlation

with comprehension specific to GPT-2 ($R[M''(\text{GPT2})] = 0.52$) is slightly higher than the one in the paper ($R[M'(\text{GPT2}) - M(\text{Word})] = 0.31$).

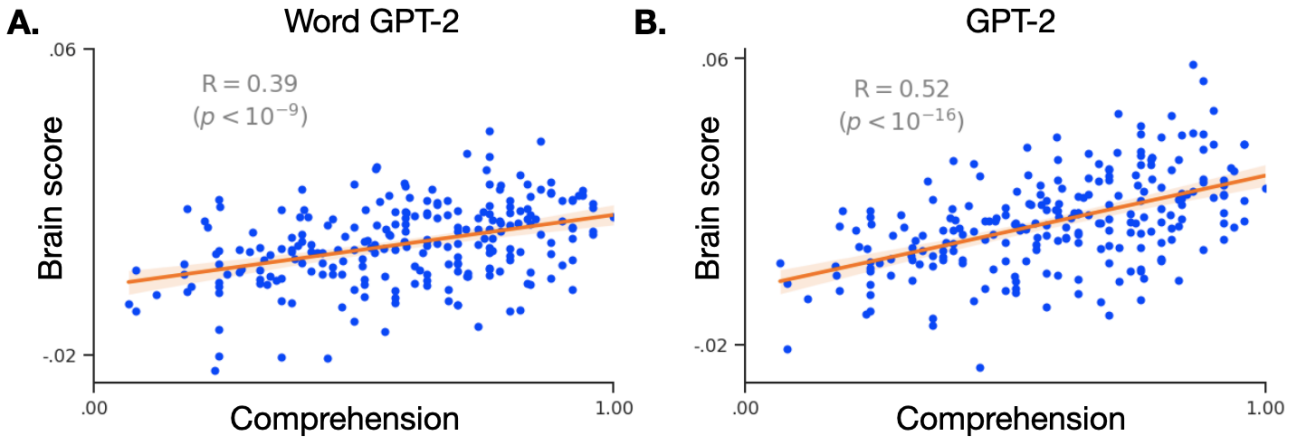


Figure S12: Same as Figure 2.5D but using partial correlation analysis: a model is fitted using both Word vectors and GPT-2 as input, we then evaluate the brain score accounted for by each submodel specifically. **A)** Brain scores of the Word vectors specifically, averaged across voxels. **B)** Brain scores of the eight layer of GPT-2 specifically, averaged across voxels. In red, the correlation between comprehension scores (x-axis) and brain scores (y-axis).

6.2.8 Effect of attention processes in the brain

Is the correlation between comprehension and GPT-2's representations solely due to attentional fluctuation? Indeed, attention can modulate both (i) comprehension and (ii) the average BOLD activity (Sabri et al., 2008; Kok et al., 2012) and thus lead to an indirect correlation between these last two variables. To address this issue, we first qualitatively compare our results to those of a meta-analysis covering 6,201 subjects recorded with fMRI during a study related to speech-based or auditory-based attention (Figure S13). The results suggest that these attentional mechanisms are associated with a restricted set of temporal and sensory-motor areas. Furthermore, our analysis of the average BOLD response and its correlation with comprehension highlight a similar cortical network (Figure S14). In both cases, however, these neural bases of attention appear much less distributed than those obtained with GPT-2. In particular, the activations in the prefrontal and parietal cortices as well as in the inferior temporal gyri seem to be specifically accounted for by GPT-2's representations. Overall, while these results call for more direct manipulations of subjects' attention, they suggest that the link between GPT-2 and the brain bases of comprehension is not trivially reducible to attention.

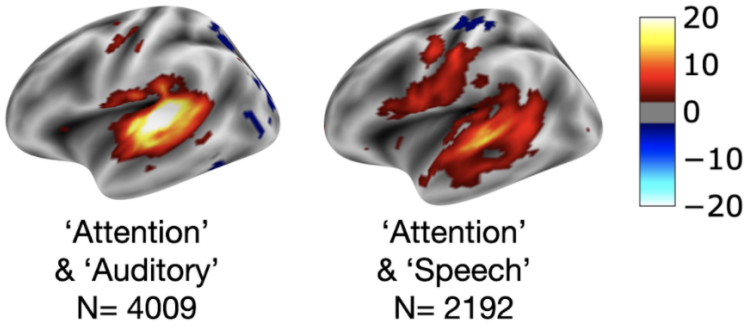


Figure S13: Meta-analyses from NeuroQuery. Brain networks associated with the concepts of"attention" combined with "auditory" or "speech".

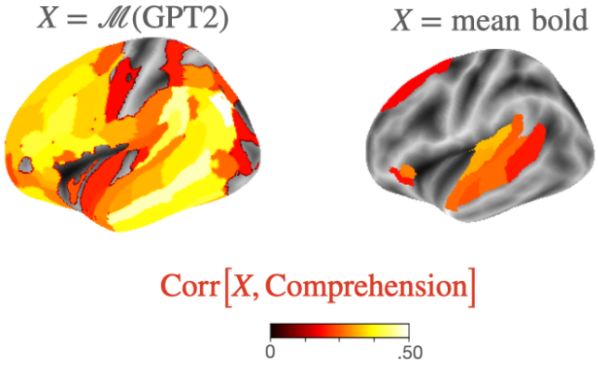


Figure S14: Correlation between comprehension scores and (a) brain scores of GPT-2 for each (subject, story) pair, (b) the BOLD magnitude, averaged across scans for each subject and story separately.

6.2.9 fMRI preprocessing

Our analyses rely on the already pre-processed data from Nastase et al. 2020 (Nastase et al., 2020), unsmoothed version. Below, the pre-processing pipeline, as stated in the original paper.

"The functional MRI data were preprocessed in the following way. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. A deformation field to correct for susceptibility distortions was estimated using fMRIPrep's fieldmap-less approach. The deformation field results from co-registering the BOLD reference to the same-subject T1w-reference with its intensity inverted (Huntenburg, 2014; Wang et al., 2017). Registration was performed with antsRegistration (ANTs 2.2.0), and the process was regularized by constraining deformation to be nonzero only along the phase-encoding direction, and modulated with an average fieldmap template (Treiber et al., 2016). Based on the estimated susceptibility distortion, a corrected EPI reference was calculated for more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer 6.0.1), which implements boundary-based registration (Greve

and Fischl, 2009). Co-registration was configured with six degrees of freedom. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9; Jenkinson et al., 2002, 2012; Smith et al., 2004). BOLD runs were slice-time corrected using 3dTshift from AFNI (20160207; Cox and Hyde, 1997). The BOLD time-series were resampled onto the following surfaces: fsaverage , fsaverage6 , fsaverage5 . The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series are referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD time-series were resampled into two volumetric standard spaces, correspondingly generating the following spatially-normalized, preprocessed BOLD runs: MNI152NLin2009cAsym , MNI152NLin6Asym . A reference volume and its skull-stripped version were first generated using a custom methodology of fMRIPrep. All resamplings were performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs 2.2.0), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer 6.0.1).

Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS, and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor; Behzadi et al., 2007). Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). The tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. All of these confound variables are provided with the dataset for researchers to use as they see fit. HTML files with quality control visualizations output by fMRIPrep are available via DataLad.

We next temporally filtered the functional data to mitigate the effects of confounding variables. Unlike traditional task fMRI experiments with a well-defined event structure, the goal of regression was not to estimate regression coefficients for any given experimental conditions; rather, similar to resting-state functional connectivity analysis, the goal of regression was to model nuisance variables,

resulting in a "clean" residual time series. However, unlike conventional resting-state paradigms, naturalistic stimuli enable intersubject analyses, which are less sensitive to idiosyncratic noises than within-subject functional connectivity analysis typically used with resting-state data (Simony et al., 2016; Simony and Chang, 2019). With this in mind, we used a modest confound regression model informed by the rich literature on confound regression for resting-state functional connectivity (e.g. Ciric et al., 2017; Parkes et al., 2018). AFNI's 3dTproject was used to regress out the following nuisance variables: six head motion parameters (three translation, three rotation), the first five principal component time series from an eroded CSF and a white matter mask (Behzadi et al., 2007; Muschelli et al., 2014), cosine bases for high-pass filtering (using a discrete cosine filter with cutoff: 128 s, or .0078 Hz), and first- and second-order detrending polynomials. These variables were included in a single regression model to avoid reintroducing artifacts by sequential filtering (Lindquist et al., 2019). The scripts used to perform this regression and the residual time series are provided with this data release. This processing workflow ultimately yields smoothed and non-smoothed versions of the "clean" functional time series data in several volumetric and surface-based standard spaces."

6.3 Disentangling syntax and semantics in the brain with deep networks

6.3.1 Deep Neural Networks' Activations

Pre-trained tansformers In Section 3.1.5, we extract the activations of GPT-2 (Radford et al., 2019) and five transformer architectures: BERT (Devlin et al., 2019), XLnet (Yang et al., 2020), Roberta (Liu et al., 2019), AlBert (Lan et al., 2020) and DistilGPT-2. We use the pre-trained models from Huggingface (Wolf et al., 2020): 'bert-base-cased', 'xlnet-base-cased', 'roberta-base', 'albert-base-v1', and 'distilGPT-2' respectively. In Figure 3.7, we focus on one middle layer of these transformers ($l = n_{\text{layers}} \times 2/3$), because it has shown to best encode brain activity (Caucheteux & King, 2022) and to encode relevant linguistic properties (Manning et al., 2020; Jawahar et al., 2019).

Text formatting and tokenization To extract the activations elicited by one story, we proceed as follows: we first format and lower case the text (replacing special punctuation marks such as "–" and duplicated marks "?." by dots), then apply the tokenizer provided by Huggingface (Wolf et al., 2020) to convert the transcript into either word-level or sub-word-level tokens called "Byte Pair Encoding" (BPE) (Sennrich et al., 2016). Here, more than 99.5% of BPE-level tokens were complete words. The tokens are then split into sections of 256 tokens (this length is constrained by GPT-2's architecture) and input to the deep network one story at a time. The activations of each layer are finally extracted, resulting in n_{layers} vectors of 768 activations for each token of each story transcript. In the 0.5% case where BPE are not complete words, BPE-features are summed between successive words, to obtain n_{layers} vectors per word per story.

6.3.2 Convergence of the Method to Build $\overline{X}$

In Section 3.1.4 and 3.1.4, we compute the syntactic component $\overline{X}$ of GPT-2 activations X elicited by a sentence w. $\overline{X}$ is approximated by $\overline{X_k}$, the average activations across k sentences with the same syntax as

w. Here, we sample $k = 10$ sentences. We check in Figure S15 that the method has converged before $k = 10$. We compute the cosine similarity between $\overline{X}_k$ and $\overline{X}_{k-1}$ for k between 1 and 15. The syntactic embeddings stabilize with at least eight sampled sentences.

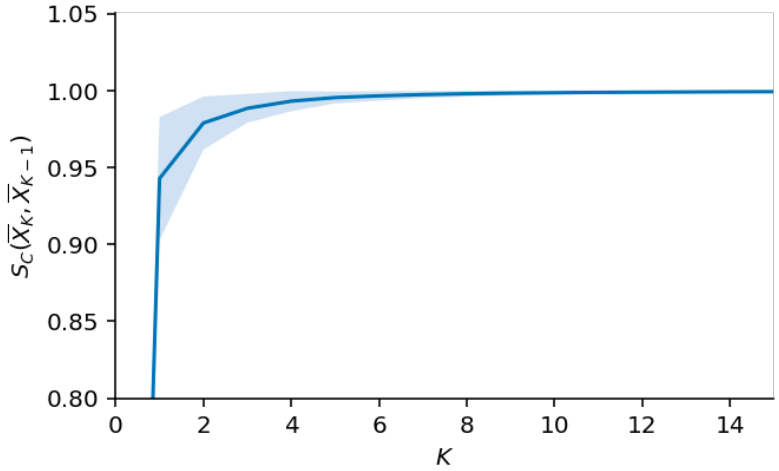


Figure S15: Convergence of the method to build syntactic embeddings. Cosine similarity S_C between the syntactic component $\overline{X}$ of GPT-2 activations induced by a sequence w, when computed with K and $K-1$ syntactically equivalent sequences. The syntactic embeddings $\overline{X}_K$ and $\overline{X}_{K-1}$ are computed for 100 Wikipedia sentences ($\approx 2,800$ words), and the similarity scores are averaged across embeddings. In shaded, the 95% confidence interval across embeddings.

6.3.3 Evaluating the Level of Semantic and Syntactic Information in $\overline{X}$

In Section 3.1.4 and Figure 3.3, we check that the syntactic embedding $\overline{X}$ extracted from GPT-2 only contains syntax. To this aim, we evaluate the ability of a linear decoder to predict two syntactic features and three semantic features from $\overline{X}$.

Semantic and syntactic features The two syntactic features derived from the stimulus are:

- The part-of-speech of the words (categorical feature), as defined by Spacy tags (Honnibal et al., 2020).
- The depth of the syntactic tree (continuous feature). The syntactic tree is extracted with the state-of-the-art Supar dependency parser (Y. Zhang et al., 2020).

The three semantic features are only computed for verbs, nouns and adjectives (as defined by Spacy part-of-speech tags) and are the followings:

- Word frequency (labeled as 'Word freq' in Figure 3.3, continuous feature). We use the 'zipf_frequency' from the wordfreq[1] python library.

[1]https://pypi.org/project/wordfreq/

- Word embedding (continuous feature), computed using the pre-trained model from Spacy (Honnibal et al., 2020) ('en_core_web_lg', 300 dimensions).
- Semantic category (categorical feature). We used the 47 semantic categories[2]. Categories are not available for all the 2,800 Wikipedia words studied here. Thus, we first train a linear model (scikit-learn 'RidgeCVClassifier') to predict the semantic category of the 535 labeled words used in (Binder et al., 2016), given their Spacy word embedding (300 dimensions). We then label the 2,800 Wikipedia words using the semantic category predicted by the classifier.

Linear decoder To evaluate the ability of a linear decoder to predict the five linguistic features from $\overline{X}$, we:

- Build syntactic embeddings $\overline{X}$ for 100 Wikipedia sentences ($\approx 2,800$ words), following Section 3.1.4, using the ninth layer of GPT-2.
- Build the three semantic and two syntactic features described above from the 2,800 Wikipedia words Wikipedia words.
- Fit a ℓ_2-regularized linear model to predict the five features given the syntactic embeddings. We use the 'RidgeCV' regressor (resp. 'RidgeClassifierCV' classifier) from scikit-learn (Pedregosa et al., 2011) to predict the continuous (resp. categorical) features, with ten possible penalization values log-spaced between 10^{-3} and 10^{6}.
- Evaluate the linear model on held out data, using a 10 cross-validation setting ('KFold' cross-validation from scikit-learn). Performance is assessed using *adjusted* accuracy ('balanced_accuracy_score' from scikit-learn) for the categorical features, and R^2 for the continuous features. Thus, the chance level is zero for both types of features, and the best score is one.
- Report the average decoding performance in Figure 3.3 (red bars), and the standard-error of the means across the ten test folds.

For comparison, we repeat the exact same procedure with the full GPT-2 activations X (instead of their syntactic component $\overline{X}$), and report the results in Figure 3.3 (grey bars).

6.3.4 Temporal Alignment g between X and Y

In Section 3.1.4, we map the network's activations X (of length M, the number of words) and the brain response Y (of length N, the number of fMRI measurements) induced by the same story w (of M words). M is usually greater than N. To align the two spaces, we first sum the features between successive fMRI measurements, and then apply a finite impulse response (FIR) model. We denote g this transformation.

[2]Categories are: abstract, action, animal, auditory, body, building, cognitive, construct, creative, device, distant, document, electronic, emotion, emotional, entity, event, food, furniture, general, geological, group, human, instrument, locative, mental, miscellaneous, multimodal, object, part, perceptual, period, physical, place, plant, property, social, somatosensory, sound, spatial, state, temporal, time, tool, vehicle, visual, weather

Specifically, for each fMRI time sample $i \in [1 \dots N]$, g_i combines word features within each acquisition interval as follows:

$$\begin{aligned} g_i : \mathbb{R}^{M \times d} &\to \mathbb{R}^{5d} \\ u &\mapsto \left[\widetilde{u_i}, \widetilde{u_{i-1}}, \dots, \widetilde{u_{i-4}}\right] \\ \widetilde{u_i} &= \sum_{\substack{m \in [1 \dots M] \\ \mathcal{T}(m) = i}} u_m \end{aligned}$$

with

$$\begin{aligned} \mathcal{T} : [1 \dots M] &\to [1 \dots N] \\ m &\mapsto i \quad / \quad |t_{y_i} - t_{x_j}| = \min_{k \in [1 \dots N]} |t_{y_k} - t_{x_m}| \end{aligned}$$

with $\tilde{u}$ the summed activations of words between successive fMRI time samples, u the five lags of FIR features, $(t_{x_1}, \dots, t_{x_M})$ the timings of the M words onsets, and $(t_{y_1}, \dots, t_{y_N})$ the timings of the N fMRI measurements.

6.3.5 Brain Parcellation

In Figure 3.6, brain scores are averaged across voxels within regions of interest using the Brodmann's areas from the PALS parcellation of freesurfer[3]. To gain in precision, we split the superior temporal gyrus (BA22) into its anterior, middle and posterior parts. In Figure 3.6, we report the top ten areas of the left hemisphere in term of average brain score. Certain areas are renamed for clarity, as specified in the table below:

Label	Corresponding Brodmann's areas
A1	BA41 / BA42
Fusiform	BA37
Angular	BA39
aSTG	BA22-anterior
mSTG	BA22-middle
pSTG	BA22-posterior
M1	BA4
Supramarginal	BA40
IFG (Op)	BA44
IFG (Tri)	BA45
IFG (Orb)	BA47
Middle-frontal	BA46
V1	BA17
Fronto-polar	BA10
Temporo-polar	BA38
Precuneus	BA7
Cingulate	BA23 / BA26 / BA29 / BA30 / BA31

[3]https://surfer.nmr.mgh.harvard.edu/fswiki/PALS_B12

6.3.6 Control for Low-level Linguistic Features

In Section 3.1.6 and Figure 3.7, we check that the brain scores are not driven by low-level linguistic features. Thus, we compute the R scores of GPT-2 activations (ninth layer) induced by modified versions of the stimulus:

- Random words sampled from the same story. Words are uniformly sampled from the words of the story, tokenized using Spacy (Honnibal et al., 2020). Punctuation marks are considered as words. Upper-cases are kept.
- Random sentences from Wikipedia, of the same length as the sentences of the stimulus. We first build a dictionary of (length, list of match-length sentences) pairs out of 10K sentences from Wikipedia ($\approx$ 577K words). Then, for each sentence of the stimulus, a sentence is uniformly sampled from the set of Wikipedia match-length sentences.
- The sentences of the stimulus, but with random word order. Words are shuffled *within* each sentence.

Then, we extract the corresponding GPT-2 activations and compute the R scores following Section 3.1.5. R scores are evaluated for each subject and reported in Figure 3.7.

6.4 Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects

To replicate Lerner et al.'s findings, we compute the model-to-brain correlation (*cf.* Section 3.2.3):

$$R = \rho\big(y, f_\theta(x^*)\big) \ ,$$

for the acoustic, word, sentence and paragraph level respectively. Here, we provide additional details on how to extract the brain signals y and estimate the mapping function f_θ in order to reproduce the experimental setting used in Section 3.2.4.

6.4.1 Brain signals

Functional MRI dataset We use the fMRI recordings of the Narratives dataset (Nastase et al., 2020)[4], a publicly available dataset gathering the brain recordings of 305 subjects listening to narratives. We use the unsmoothed version of the fMRI recordings, already preprocessed in the original dataset. As suggested in the original paper, we reject subject / narrative pairs because of noisy recordings, resulting in 617 unique (story, subject) pairs and 4.1 hours of audio stimulus in total. To replicate the results of Lerner et al. (2011), we restrict the analyses to the 75 subjects listening to the 'Pieman' story (7 min long), including the seven subjects analysed in the original paper (only the data for non-scrambled stimuli are publicly available). Then, we extend the analyses to the brain recordings of 305 subjects listening to

[4] http://datasets.datalad.org/?dir=/labs/hasson/narratives

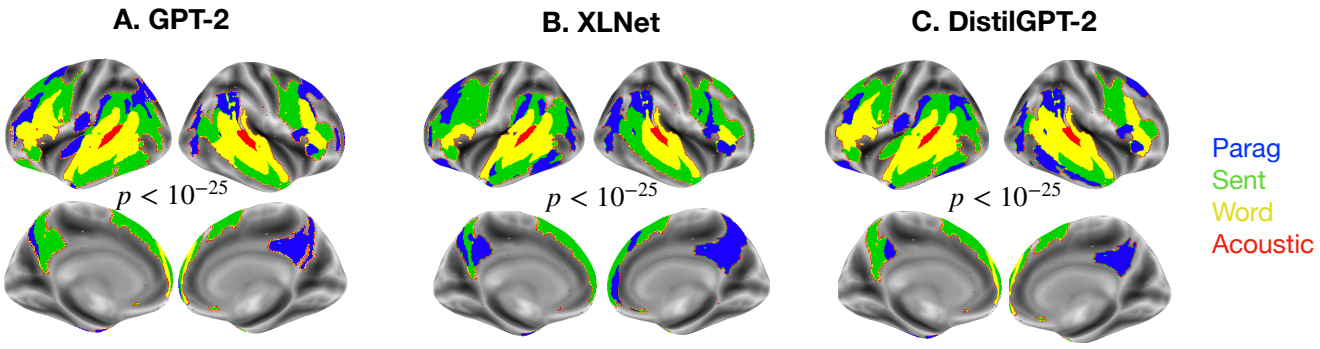


Figure S16: Replication to two other architectures. Same as Figure 3.9.C but using the intermediate layers of XLNet and Distilgpt2 causal architectures ($l = 4$ for Distilgpt2, out of 6 layers in total and $l = 8$ for XLNet, out of 12 layers in total). As in Figure 3.9.C, the significance threshold is set to $p < 10^{-25}$.

fifteen narratives (from 3 min to 57 min), from the same dataset (Nastase et al., 2020). For both analyses, we only have access and thus use the brain recordings elicited by regular –i.e non scrambled– version of the stimuli.

6.4.2 Encoding features

Deep language models' activations In Section 3.2.4, we extract the activations of GPT-2 ($\mathcal{A}$), a deep neural language model trained to predict a word given its past context. It consists of 12 transformer layers of dimensionality 768, 8 heads, and has 1.5 billion parameters in total. We use the model provided by Huggingface (Wolf et al., 2020), trained on a dataset of 8 million web pages.

To extract the activations elicited by a sequence w of M words from a layer l, we proceed as follows: we tokenize the sequence into sub-words called "Byte Pair Encoding" (BPE) (Sennrich et al., 2016) using the GPT-2 tokenizer provided by Huggingface. Then, we feed the network with the M' BPE tokens ($M' \geq M$, up to 256 tokens in memory) and extract the corresponding activations from layer l, of shape $(M' \times D)$ with $D = 758$. Then, we sum the activations over the BPEs of each word to obtain a vector of size $(M \times D)$.

All our analyses are based on the eighth layer of GPT-2. We choose GPT-2 because it has been shown to best encode the brain activity elicited by language stimuli (Schrimpf et al., 2021). We choose its eighth layer because the intermediate layers of transformers have shown to encode relevant linguistic features (Jawahar et al., 2019; Manning et al., 2020) and to better encode brain activity than input and output layers (Caucheteux & King, 2022; Toneva & Wehbe, 2019).

Scrambling the stimulus at the word, sentence and paragraph level Words and sentences of the stimulus are delimited using Spacy tokenizer (Honnibal et al., 2020). Note that punctuation marks are not considered as words (*e.g.,* 'time.' forms *one* token, not two). We define paragraphs as contiguous chunks of eight sentences. To 'scramble' a sequence at the word (resp. sentence, paragraph) level, we uniformly shuffle the indices of its words (resp. sentences, paragraphs) and form the new sequence accordingly.

Computation of x^* for the word, sentence and paragraph conditions In Section 3.2.3, we compute a context-free representation x^* for the word, sentence and paragraph condition. In short, x^* are the activations of GPT-2, averaged over several scrambled contexts. For clarity, we focus on the sentence level to detail the approach. To build the sentence-level representation x^* of the stimulus, we use the approximation introduced in equation (3.7). For each sentence s of one story w, we i) generate K=10 sequences ending with s, but with scrambled previous context. The scrambled context is uniformly sampled from the other sentences in the same story w. Then, ii) we extract the K corresponding activations from GPT-2 (as described in the previous section) and iii) average the activations across the K samples. GPT-2 activations are extracted for each word. Thus, for each of the M_s words of sentence s, we obtain a vector x_s^* of shape $M_s \times D$. We concatenate these vectors to obtain x^*, a sentence-level representation of the whole story w, of shape $M \times D$. This method is adapted from (Caucheteux et al., 2021a), in which the authors compute the average over GPT-2 activations to extract syntactic representations from the input sequence.

Acoustic features GPT-2 takes words as input and not sounds. To build x^* at the acoustic level, we simply use non-contextual acoustic features: the word rate ($D = 1$), phoneme rate ($D = 1$) phonemes, stress, and tone (categorical, $D = 117$). For the latter, we use the annotations provided the original Narratives dataset (Nastase et al., 2020).

6.4.3 Mapping x^* onto the brain

The linear function f_θ maps x^* onto y, the fMRI recordings of one subject at one voxel. Vector y is of length T, the number of fMRI time samples, whereas x^* is of length M, the number of words (or phonemes for acoustic features) in the story. To align the two time domains, we apply the function $g : \mathbb{R}^{M \times D} \mapsto \mathbb{R}^{T \times 5D}$ that i) sums the features x^* between the successive fMRI time samples, and ii) uses a Finite-Impulse Response model (FIR) with five delays. Thus, $f_\theta = f'_\theta \circ g$, with f_θ a linear function whose parameters θ are learned, and g a temporal alignment function.

To estimate θ, we fit an ℓ_2-penalized linear regression to predict y given $g(x^*)$ on a training set of time samples. θ thus minimizes

$$\underset{\theta' \in \Theta}{\operatorname{argmin}} \|y_{\text{train}} - f_{\theta'} \circ g(x^*_{\text{train}})\|^2 + \lambda \|\theta'\|^2 \ ,$$

with λ the regularization parameter. We assess the mapping with a Pearson correlation score evaluated on the left out times samples:

$$R = \rho\Big(y_{\text{test}}, f_\theta \circ g(x^*_{\text{test}})\Big) \ .$$

In practice, x^* and $g(x^*)$ are standardized (0-mean, 1-std) and brain signals y are scaled based on quantiles using scikit-learn RobustScaler (Pedregosa et al., 2011) with quantile range $(.01, .99)$. We use the RidgeCV implementation of scikit-learn with a pool of twenty possible penalization parameters between 10^{-3} and 10^6. We learn f_θ on 90% of the T time samples, and compute the correlation scores R on the 10% left out data. We repeat the procedure on 10 test folds using a cross-validation setting, following the KFold implementation of scikit-learn without shuffling. Finally, we average the R over the 10 folds to obtain one model-to-brain correlation score per subject, voxel and feature space x^*.

6.4.4 Brain parcellation

In Figure 3.9, we use a subdivision of Destrieux' atlas (Destrieux et al., 2010). Regions of more than 200 vertices are split into smaller regions, so that each region contains at most 200 vertices. Thus, from the 75 regions of Destrieux' atlas (in each hemisphere), we obtain a parcellation of 465 brain regions per hemisphere.

6.4.5 Significance

In Figure 3.9, we test whether the model-to-brain correlations (R) are significantly different from zero. To this aim, we use a two-sided Wilcoxon test across subjects ($N = 75$ in Figure 3.9B, $N = 305$ in Figure 3.9A), corrected using False Discovery Rate (FDR) across the 465 region of interests in each hemisphere.

6.4.6 Generalization to other transformer architectures

In Figure S16 (B and C), we replicate our results (Figure 3.9.C) on the activations of two other causal transformer architectures: XLNet (Yang et al., 2020) and Distilgpt2 (Figure S16.C), using the implementation from Huggingface[5].

[5]https://huggingface.co/

6.5 Toward a realistic model of speech processing in the brain with self-supervised learning

6.5.1 Self-supervised loss formula

Wav2vec 2.0, when trained in a self-supervised way, uses a loss (L) which is the weighted combination of two losses: one diversity loss (L_d), which pushes the quantization module to contain representations that are as diverse as possible, and one Contrastive Predictive Coding loss (L_m), which pushes the model to choose, from the context network output c, the right quantized representation (q) of some masked input, among other possible representations. L_m has the following formula, for some masked time step t:

$$\mathcal{L}_m = -\log \frac{\exp\left(\text{sim}\left(\mathbf{c}_t, \mathbf{q}_t\right)/\kappa\right)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp\left(\text{sim}\left(\mathbf{c}_t, \tilde{\mathbf{q}}\right)/\kappa\right)} \tag{6.1}$$

with $\text{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathbf{b}/\|\mathbf{a}\|\|\mathbf{b}\|$, κ the temperature, which is constant during training, Q_t the set of $K+1$ quantized candidate the model has to choose from, including the right one, i.e. q_t.

L_d is included to encourage the equal use of the V possible entries of each of the G codebooks of the quantization module. The goal is to maximize the entropy of the averaged softmax distribution over the codebook entries for each codebook $\bar{p}_g$, across a set of utterances:

$$\mathcal{L}_d = \frac{1}{GV}\sum_{g=1}^{G} -H\left(\bar{p}_g\right) = \frac{1}{GV}\sum_{g=1}^{G}\sum_{v=1}^{V} \bar{p}_{g,v}\log \bar{p}_{g,v} \tag{6.2}$$

6.5.2 Supervised loss formula

When trained in a supervised way, wav2vec 2.0 is trained to optimise a Connectionist Temporal Classification loss parameterized over θ:

$$\text{argmin}_\theta \ -\log \sum_{a \in a_{U,V}} \prod_{t=1}^{d_t} p_{\text{CTC}}\left(a_t \mid m_\theta(U)\right) \ , \tag{6.3}$$

where $m_\theta(U) \in \mathbb{R}^{d_\tau \times d_v}$ are the probabilistic predictions of the model at each τ time sample given the input raw waveform $U \in \mathbb{R}^{d_\tau \times d_u}$, $V \in \mathbb{R}^{d_t \times d_v}$ are the true transcriptions of U, and $a_{U,V}$ is the set of all possible alignments between U and V.

6.5.3 Preprocessing of the model's activations

The activations of the network $X \in \mathbb{R}^{d_{\hat{t}} \times d_x}$ are first normalized to be between $[0, 1]$ for each listening session. Then, we use nistats (Abraham et al., 2014) `compute_regressor` function with the 'glover' model to temporally convolve ($h \in \mathbb{R}^{d_{\hat{t}}}$) and temporally down-sample (using $g : \mathbb{R}^{d_{\hat{t}}} \to \mathbb{R}^{d_t}$) each artificial neuron j:

$$\hat{x}^{(j)} = g\Big(x^{(j)} * h\Big) \ . \tag{6.4}$$

6.5.4 Penalized linear model - Ridge regression

For each split s, we fit an ℓ_2-penalized linear model $V \in \mathbb{R}^{d_x \times d_z}$ trained to predict the transformed BOLD time series from the model activations for each dimension independently. The formula of the optimization is the following:

$$\text{argmin}_V \sum_{i \in \text{train}_s} (V^\top \hat{X}_i - y_i)^2 + \lambda \|V\|^2 \ . \tag{6.5}$$

6.5.5 Probing the linguistic features encoded in wav2vec2 activations

Interpreting the representations of deep learning models is notoriously difficult. To address this issue, (Pasad et al., 2021) explored the encoding of local acoustic features, phone identity, word identity and word meaning across layers. Similarly, (Millet et al., 2021) compared representations to human behavioural data to assess whether they better captured listeners' perception of higher-level phonemic properties or of lower-level subphonemic properties of speech stimuli. Finally, (Vaidya et al., 2022) recent study explores filter banks, spectrograms, phonemes and words across layers. Here, we complement these analyses by showing that self-supervised learning allows wav2vec 2.0 to learn represents, along its hierarchy the representations of MEL spectrograms, phonetic categories and word embeddings (Figure S17).

For this, we perform a ridge regression on the Timit dataset[6] to predict five auditory and linguistic features from the activation functions of each layer and model of the present paper. We study the following features:

- the MEL spectrogram of the audio, computed using librosa (d=128)
- the phonemes (categorical features). We use the transcripts and alignments provided in Timit.
- the word embedding and part-of-speech of the words. The time alignments for words are provided by Timit. We use spaCy to compute the word embedding (medium model, d=300), and their part-of-speech (categorical feature, d=19).
- the sentence embedding of each sample, provided by Laser.

We use a subset of 1,680 samples from Timit, each sample being an audio recording of a short sentence (¡10 seconds) from 24 speakers. The model's activations were mean-pooled to the sampling rate of each feature.

The results show that the layers of wav2vec 2.0 partially follow the hierarchy predicted from neurolinguistics (Hickok & Poeppel, 2007) (Table S6): the first layers of the transformer best account for the spectro-temporal information, whereas deeper layers best account for the phonetic, word-level and sentence level information. While all of these features emerge with training (Figure S17), only the highest-level features (phone, word and sentence-level) appear to be specific to speech and to the language with which wav2vec 2.0 was trained (Figure S17).

Interestingly, the word and sentence-level features are encoded deeper in the supervised network (best layer=18 in Table S6) compared to the unsupervised network (best layer=14), which suggests that self-supervised learning generates a reservoir representations in its middle layers, reservoir which may

[6] https://catalog.ldc.upenn.edu/LDC93S1

partly overlap with the labels used in supervised learning. Together with our ABX tests, and layer-wise tuning of each voxel (Figure 3.12), these elements suggest that the representations of speech shaped by our experience are learnt and instantiated in the superior temporal gyrus and sulcus. These elements, consistent with previous electrophysiological studies (Mesgarani et al., 2014), thus provide a coherent spectrum of evidence for the location of acquired speech representations in the brain.

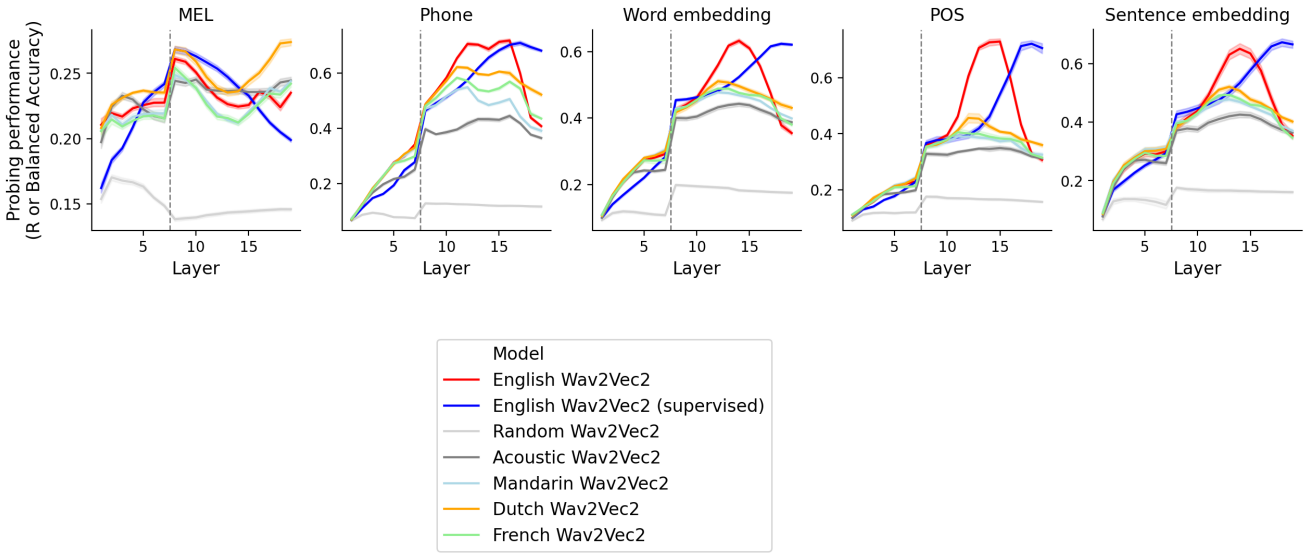


Figure S17: Linguistic features encoded in each layer of the networks. For each layer of each network, we train a l2-penalized linear model from scikit-learn (Pedregosa et al., 2011) to predict several linguistic categories given the embedding. The tested categories are the following: MEL (the MEL spectrogram of the audio, d=128), phone (the phoneme, categorical, d=39), the word embedding of the word (computed with spaCy (`https://spacy.io`) English model, d=300), the Part-Of-Speech (POS) of the word provided by spaCy (categorical feature, n=19), and the embedding of the sentence, computed using Laser (`https://github.com/facebookresearch/LASER`) (d=1,024). We train and test the linear probe on a subset of Timit data (`https://catalog.ldc.upenn.edu/LDC93s1`), using a 10-folds cross-validation scheme, and report the probing accuracy (either R for continuous variables or balanced accuracy for categorical variables) for each possible target feature. We average the corresponding probing performances across the 10 folds. Error bars are standard errors of the mean across folds.

6.5.6 Noise ceiling analysis

The noise in fMRI recordings is inevitable. To estimate the maximum explainable signal given this level of noise, we follow previous studies and employ a shared-response model, or "noise ceiling" (Huth, de Heer, et al., 2016; Caucheteux & King, 2022; Caucheteux et al., 2022). Precisely, we predict the brain signals of one subject given the brain activity of the other subjects, in response to the same audio recording. In practice, we apply the same evaluation as Equation (3.3.3), for one subject s and one voxel v, but we use the average brain signals of other subjects' brains $\overline{Y}^{(s)} = \frac{1}{|\mathcal{S}|}\sum_{s' \neq s} Y^{(s')}$ instead of the

activations X. As a result, the "noise ceiling" of one subject (s) and one voxel (v) is computed as follows:

$$R_{\text{noiseceil}} = \text{Corr}(W \cdot \overline{Y}^{(s)}, Y^{(s,v)}) \quad , \tag{6.6}$$

where W is an ℓ_2-penalized linear regression fitted on separate train data, using a cross validation setting with five test folds.

We compute such noise ceiling on 290 subjects of the Narrative dataset listening to the same stories (Figure S18). We report the noise ceiling across voxels in Figure S18, and, in Table S3, the brain scores of the networks studied in the main paper normalised by the noise ceiling. Precisely, for each voxel, we divide the average brain scores by the noise ceiling for this particular voxel. While low on average, the unsupervised wav2vec2 model reaches 74% of the noise ceiling in Heschel, and more than 20% in STS, STS and IFG.

	Average	Top10	Heschl	STG	STS	IFG	Motor
Random wav2vec2	13.9%	29.0%	66.9%	32.0%	21.8%	15.9%	11.9%
Non-Speech	16.4%	33.9%	71.0%	36.8%	26.9%	19.0%	11.7%
Non-Native	17.6%	35.9%	73.0%	39.0%	29.1%	21.0%	12.9%
Native, Supervised	18.3%	36.7%	74.2%	39.6%	29.8%	21.2%	13.6%
Native, Unsupervised	18.8%	37.9%	74.4%	40.3%	31.3%	22.8%	13.8%
Noise ceiling	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Table S3: Brain scores *with* noise ceiling normalisation. Brain scores divided by the noise ceiling, for the Narrative dataset, on average across all voxels ('Average'), for the 10% best voxels of the noise ceiling ('Top10', Figure 6.5.6) and the voxels of five regions of interests.

	Average	Top10	Heschl	STG	STS	IFG	Motor
Random wav2vec2	0.019	0.069	0.192	0.071	0.044	0.024	0.011
Non-Speech	0.022	0.080	0.205	0.081	0.055	0.028	0.011
Non-Native	0.024	0.085	0.211	0.086	0.059	0.031	0.012
Native, Supervised	0.025	0.086	0.213	0.087	0.060	0.032	0.013
Native, Unsupervised	0.025	0.089	0.214	0.089	0.063	0.034	0.013
Noise ceiling	0.117	0.219	0.287	0.181	0.196	0.149	0.094

Table S4: Brain scores *without* noise ceiling normalisation Same as Table S3, but without dividing by the noise ceiling estimates.

Below, we report the brain scores of our models, normalised by such noise ceiling. Precisely, we compute the brain scores for each subject and voxels

	Avg	Top10NoiseCeil	Heschl	STG	STS	IFG	Motor
Unsupervised	0.03 +/- 0.001	0.09 +/- 0.002	0.21 +/- 0.007	0.09 +/- 0.003	0.06 +/- 0.002	0.03 +/- 0.001	0.01 +/- 0.001
Supervised	0.02 +/- 0.001	0.09 +/- 0.002	0.21 +/- 0.007	0.09 +/- 0.003	0.06 +/- 0.002	0.03 +/- 0.001	0.01 +/- 0.001
Noise ceiling	0.12 +/- 0.006	0.22 +/- 0.006	0.29 +/- 0.008	0.18 +/- 0.006	0.20 +/- 0.006	0.15 +/- 0.006	0.09 +/- 0.006
Ratio	0.19 +/- 0.006	0.38 +/- 0.010	0.74 +/- 0.025	0.40 +/- 0.013	0.31 +/- 0.011	0.23 +/- 0.010	0.14 +/- 0.014

Table S5: Brain scores and noise ceiling estimates. Ratio indicate the unsupervised model divided by the noise ceiling. Scores are averaged across subjects and either all the voxels ('Avg') or the voxels of the selected regions of interests.

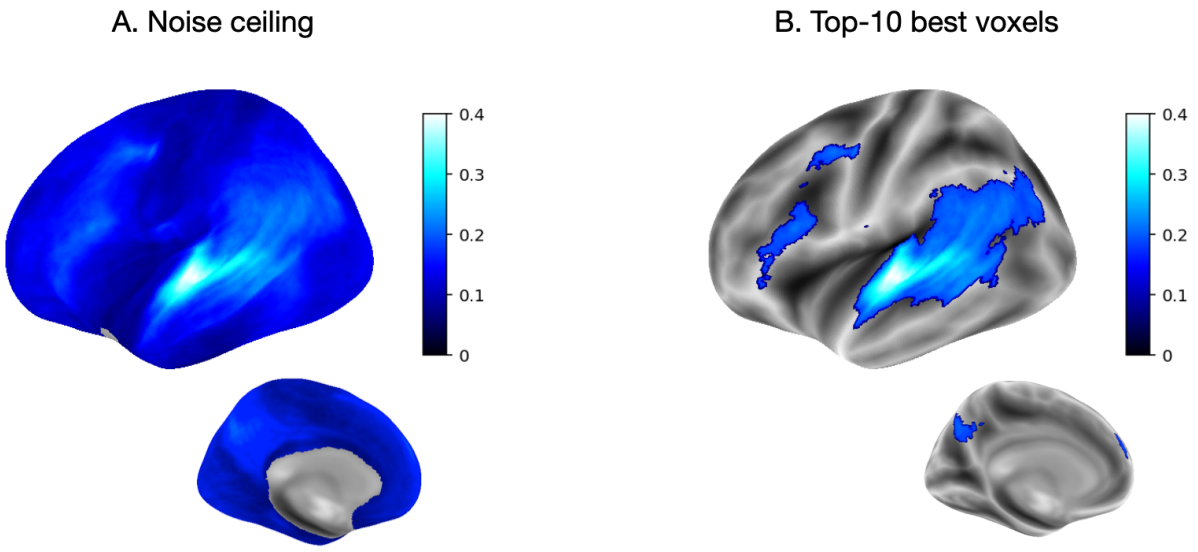


Figure S18: Noise ceiling. A. Noise ceiling estimates computed on 290 subjects of the Narratives dataset, averaged across subject. We only display the significant voxels across subjects ($p < 10^{-18}$). **B.** Same as A, but we only display the 10% voxels with the best noise ceiling estimates on average across subjects.

	MEL	Phone	Wordemb	POS	Sentemb	Average
Random wav2vec2	2.0	8.7	8.0	8.9	8.1	7.1
Acoustic wav2vec2	12.5	15.7	14.0	14.4	14.2	14.2
Mandarin wav2vec2	9.1	11.9	12.2	11.9	13.0	11.6
French wav2vec2	8.0	11.0	12.7	11.8	13.0	11.3
Dutch wav2vec2	18.9	11.4	12.0	12.4	13.0	13.5
English wav2vec2	8.0	15.2	14.0	14.4	14.0	13.1
English wav2vec2 (supervised)	8.0	16.9	18.0	18.0	18.0	15.8
Avg	9.5	13.0	13.0	13.1	13.3	12.4

Table S6: For each model (row) and target (column), the layer that maximizes probing performance, averaged across the 10 cross-validation folds.

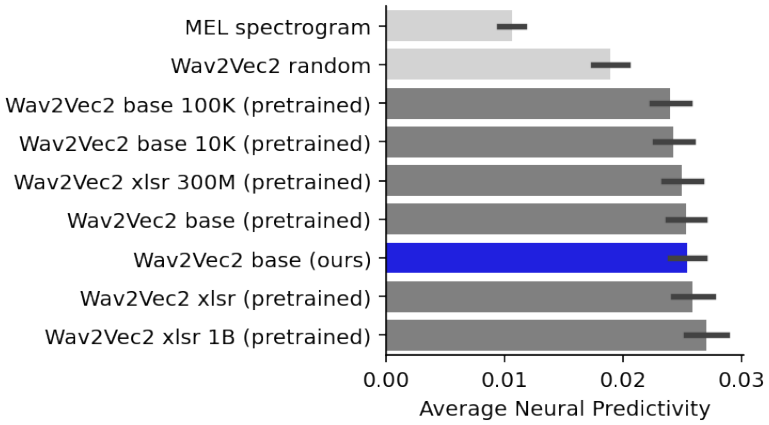


Figure S19: Brain scores of self-supervised pre-trained models. Brain scores, averaged across all voxels and subjects, for the MEL spectrogram, a wav2vec2 (base) architecture with random weights, wav2vec 2.0 (base) pre-trained with self supervised learning on 100K hours from Voxpopuli (Wang, 2021) ('wav2vec2-base-100k-voxpopuli' from huggingface), on 10K hours from Voxpopuli ('wav2vec2-base-10k-voxpopuli'), on 53K hours of english ('wav2vec2-base'), two models pre-trained on the same multilingual corpus of 436K hours, with 300M ('wav2vec2-xls-r-300m') and 1B parameters ('wav2vec2-xls-r-1b'), respectively, and our model trained on 600 hours of english speech (in blue). +/- refers to standard errors of the mean across subjects.

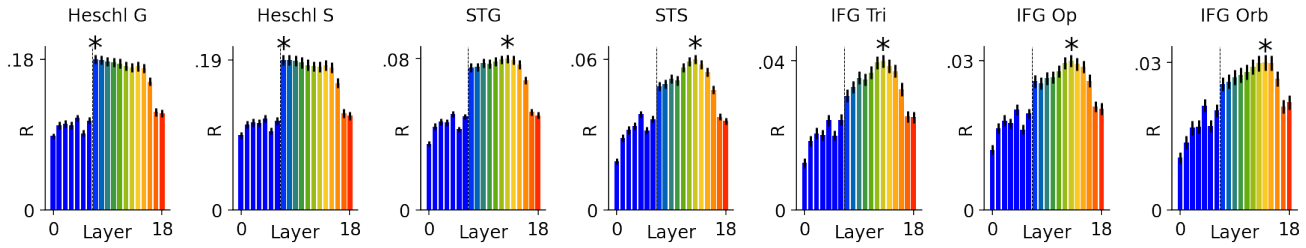


Figure S20: Brain scores for each layer of wav2vec 2.0 Same as figure 3B, but for different regions of the brain. Brain scores are averaged across all voxels in each regions.

6.6 Evidence of a predictive coding hierarchy in the human brain listening to speech

6.6.1 Scores per region of interest

For clarity, we report in Supplementary Figure S21 the average brain scores, forecast scores and forecast distances for each region of interest in both the left and right hemispheres. We also report scores in the less noisy voxels, for the subjects with the highest brainscore (Supplementary Table S7), their corresponding p-values computed across subjects (Supplementary Table S9) and the scores normalized by the noise ceiling (Supplementary Table S8).

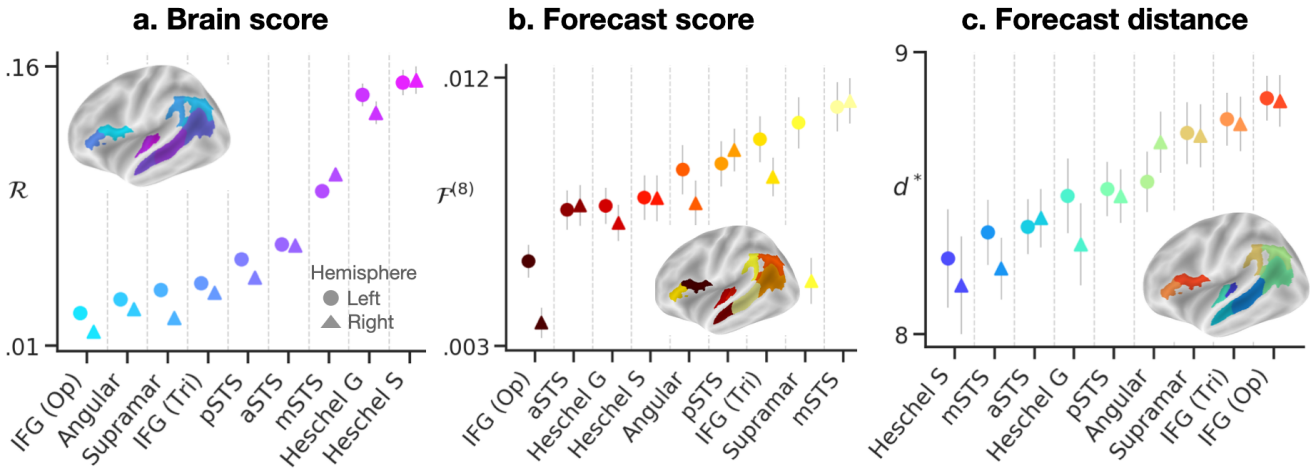


Figure S21: Scores per region of interest. a-c. Brain scores (Figure 4.2**a**, Methods 4.1.5), forecast scores (Figure 4.2**c**, Methods 4.1.5) and forecast distance (Figure 4.2**e** Methods 4.1.5) for nine regions of interests in both the left (circle) and right (triangle) hemispheres. Scores are averaged across voxels within each region of interest and across subjects. Error bars are the standard errors of the mean across subjects. Regions are ordered with respect to their average score in the left hemisphere.

Table S7: Brain and forecast scores in language areas. Scores averaged across all voxels in the brain (Avg), across the ten percent less noisy voxels (w.r.t the noise ceiling, Top10Vox), for the ten percent subjects with the highest brainscore (Top10sub), and averaged across voxels in representative language areas (Heschl, STG, STS and IFG). The last row is the relative improvement of $R(X + \widetilde{X})$ over $R(X)$.

	Avg	Top10Vox	Top10Sub	Heschl	STG	STS	IFG
Brain score, $R(X)$	0.023	0.084	0.049	0.145	0.072	0.072	0.037
Forecast score, $F^{(8)}(X)$	0.005	0.010	0.006	0.008	0.008	0.010	0.008
Relative improvement, $\frac{F^{(8)}(X)}{R(X)}$	23%	13%	39%	5%	21%	13%	18%

Table S8: Brain and forecast scores in language areas, with noise ceiling normalization. Same as Table S7, but, for each voxel, scores are divided by the average noise ceiling.

	Avg	Top10Vox	Top10Sub	Heschl	STG	STS	IFG
Brain score $R(X)$	17%	37%	37%	50%	32%	36%	24%
Forecast score $F^{(8)}(X)$	4%	5%	5%	3%	4%	5%	5%

Table S9: Brain and forecast scores' significance. Same as Table S7, but we indicate the p-values computed across subjects, testing whether the scores (either $R(X)$, $R(X+\widetilde{X})$ or $F(X)$) are different from zero. We use a two-sided Wilcoxon test provided by Scipy. The p-values for the Top10Sub columns are higher because we restrict ourselves to the 10 percent less noisy subjects.

	Avg	Top10Vox	Top10Sub	Heschl	STG	STS	IFG
Brain score $R(X)$	10^{-50}	10^{-51}	10^{-6}	10^{-51}	10^{-51}	10^{-50}	10^{-45}
Forecast score $F^{(8)}(X)$	10^{-35}	10^{-37}	10^{-4}	10^{-32}	10^{-37}	10^{-32}	10^{-29}

6.6.2 Generalisation to other architectures

The analyses in the main manuscript focus on one representative deep neural network: GPT-2 (Radford et al., 2019). Here, we replicate our results with the activations extracted from seven other transformer architectures. We only analyse *causal* models, trained to predict a word from their *previous* context. Note that XLNet is trained to predict both left and right context (Yang et al., 2020), but, here, we only input the model with left context when extracting the activations. Similarly as with GPT-2, we use the pretrained models from Huggingface (labeled 'distilgpt2', 'gpt2', 'gpt2-medium', 'gpt2-large', 'gpt2-large', 'gpt2-xl', 'transfo-xl-wt103', 'xlnet-base-cased', 'xlnet-large-cased'), based on GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2020) and Transformer-XL (Dai et al., 2019) architectures, and focus on one intermediate-to-deep layer of the model ($l = \frac{2}{3} \times n_{\text{layers}}$). For each architecture, we 1) extract the activations corresponding to the subjects' stories (Methods 4.1.5) 2) compute the corresponding brain scores (Methods 4.1.5) and forecast scores (Methods 4.1.5) for each voxel, subject, and forecast distance. As displayed in Supplementary Figure S22, the seven architectures accurately map onto brain activity (Supplementary Figure S22**a**), and the mapping is improved when adding information about around eight words in the future (Supplementary Figure S22**b**).

6.6.3 Robustness of the forecast effect

Below, we show that the forecast effect holds without PCA, with different window sizes, when using *banded* ridge regression (Nunez-Elizalde et al., 2019; Tour et al., 2022) instead of ridge regression, when averaging instead of summing vectors within each TR, when matching the TR with the word onset instead of word offset, when accounting for low-level speech features and when testing for significance across windows at the single-subject level.

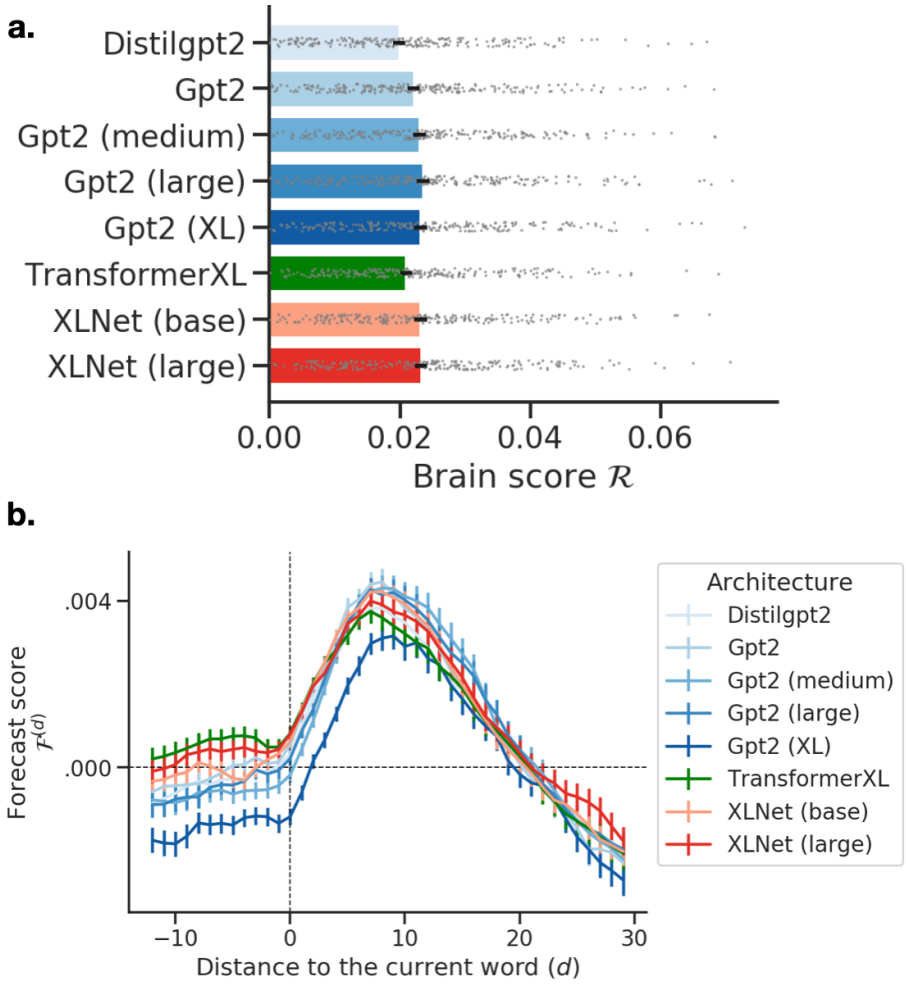


Figure S22: Generalisation to other architectures. a. Brain scores (cf. Figure 4.1**b**, Methods 4.1.5) of eight transformer models, based on XLNet (Yang et al., 2020), TransformerXL (Dai et al., 2019) and GPT-2 (Radford et al., 2019) architectures. We use the pre-trained models from Huggingface and proceed similarly as with GPT-2 (Methods 4.1.5). Brain scores are averaged across voxels and subjects, error bars are the standard errors of the mean across subjects (n=304). **b.** Same as Figure 4.2**d** for the eight transformer architectures.

Replication with banded ridge regression In the main manuscript, we use ℓ_2-regularized ridge regression (as in e.g. (Huth, de Heer, et al., 2016)) followed by a hierarchical comparison of the brain scores: i.e. computing the brain score of the two sets of features (here, X vs. $X \oplus \widetilde{X}$) and then subtracting the scores ($R[X \oplus \widetilde{X}] - R[X]$). To our knowledge, this approach is most conservative when it comes to assess the explained variance of the highest level: the explainable variance shared by two sets of features is by definition fully attributed to the lower-level feature set (i.e. X). Thus, in the worst case scenario, our method underestimates the variance specific to $\widetilde{X}$. This is what happens when the sliding window contains far off future words that are no longer relevant for prediction, and $R[X \oplus \widetilde{X}]$ becomes smaller than $R[X]$.

We replicate our results with banded ridge regression (Nunez-Elizalde et al., 2019; Tour et al., 2022) using the Himalaya (https://github.com/gallantlab/himalaya) package (Tour et al., 2022). Both X and $\widetilde{X}$ models are fitted simultaneously with a specific penalization term learnt for each submodel. We then evaluate the unique variance accounted for by each submodel by zeroing-out either X or $\widetilde{X}$ at test time, predicting Y, and computing Pearson's correlation between predicted and actual Y after zeroing out the specific features. We use the same cross-validation setting as in the paper.

Supplementary Figure S23**a** below displays the brain scores obtained with banded ridge when adding the window for each future word, and Supplementary Figure S23**b** shows the brains scores specifically attributed to the contextual words in $\widetilde{X}$ after zero-ing out X. We obtain similar results as in the original paper, but the forecast effect specific to $\widetilde{X}$ is higher than the one in the paper ($R''[\widetilde{X}]$ peaks at 0.027, while ($R[X \oplus \widetilde{X}] - R[X]$ peaks at 0.004).

Replication without PCA In the manuscript, we apply PCA to the GPT-2 features before applying the FIR and regression (Supplementary Figure S28 and S29). We show in Supplementary Figure S23**c** that the forecast effect holds without applying PCA.

Replication without silent periods and with confounding variables In the main manuscript, we cut the TRs that do not contain words at the beginning of the stories, and do not add to the GPT-2 features confounding variables such as the phoneme rate and word rate. In Supplementary Figure S23**h**, we show that the results hold when the brain and forecast score are computed:

- When removing the empty TRs both at the beginning and end of the recordings (we thus cut the recordings between the first and last word of the story before fitting the ridge regression)
- When including the Word and Phoneme rates as confound variables. These are one-dimensional variables indicating the presence or absence of a word/phoneme.

Replication with different word aggregation in FIR In Supplementary Figure S23**d**, we show that results hold when averaging instead of summing vectors within each TR and when matching the TR with the word onset instead of word offset.

Testing for significance at the single-subject level In the main manuscript, we compare $R(X + X^{(i)})$ to $R(X)$ within each subject and then test the significance *across subjects* ($H_0 : R(X + X^{(d)}) < R(X)$). We show the results hold when testing for significance with a bootstrap test *across windows*, at the single-subject level ($H0 : R(X + X^{(d)}) < R(X + X^{(i)}), i \neq d$). Precisely, for each subject and each distance d, we compute $R(X + X^{(i)}), i \neq d$, with $X^{(i)}$ a sliding window randomly sampled from the stories. We repeat the procedure 1000 times and then estimate the probability of sampling $X^{(i)}$, such that $R(X + X^{(d)}) < R(X + X^{(i)})$. This results in a p-value for each subject and distance d, assessing the significance of $R(X + X^{(d)})$ being greater than $R(X + X^{(i)}), i \neq d$.

In Supplementary Figure S23**f-g**, we show that testing for significance at the single-subject level yields to similar conclusions as across subjects.

Effect of window size In the main manuscript, we use a fixed window size of seven words because it led to the best brain score when varying the length of the window (Supplementary Note 6.6.4). To further assess the impact of the window length on the forecast effect, we compute the forecast scores

for different window sizes (from a size of 5 to 27 words). In Supplementary Figure S23**e**, we find that window length slightly but significantly affects the results. The distances maximizing the forecast scores are on average concentrated between 6 and 12 words, and brain scores are highest for a window of 7-9-11 words. The peak varies with the window length. This phenomenon is partly expected: words that are close to the current word likely carry relevant information (e.g. word n+1). Thus, for short window sizes, not including the closest words is expected to decrease the brain score. This confirms that the forecast result can be found regardless of the window size, and further suggests that forecasts are likely to be slightly longer-term than 8 words.

6.6.4 Controls with a growing window analysis

Testing different window sizes In the previous paragraphs, we use a sliding forecast window with a *fixed* number of words in order to compare the brain scores of representations with the same dimensionality. Here, we test different window sizes by a growing window analysis. Precisely, we build the forecast window $\tilde{X}^{(d)}$ by concatenating the d words succeeding the current word. The size of the window thus varies and d corresponds to both the number of words in the window, and the distance between the last word and the current word. We proceed similarly as in the main manuscript, build forecast window for different distances d and the corresponding forecast scores. As displayed in Supplementary Figure S24, the forecast score is maximal for a window of 8 future words ($d^* = 7.9 \pm 0.5$ on average across subjects), which is consistent with the previous results (Figure 4.2**c**, where $d^* = 8$).

Using random forecast representations We use the same growing window framework and check that adding a forecast window composed of random words does not improve the brain score (Supplementary Figure S24). Precisely, we randomly pick words out of all stories, concatenate the GPT-2 activations of random words to build the forecast windows $\tilde{X}^{(d)}$, and compute the corresponding forecast scores for different distances d. Supplementary Figure S24 shows that random forecast windows do *not* improve our ability to predict brain activity.

Using GPT-2 generations as forecast representations To what extent are the improvements in brain score due to (1) additional information about future words and/or (2) a different way to represent past words? To address this question, we repeat the same analysis with a forecast window input, not of the *true* future words, but with the words *generated* by GPT-2. Specifically, for each word w_k, we 1) GPT-2 with its past context $w_0, \dots w_k$, 2) generate future words $w'_{k+1}, \dots w'_{k+n}$ using different decoding methods (greedy and sampling schemes), 3) extract the corresponding activations $X'_{k+1}, \dots X'_{k+n}$, 4) build the growing windows from these activations and 5) compute their forecast scores. Thus, the brain signals, the current activations X_k and the activations of generated words $X'_{k+n} \dots X'_{k+n}$ are all distinct transformations of the same past words $w_0, \dots w_k$. Note that for step 2), we use Huggingface's sampling scheme with topk=50 and topp=0.95, do_sample=True, max_length=100. For the greedy scheme, we simply set do_sample to False, topp and topk to 1. (Holtzman et al., 2020)). The results show that a window made of *generated* words improves the brain score, although less than a window made of the *true* words of the stories (Supplementary Figure S24), confirming that GPT-2 is an imperfect forecaster.

6.6.5 Contribution of each future word in the forecast effect

In Figure 4.2**b**, we show that adding a sliding window containing future words improves our ability to predict brain activity. To interpret the impact of each word in this improvement, we launch a zero-out analysis. Precisely, we proceed as follows:

- At train time, we proceed similarly as in the main analysis (Figure 4.2**b**) and fit the regression using the current word embedding, concatenated to the sliding window.
- At test time, we zero out the features corresponding to all words after word k (i.e. we replace their embeddings by zeros).
- Finally, we report the Pearson correlation between predicted and actual brain data, when zeroing out words after word k.

This evaluates the importance of the words after word k in the prediction. We repeat the procedure for $k = 1$ to $k = 17$. Note that if the words-to-TR transform had been linear, this analysis would have been identical to an analysis of the coefficients.

We find that zeroing out future words triggers a clear drop in performance (Supplementary Figure S25). This demonstrates that each future word significantly contributes to the prediction in the ridge regression.

To further address this issue, we compute the brain scores when concatenating different continuations to the current word embedding. Specifically, we run the exact same analysis as Figure 4.2**b**, but replacing future words by either zeros or random continuations. These continuations are sensible phrases, of the same length as the true continuations, but randomly sampled from all stories. Supplementary Figure S26 below confirms that adding random continuations does not improve the brain scores.

Overall, Supplementary Figure S25 and S26 show that each future word up to ≈ 10 plays a significant role in the ridge regression.

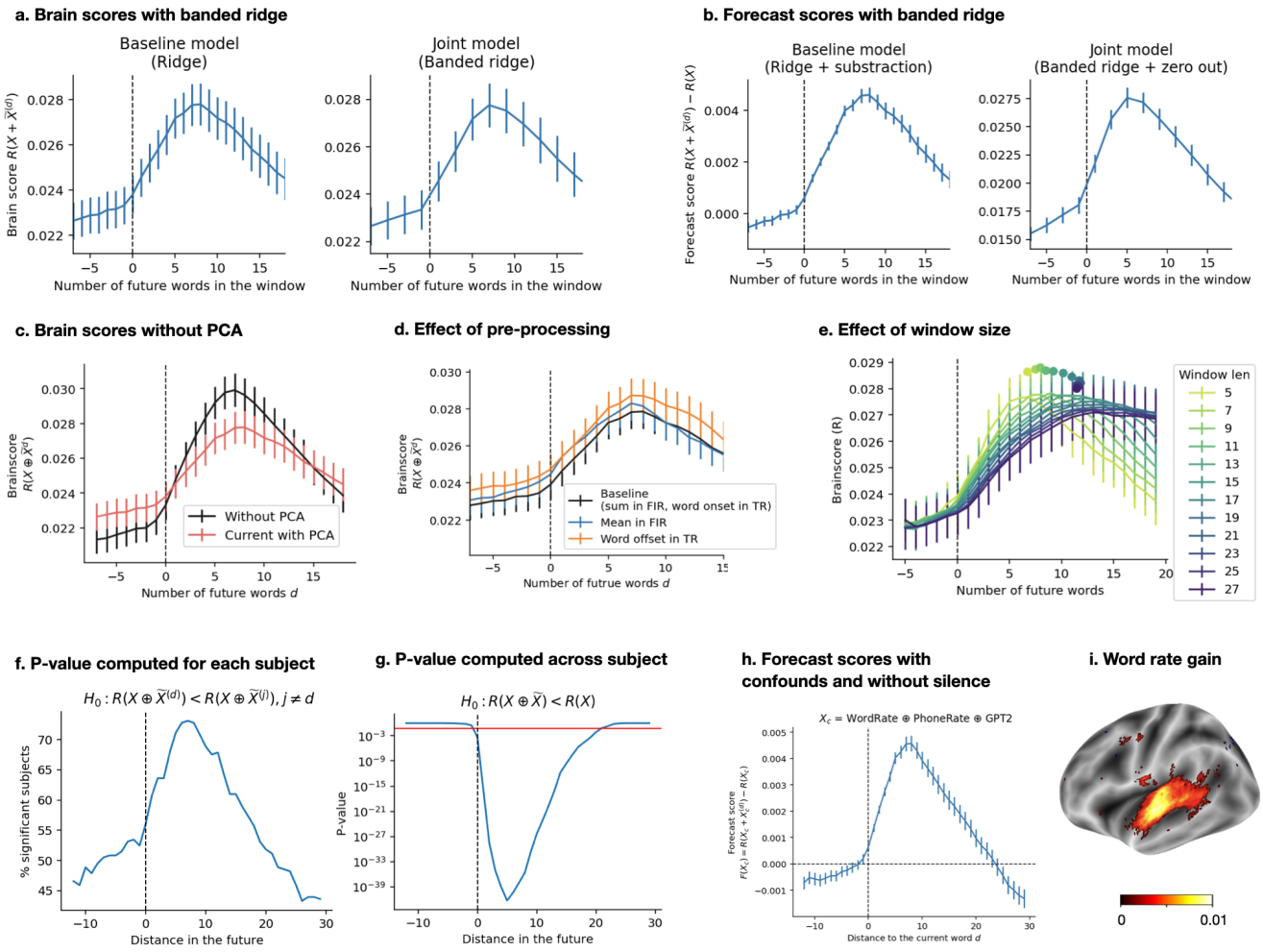


Figure S23: Robustness of the forecast effect. **a.** Replication with banded regression. Brain scores computed with ridge regression (left, same as 4.2 and banded ridge regression (right) (Nunez-Elizalde et al., 2019; Tour et al., 2022). **b.** Forecast scores computed with ridge regression followed by subtraction (left, same as Figure 4.2**d**) and banded ridge regression followed by zero-out (right) (Nunez-Elizalde et al., 2019; Tour et al., 2022). In a banded ridge regression, a model is fitted using both X and $\widetilde{X}$ as input with regularization parameters specific to X and $\widetilde{X}$. We then evaluate the brain score accounted for by the context window $\widetilde{X}$ specifically by zero-ing out X at test time (and the present word in the window). **c.** Forecast scores without PCA. Brain scores when adding the sliding forecast window (Same as Figure 4.2**b**), but without applying PCA before fitting the ridge regression. **c.** Impact of pre-processing parameters. Brain scores when adding the sliding window for different distances d (same as Figure 4.2) (black), but averaging words within TR instead of summing them (blue) and matching the word offset with the TR boundary instead of the word onset (orange). **d.** Effect of window size. Brain scores when adding the forecast window (same as Figure 4.2**b**) computed with a sliding window of size 5 to 27 words. Average peaks across subjects are indicated with a dot. **f-g.** Significance of the forecast effect. In **f.**, the percentage of subjects with a significant bootstrap test for each distance d ($p < 0.05$). For each subject and each distance d, we compute $R(X + X^{(i)}), i \neq d$, with $X^{(i)}$ a sliding window randomly sampled from the stories. We repeat the procedure 1000 times and then estimate the probability of sampling $X^{(i)}$ such that $R(X + X^{(d)}) < R(X + X^{(i)})$, for each subject and distance d. In **g.**, the p-value computed with a one-sided Wilcoxon test across subjects, testing whether the sliding window improves the brain score ($R(X + X^{(d)}) > R(X)$). The red bar indicates the significance threshold ($p = 0.05$). **h.** Forecast scores with confounds and without silent periods. Forecast scores averaged across subjects and voxels (same as 4.2**b**) when (1) including two confounding variables (the word and phone rates) and (2) removing periods without words at the beginning and end of the recordings. The word and phone rates are one-dimensional variables indicating the absence/presence of a word and phoneme. **i.** Word rate gain.

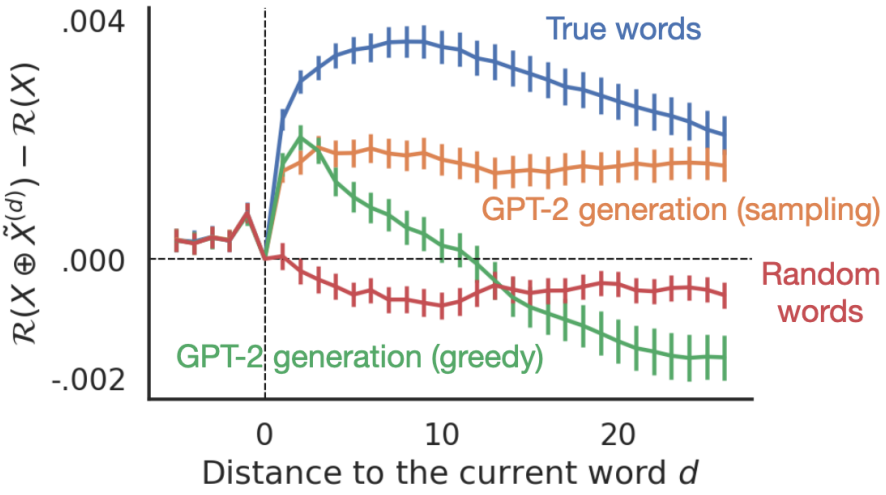


Figure S24: Controls with a growing window analysis. Forecast scores for different types of forecast representations $\widetilde{X}$. Here, we use a growing window analysis: $\widetilde{X}^{(d)}$ is the concatenation of the activations of $|d|$ future ($d > 0$) or past ($d < 0$) words; the size of the window thus varies with the distance. The forecast score is the gain in brain score when concatenating the forecast window (cf. (4.3)). In blue, $\widetilde{X}$ is built out of the true words of the story. In red, $\widetilde{X}$ is built out of randomly picked words from all stories. In green and orange, $\widetilde{X}$ is built out of words generated by GPT-2. Precisely, GPT-2 is input with the current word and its previous context, and we use greedy (green) and sampling (orange) decoding schemes to generate a sequence of expected words. For simplicity, when $d < 0$, $\widetilde{X}$ is the concatenation of d the *true* past words. When $d > 0$, $\widetilde{X}$ is the concatenation of d future words (either true, generated or random words).

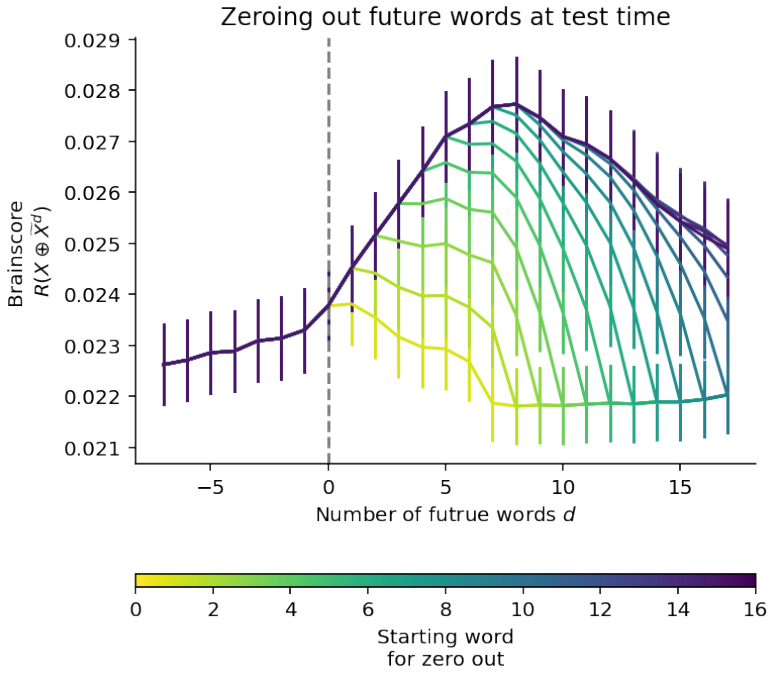


Figure S25: Contribution of future words in the ridge regression. We proceed similarly as in Figure 4.2**b** and fit a ridge regression to predict the fMRI given X and the sliding window $\widetilde{X}^{(d)}$. Yet, at test time, we set to zero (or "zero-out") the dimensions corresponding to all words after word k. We then evaluate the prediction given the zeroed-out input (Pearson's correlation between predicted and true fMRI). On the x-axis, the last word that is not zeroed-out (k, i.e. all words ¿ k are zeroed-out). On the y-axis, the corresponding Pearson correlation.

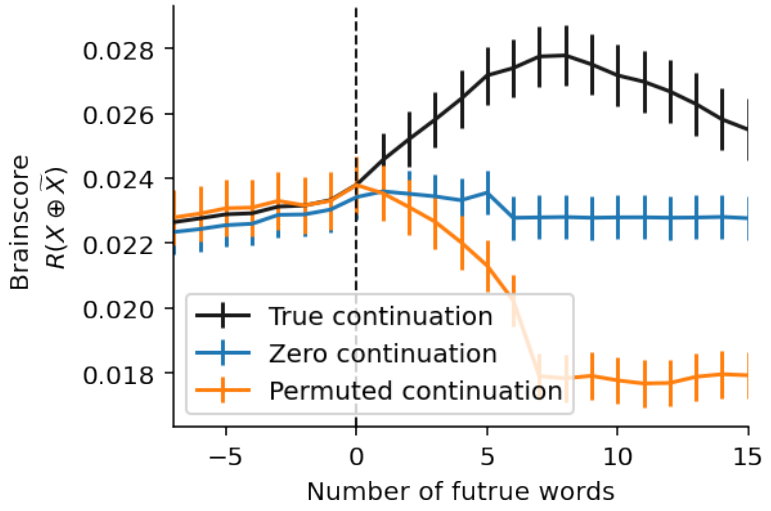


Figure S26: Brain scores when adding different continuations. Same as Figure 4.2**b**, but true continuations (black) are replaced by zeros embeddings (blue) and random continuations sampled from all stories (orange). Random continuations are sensible phrases.

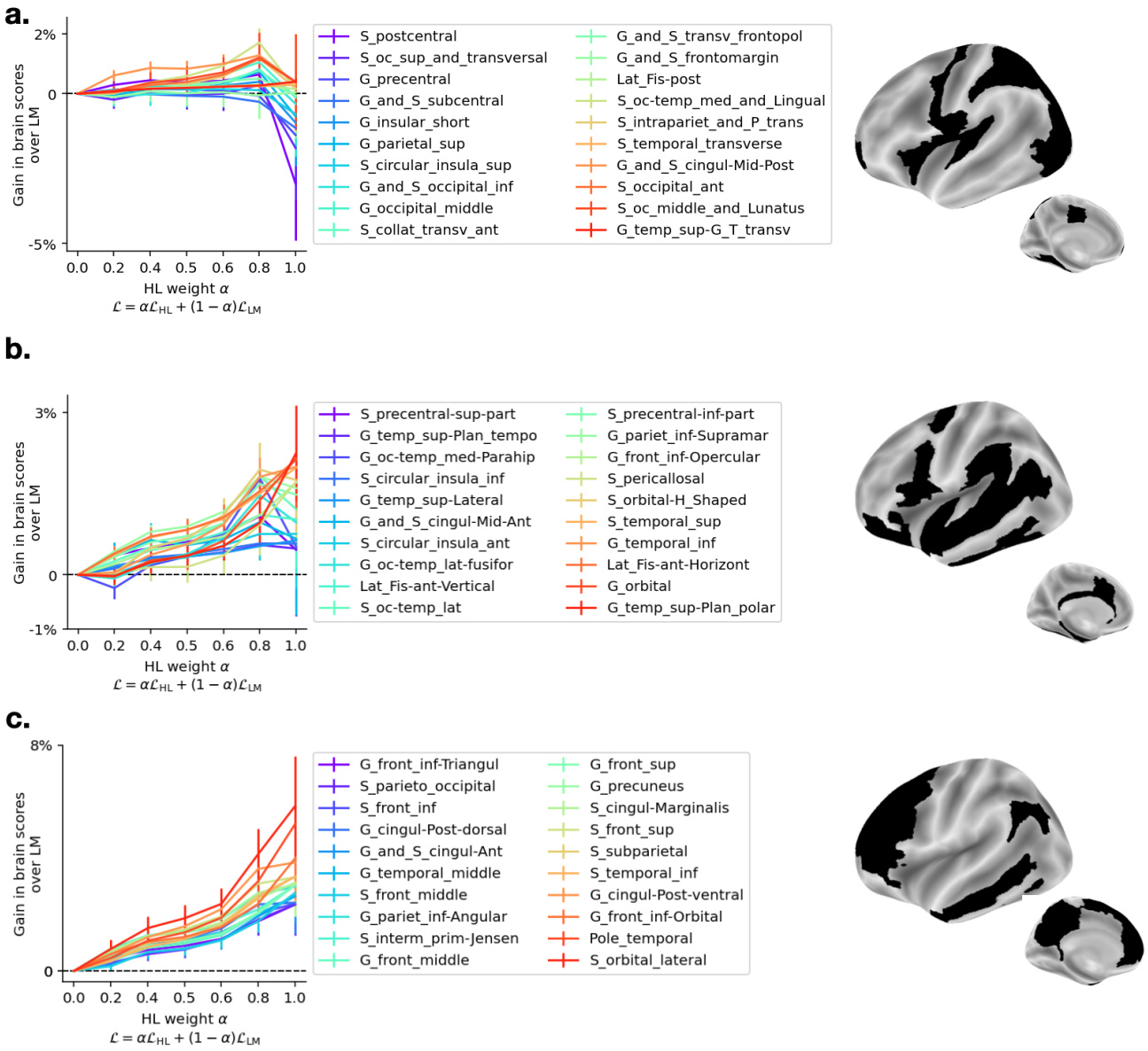


Figure S27: Gain in brain scores when fine-tuning GPT-2 with a mixture of Language Modeling (LM) and High-Level prediction (HL). Gain in brain scores when adding the HL loss, compared to LM only as a function of the weight α_{HL} (Eq. (4.8)). Regions are grouped with respect to their gain, from negative or null improvement (**a.**) to high improvement (**c**). In black, the corresponding regions in the brain. Error bars are SEM across subjects. Brain scores were computed at the voxel-level and then averaged across voxels within 75 regions of interest using Destrieux's parcellation (Destrieux et al., 2010). We only display the 60 regions with highly significant brain scores ($p < 10^{-15}$ using a two-sided Wilcoxon test after FDR correction for multiple comparison across regions).

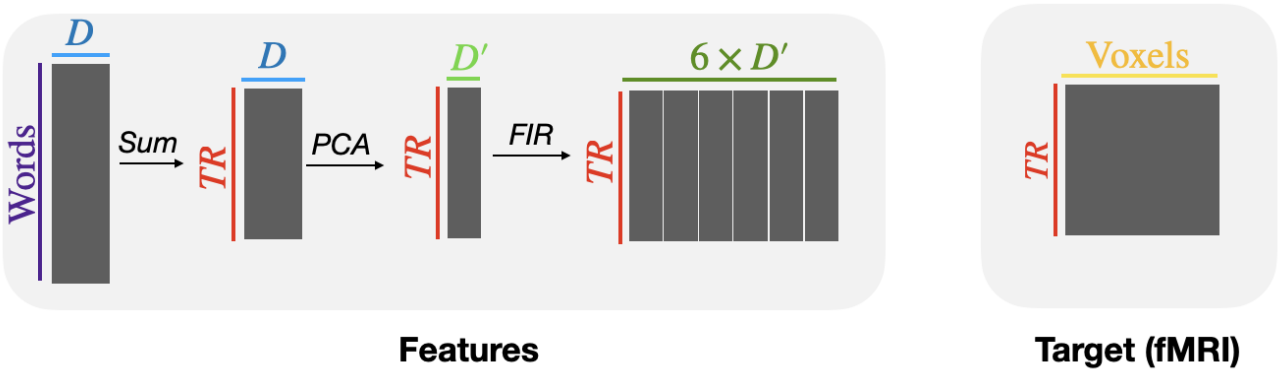


Figure S28: Data pipeline *without* sliding window. Processing steps applied to the raw data of each subject before fitting the ridge regression. The ridge regression is then trained to predict the fMRI target (on the right) given the features (on the left) using a 5-folds cross-validation setting. D is the dimensionality of the language model, here $D = 768$. Words refers to the number of words in the audio recordings the subject listened to while being scanned. If the subject listened to more than one story, the audio recordings are concatenated and Words is the sum of the words of each story. TR is the number of the corresponding fMRI scans. D' is the dimensionality after PCA reduction, here D'=20. 6 is the number of delays used in the FIR.

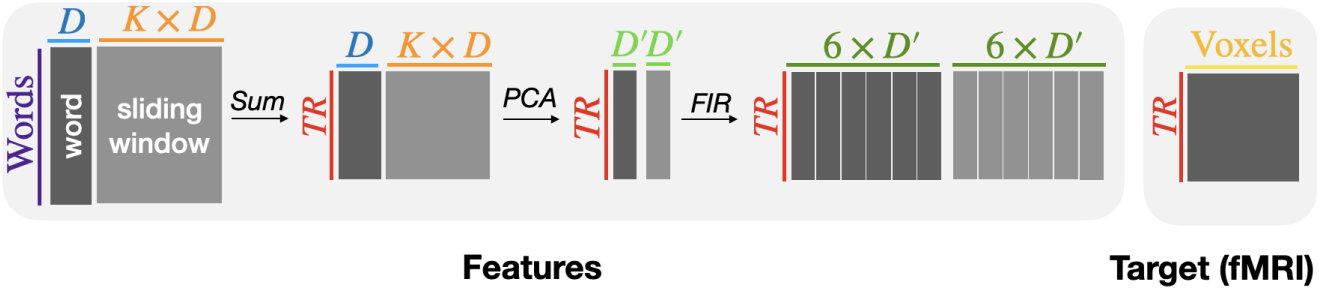


Figure S29: Data pipeline *with* sliding window. Same as Supplementary Figure 8, but we concatenate the sliding window to the current word (in orange and light grey). The sliding window contains the GPT-2 embeddings of past and/or future words. K is the number of words in the sliding window, here $K = 7$.

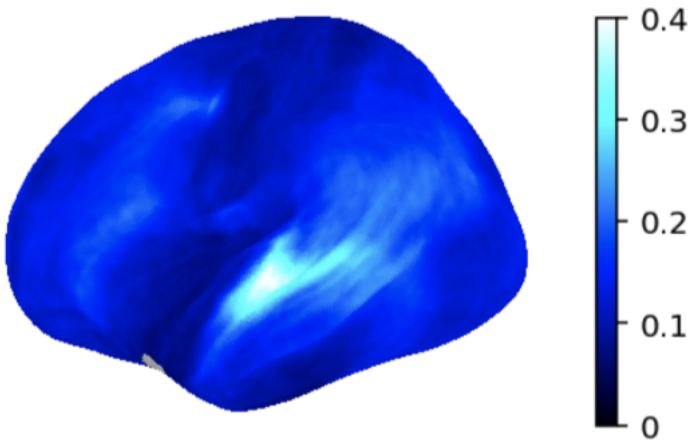


Figure S30: Noise ceiling. Noise ceiling estimates averaged across subjects, for each voxels of the left hemisphere (Methods 4.1.5).

References

Abnar, S., Beinborn, L., Choenni, R., & Zuidema, W. (2019, June). Blackbox meets blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains. *arXiv:1906.01539 [cs, q-bio]*. (arXiv: 1906.01539)

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., ... Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, *8*, 14.

Affolter, N., Egressy, B., Pascual, D., & Wattenhofer, R. (2020). Brain2word: Decoding brain activity for language generation. *arXiv preprint arXiv:2009.04765*.

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., ... Frank, C. (2023, January). *MusicLM: Generating Music From Text*. arXiv. (arXiv:2301.11325 [cs, eess])

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... Kay, K. (2022, January). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. (Number: 1 Publisher: Nature Publishing Group)

Anderson, A. J., Kiela, D., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., ... Lalor, E. C. (2021, May). Deep Artificial Neural Networks Reveal a Distributed Cortical Network Encoding Propositional Sentence-Level Meaning. *Journal of Neuroscience*, *41*(18), 4100–4119. (Publisher: Society for Neuroscience Section: Research Articles)

Anderson, A. J., Lalor, E. C., Lin, F., Binder, J. R., Fernandino, L., Humphries, C. J., ... Wang, X. (2019, June). Multiple Regions of a Cortical Network Commonly Encode the Meaning of Words in Multiple Grammatical Positions of Read Sentences. *Cerebral Cortex*, *29*(6), 2396–2411. (Publisher: Oxford Academic)

Antonello, R., & Huth, A. (2022, December). Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data. *Neurobiology of Language*, 1–16. (_eprint: https://direct.mit.edu/nol/article-pdf/doi/10.1162/nol_a_00087/2062803/nol_a_00087.pdf)

Antonello, R., Turek, J. S., Vo, V., & Huth, A. (2021). Low-dimensional structure in the space of language representations is reflected in brain responses. *Advances in Neural Information Processing Systems*, *34*.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., … Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Lrec.*

Arehalli, S., & Linzen, T. (2020). Neural language models capture some, but not all, agreement attraction effects.

Athanasiou, N., Iosif, E., & Potamianos, A. (2018). Neural Activation Semantic Models: Computational lexical semantic models of localized neural activations. , 12.

Attardi, G. (2015). *Wikiextractor.* https://github.com/attardi/wikiextractor. GitHub.

Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., … Lee, H. (2019). What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the ieee international conference on computer vision* (pp. 4715–4723).

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., & Auli, M. (2022, October). *data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language.* arXiv. (arXiv:2202.03555 [cs])

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020, October). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.* arXiv. (arXiv:2006.11477 [cs, eess])

Baillet, S. (2017, March). Magnetoencephalography for brain electrophysiology and imaging. *Nature Neuroscience*, *20*(3), 327–339. (Number: 3 Publisher: Nature Publishing Group)

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., … Fung, P. (2023, February). *A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity.* arXiv. (arXiv:2302.04023 [cs])

Bardes, A., Ponce, J., & LeCun, Y. (2022, January). *VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning.* arXiv. (arXiv:2105.04906 [cs])

Baroni, M. (2020, February). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190307. (arXiv: 1904.00157)

Bashivan, P., Kar, K., & DiCarlo, J. J. (2018, November). *Neural Population Control via Deep Image Synthesis.* bioRxiv. (Pages: 461525 Section: New Results)

Bazeille, T., DuPre, E., Richard, H., Poline, J.-B., & Thirion, B. (2021, December). An empirical evaluation of functional alignment using inter-subject decoding. *NeuroImage*, *245*, 118683.

Begus, G., Zhou, A., & Zhao, C. (2022). Encoding of speech in convolutional layers and the brain stem based on language experience. *bioRxiv*.

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

Bellman, R. (1966). Dynamic programming. *Science*, *153*(3731), 34–37.

Beltagy, I., Peters, M. E., & Cohan, A. (2020, December). *Longformer: The Long-Document Transformer.* arXiv. (arXiv:2004.05150 [cs])

Bengio, Y., Courville, A., & Vincent, P. (2014, April). *Representation Learning: A Review and New Perspectives.* arXiv. (arXiv:1206.5538 [cs])

Bengio, Y., Ducharme, R., & Vincent, P. (2001). A Neural Probabilistic Language Model. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13* (pp. 932–938). MIT Press.

Benjamini, Y. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society: series B (statistical methodology)*, *72*(4), 405–416.

Berezutskaya, J., Freudenburg, Z. V., Güçlü, U., van Gerven, M. A., & Ramsey, N. F. (2017). Neural tuning to low-level features of speech throughout the perisylvian cortex. *Journal of Neuroscience*, *37*(33), 7906–7920.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016, May). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, *33*(3-4), 130–174. (Publisher: Routledge _eprint: https://doi.org/10.1080/02643294.2016.1147426)

Bingham, E., & Mannila, H. (2001, August). Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 245–250). New York, NY, USA: Association for Computing Machinery.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., … Turian, J. (2020, November). Experience Grounds Language. *arXiv:2004.10151 [cs]*. (arXiv: 2004.10151)

Blank, I., Balewski, Z., Mahowald, K., & Fedorenko, E. (2016, February). Syntactic processing is distributed across the language system. *NeuroImage*, *127*, 307–323.

Bohn, O.-S. (2017). Cross-language and second language speech perception. *The handbook of psycholinguistics*, 213–239.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. (Place: Cambridge, MA Publisher: MIT Press)

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., … Sifre, L. (2022, February). *Improving language models by retrieving from trillions of tokens.* arXiv. (arXiv:2112.04426 [cs])

Borgholt, L., Havtorn, J. D., Edin, J., Maaløe, L., & Igel, C. (2022). A brief overview of unsupervised neural speech representation learning. *arXiv preprint arXiv:2203.01829*.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2006, November). The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychological review*, *113*, 787–821.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2013, April). Reconciling time, space and function: a new dorsal-ventral stream model of sentence comprehension. *Brain and Language*, *125*(1), 60–76.

Bowman, S. R., & Dahl, G. (2021, June). What Will it Take to Fix Benchmarking in Natural Language Understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4843–4855). Online: Association for Computational Linguistics.

Bransford, J. D., & Johnson, M. K. (1972, December). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 717–726.

Breiman, L. (2001, October). Random Forests. *Machine Learning*, *45*(1), 5–32.

Brennan, J. (2016). Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, *10*(7), 299–313.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012, February). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, *120*(2), 163–173.

Brennan, J. R., & Hale, J. T. (2019, January). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE*, *14*(1), e0207741. (Publisher: Public Library of Science)

Brennan, J. R., & Pylkkänen, L. (2017). Meg evidence for incremental sentence composition in the anterior temporal lobe. *Cognitive science*, *41*, 1515–1531.

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, *28*(24), 3976–3983.

Broderick, M. P., Anderson, A. J., Liberto, G. M. D., Crosse, M. J., & Lalor, E. C. (2018, March). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology*, *28*(5), 803–809.e3. (Publisher: Elsevier)

Broderick, M. P., Zuk, N. J., Anderson, A. J., & Lalor, E. C. (2020, December). *More than Words: Neurophysiological Correlates of Semantic Dissimilarity Depend on Comprehension of the Speech Narrative* (preprint). Neuroscience.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020, July). Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*. (arXiv: 2005.14165)

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., … DiCarlo, J. J. (2014, December). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*, *10*(12), e1003963. (Publisher: Public Library of Science)

Caucheteux, C., Gramfort, A., & King, J.-R. (2021a, July). Disentangling syntax and semantics in the brain with deep networks. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 1336–1348). PMLR. (ISSN: 2640-3498)

Caucheteux, C., Gramfort, A., & King, J.-R. (2021b, November). Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. In *EMNLP 2021 - Conference on Empirical Methods in Natural Language Processing*. Punta Cana (and Online), Dominican Republic.

Caucheteux, C., Gramfort, A., & King, J.-R. (2022, September). Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, *12*(1), 16327. (Number: 1 Publisher: Nature Publishing Group)

Caucheteux, C., Gramfort, A., & King, J.-R. (2023, March). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 1–12. (Publisher: Nature Publishing Group)

Caucheteux, C., & King, J.-R. (2022, February). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 1–10. (Number: 1 Publisher: Nature Publishing Group)

Cerullo, M. (2022). *In Defense of Blake Lemoine and the Possibility of Machine Sentience in Lamda.*

Chalmers, D. J. (2023). *Could a Large Language Model Be Conscious?*

Chehab, O., Defossez, A., Loiseau, J.-C., Gramfort, A., & King, J.-R. (2022, September). *Deep Recurrent Encoder: A scalable end-to-end network to model brain signals.* arXiv. (arXiv:2103.02339 [cs, q-bio])

Chen, H.-H., & Cherkassky, V. (2020, August). Performance metrics for online seizure prediction. *Neural Networks*, *128*, 22–32.

Chen, P.-H. C., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., & Ramadge, P. J. (2015). A Reduced-Dimension fMRI Shared Response Model. In *Advances in Neural Information Processing Systems* (Vol. 28). Curran Associates, Inc.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, June). A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709 [cs, stat]*. (arXiv: 2002.05709)

Chomsky, N. (1957). *Syntactic structures*. Oxford, England: Mouton. (Pages: 116)

Chomsky, N. (2000). Linguistics and brain science. *Image, language, brain*, 13–28.

Chomsky, N. (2006). *Language and mind*. Cambridge University Press.

Chomsky, N. (2014). *The minimalist program*. MIT press.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., … Fiedel, N. (2022, October). *PaLM: Scaling Language Modeling with Pathways.* arXiv. (arXiv:2204.02311 [cs])

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in cognitive sciences*, *23*(4), 305–317.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020, March). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv:2003.10555 [cs]*. (arXiv: 2003.10555)

Co., B. M. D. T. (2019). *Magic data chinese mandarin conversational speech.* http://www.imagicdatatech.com/index.php/home/dataopensource/data_info/id/101. doi: https://doi.org/10.35111/zrz3-fw98

Cogan, G. B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., & Pesaran, B. (2014). Sensory–motor transformations for speech occur bilaterally. *Nature*, *507*(7490), 94–98.

Cohen, L., Salondy, P., Pallier, C., & Dehaene, S. (2021). How does inattention affect written and spoken language processing? *Cortex*, *138*, 212–227.

Cohen, U., Chung, S., Lee, D. D., & Sompolinsky, H. (2020). Separability and geometry of object manifolds in deep neural networks. *Nature communications*, *11*(1), 1–13.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019, June). Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *arXiv:1901.02860 [cs, stat]*. (arXiv: 1901.02860)

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

Dehaene, S., Yann, L., & Girardon, J. (2018). *La plus belle histoire de l'intelligence: des origines aux neurones artificiels: vers une nouvelle étape de l'évolution*. Robert Laffont.

Dehghani, M., Boghrati, R., Man, K., Hoover, J., Gimbel, S. I., Vaswani, A., … Kaplan, J. T. (2017, December). Decoding the neural representation of story meanings across languages: Decoding the Neural Representation. *Human Brain Mapping*, *38*(12), 6096–6106.

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019, September). The Representation of Semantic Information Across Human Cerebral Cortex During Listening Versus Reading Is Invariant to Stimulus Modality. *Journal of Neuroscience*, *39*(39), 7722–7736. (Publisher: Society for Neuroscience Section: Research Articles)

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010, October). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, *53*(1), 1–15.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. (arXiv: 1810.04805)

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020, April). *Jukebox: A Generative Model for Music.* arXiv. (arXiv:2005.00341 [cs, eess, stat])

DiCarlo, J. J., & Cox, D. D. (2007, August). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341. Retrieved 2023-03-16, from https://www.sciencedirect.com/science/article/pii/S1364661307001593 doi: 10.1016/j.tics.2007.06.010

Dichter, B. K., Breshears, J. D., Leonard, M. K., & Chang, E. F. (2018). The control of vocal pitch in human laryngeal motor cortex. *Cell*, *174*(1), 21–31.

Dinan, E., Logacheva, V., Malykh, V., Miller, A., Shuster, K., Urbanek, J., ... Weston, J. (2019, January). *The Second Conversational Intelligence Challenge (ConvAI2).* arXiv. (arXiv:1902.00098 [cs])

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

Donhauser, P. W., & Baillet, S. (2020, January). Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron*, *105*(2), 385–393.e9.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59.

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., & King, J.-R. (2022, August). *Decoding speech from non-invasive brain recordings.* arXiv. (arXiv:2208.12266 [cs, eess, q-bio])

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017, May). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194.

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., & Goldberg, Y. (2021, December). Measuring and Improving Consistency

in Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, *9*, 1012–1031. (_eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00410/1975957/tacl_a_00410.pdf)

El-Nouby, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., … Jegou, H. (2021, June). XCiT: Cross-Covariance Image Transformers. *arXiv:2106.09681 [cs]*. (arXiv: 2106.09681 version: 2)

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., … Gorgolewski, K. J. (2019, January). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116.

Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., & Auli, M. (2019, July). *ELI5: Long Form Question Answering*. arXiv. (arXiv:1907.09190 [cs])

Fan, A., Lewis, M., & Dauphin, Y. (2018, May). Hierarchical Neural Story Generation. *arXiv:1805.04833 [cs]*. (arXiv: 1805.04833)

Fedorenko, E., Blank, I., Siegelman, M., & Mineroff, Z. (2020, February). Lack of selectivity for syntax relative to word meanings throughout the language network. *bioRxiv*, 477851. (Publisher: Cold Spring Harbor Laboratory Section: New Results)

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010, August). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194.

Fedorenko, E., Nieto-Castanon, A., & Kanwisher, N. (2012). Lexical and syntactic representations in the brain: an fmri investigation with multi-voxel pattern analyses. *Neuropsychologia*, *50*(4), 499–513.

Fedorenko, E., Scott, T., Brunner, P., Coon, W., Pritchett, B., Schalk, G., & Kanwisher, N. (2016, September). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences of the United States of America*, *113*.

Felleman, D. J., & Van Essen, D. C. (1991, February). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, *1*(1), 1–47.

Fiebach, C., Schlesewsky, M., Lohmann, G., von Cramon, D., & Friederici, A. (2005). Revisiting the role of Broca's area in sentence processing: Syntactic integration versus syntactic working memory. *Human Brain Mapping*, *24*(2), 79–91. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.20070)

Fischl, B. (2012). Freesurfer. *Neuroimage*, *62*(2), 774–781.

Forseth, K. J., Hickok, G., Rollo, P. S., & Tandon, N. (2020, October). Language prediction mechanisms in human auditory cortex. *Nature Communications*, *11*(1), 5240. (Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cortex;Language;Neural encoding Subject_term_id: cortex;language;neural-encoding)

Frankle, J., & Carbin, M. (2019, March). *The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks.* arXiv. (arXiv:1803.03635 [cs])

Friederici, A. D. (1999). The neurobiology of language comprehension. In *Language comprehension: A biological perspective* (pp. 265–304). Springer.

Friederici, A. D. (2011, October). The Brain Basis of Language Processing: From Structure to Function. *Physiological Reviews*, *91*(4), 1357–1392.

Friederici, A. D., Opitz, B., & Von Cramon, D. Y. (2000). Segregating semantic and syntactic aspects of processing in the human brain: an fmri investigation of different word types. *Cerebral cortex*, *10*(7), 698–705.

Friston, K., & Kiebel, S. (2009, May). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1521), 1211–1221.

Friston, K. J., & Stephan, K. E. (2007, December). Free-energy and the brain. *Synthese*, *159*(3), 417–458.

Gallant, J. (2013). in 'reading minds'. *Nature*, *502*(7472), 428.

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: a review of underlying mechanisms. *Clinical neurophysiology*, *120*(3), 453–463.

Gauthier, J., & Ivanova, A. (2018, June). Does the brain represent words? An evaluation of brain decoding studies of language understanding. *arXiv:1806.00591 [cs]*. (arXiv: 1806.00591)

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

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., … Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 776–780).

Geschwind, N. (1965, June). Disconnexion syndromes in animals and man. I. *Brain: A Journal of Neurology*, *88*(2), 237–294.

Gifford, A. T., Lahner, B., Saba-Sadiya, S., Vilas, M. G., Lascelles, A., Oliva, A., … Cichy, R. M. (2023, January). *The Algonauts Project 2023 Challenge: How the Human Brain Makes Sense of Natural Scenes.* arXiv. (arXiv:2301.03198 [cs, q-bio])

Gilbert, S. J., & Burgess, P. W. (2008, February). Executive function. *Current biology: CB*, *18*(3), R110–114.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., … Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology*, *26*(2), 248–265.

Givón, T. (2001). *Syntax: an introduction* (Vol. 1). John Benjamins Publishing.

Goertzel, B. (2014, December). Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*, *5*(1), 1–48.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., … Hasson, U. (2022, March). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. (Number: 3 Publisher: Nature Publishing Group)

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., … Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*. (Publisher: Frontiers)

Graves, A. (2012). Connectionist temporal classification. In *Supervised sequence labelling with recurrent neural networks* (pp. 61–93). Springer.

Graves, R. E. (1997, April). The Legacy of the Wernicke-Lichtheim Model*. *Journal of the History of the Neurosciences*, *6*(1), 3–20.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., … Valko, M. (2020, September). *Bootstrap your own latent: A new approach to self-supervised Learning.* arXiv. (arXiv:2006.07733 [cs, stat])

Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2020). Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. *bioRxiv*.

Güçlü, U., & Gerven, M. A. J. v. (2015, July). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014. (Publisher: Society for Neuroscience Section: Articles)

Hagoort, P. (2005). On broca, brain, and binding: a new framework. *Trends in cognitive sciences*, *9*(9), 416–423.

Hagoort, P. (2013). MUC (Memory, Unification, Control) and beyond. *Frontiers in Psychology*, *4*.

Hagoort, P., Baggio, G., & Wlllems, R. M. (2009). Semantic unification. In *The cognitive neurosciences, 4th ed* (pp. 819–835). Cambridge, MA, US: Massachusetts Institute of Technology.

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

Hale, J., Campanelli, L., Li, J., Bhattasali, S., Pallier, C., & Brennan, J. (2021). Neurocomputational models of language processing. *Annual Review of Linguistics*.

Hale, J., Dyer, C., Kuncoro, A., & Brennan, J. (2018, July). Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2727–2736). Melbourne, Australia: Association for Computational Linguistics.

Hamilton, L., & Huth, A. (2018, July). The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*, 1–10.

Hart, B., & Risley, T. R. (1992). American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental psychology*, *28*(6), 1096.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017, July). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258. (Publisher: Elsevier)

Hasson, U., Egidi, G., Marelli, M., & Willems, R. M. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition*, *180*, 135–157. (Place: Netherlands Publisher: Elsevier Science)

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020, June). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife*, *9*, e56601. (Publisher: eLife Sciences Publications, Ltd)

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020, March). Momentum Contrast for Unsupervised Visual Representation Learning. *arXiv:1911.05722 [cs]*. (arXiv: 1911.05722)

Hebb, D. O. (1949). *The organization of behavior; a neuropsychological theory*. Oxford, England: Wiley. (Pages: xix, 335)

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

Heilbron, M., & Chait, M. (2018, October). Great Expectations: Is there Evidence for Predictive Coding in Auditory Cortex? *Neuroscience*, *389*, 54–73.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010, June). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems* (Vol. 28). Curran Associates, Inc.

Hermes, D., Rangarajan, V., Foster, B. L., King, J.-R., Kasikci, I., Miller, K. J., & Parvizi, J. (2017). Electrophysiological responses in the ventral temporal cortex during reading of numerals and calculation. *Cerebral cortex*, *27*(1), 567–575.

Hickok, G., & Poeppel, D. (2007, May). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. (Number: 5 Publisher: Nature Publishing Group)

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... Lerchner, A. (2017). β-VAE: LEARNING BASIC VISUAL CONCEPTS WITH A CONSTRAINED VARIATIONAL FRAMEWORK.

Hochreiter, S., & Schmidhuber, J. (1997, December). Long Short-term Memory. *Neural computation*, *9*, 1735–80.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... Sifre, L. (2022, March). *Training Compute-Optimal Large Language Models.* arXiv. (arXiv:2203.15556 [cs])

Hollenstein, N., de la Torre, A., Langer, N., & Zhang, C. (2019, November). CogniVal: A Framework for Cognitive Word Embedding Evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (pp. 538–549). Hong Kong, China: Association for Computational Linguistics.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020, February). The Curious Case of Neural Text Degeneration. *arXiv:1904.09751 [cs]*. (arXiv: 1904.09751)

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python.* Zenodo.

Hu, Z., Chan, H. P., Liu, J., Xiao, X., Wu, H., & Huang, L. (2022, March). PLANET: Dynamic Content Planning in Autoregressive Transformers for Long-form Text Generation. *arXiv:2203.09100 [cs]*. (arXiv: 2203.09100)

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., & Xing, E. P. (2018, September). *Toward Controlled Generation of Text.* arXiv. (arXiv:1703.00955 [cs, stat])

Huang, N., Slaney, M., & Elhilali, M. (2018). Connecting deep neural networks to physical, perceptual, and electrophysiological auditory signals. *Frontiers in neuroscience*, *12*, 532.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*, 574–591. (Place: United Kingdom Publisher: Cambridge University Press)

Humphreys, L. G. (1939). Acquisition and extinction of verbal expectations in a situation analogous to conditioning. *Journal of Experimental Psychology*, *25*, 294–301. (Place: US Publisher: American Psychological Association)

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2006). Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *Journal of Cognitive Neuroscience*, *18*(4), 665–679. (Place: US Publisher: MIT Press)

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2007, July). Time course of semantic processes during sentence comprehension: An fMRI study. *NeuroImage*, *36*(3), 924–932.

Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020). Compositionality decomposed: how do neural networks generalise? *Journal of Artificial Intelligence Research*, *67*, 757–795.

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

Huth, A. G., Lee, T., Nishimoto, S., Bilenko, N. Y., Vu, A. T., & Gallant, J. L. (2016, October). Decoding the Semantic Content of Natural Movies from Human Brain Activity. *Frontiers in Systems Neuroscience*, *10*.

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

Hénaff, O. J., Srinivas, A., De Fauw, J., Razavi, A., Doersch, C., Eslami, S. M. A., & Oord, A. v. d. (2019, December). Data-Efficient Image Recognition with Contrastive Predictive Coding. *arXiv:1905.09272 [cs]*. (arXiv: 1905.09272)

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., … Grave, E. (2022, November). *Atlas: Few-shot Learning with Retrieval Augmented Language Models.* arXiv. (arXiv:2208.03299 [cs])

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press. (Publication Title: Foundations of Language)

Jain, S., & Huth, A. G. (2018, May). *Incorporating Context into Language Encoding Models for fMRI* (preprint). Neuroscience.

Jain, S., Vo, V., Wehbe, L., & Huth, A. (2023, January). Computational language modeling and the promise of in silico experimentation. *Neurobiology of Language*, 1–65.

Jat, S., Tang, H., Talukdar, P., & Mitchell, T. (2019, June). Relating Simple Sentence Representations in Deep Neural Networks and the Brain. *arXiv:1906.11861 [cs]*. (arXiv: 1906.11861)

Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3651–3657). Florence, Italy: Association for Computational Linguistics.

Jernite, Y., Bowman, S. R., & Sontag, D. (2017, April). Discourse-Based Objectives for Fast Unsupervised Sentence Representation Learning. *arXiv:1705.00557 [cs, stat]*. (arXiv: 1705.00557)

Jiang, A. Q., Welleck, S., Zhou, J. P., Li, W., Liu, J., Jamnik, M., … Lample, G. (2022, November). *Draft, Sketch, and Prove: Guiding Formal Theorem Provers with Informal Proofs.* arXiv. (arXiv:2210.12283 [cs])

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020, January). SpanBERT: Improving Pre-training by Representing and Predicting Spans. *arXiv:1907.10529 [cs]*. (arXiv: 1907.10529)

Kanwisher, N., Khosla, M., & Dobs, K. (2023, March). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, *46*(3), 240–254. Retrieved 2023-03-16, from https://www.sciencedirect.com/science/article/pii/S0166223622002624 doi: 10.1016/j.tins.2022.12.008

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., … Amodei, D. (2020, January). *Scaling Laws for Neural Language Models.* arXiv. Retrieved 2023-03-16, from http://arxiv.org/abs/2001.08361 (arXiv:2001.08361 [cs, stat]) doi: 10.48550/arXiv.2001.08361

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008, March). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355. (Number: 7185 Publisher: Nature Publishing Group)

Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: windows onto the role of task constraints. *Current opinion in neurobiology*, *55*, 121–132.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018, May). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630–644.e16.

Keller, G. B., & Mrsic-Flogel, T. D. (2018, October). Predictive Processing: A Canonical Cortical Computation. *Neuron*, *100*(2), 424–435.

Kellis, S., Miller, K., Thomson, K., Brown, R., House, P., & Greger, B. (2010). Decoding spoken words using local field potentials recorded from the cortical surface. *Journal of neural engineering*, *7*(5), 056007.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014, November). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, *10*(11), e1003915. (Publisher: Public Library of Science)

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2018). Deep neural networks in computational neuroscience. *BioRxiv*, 133504.

King, J.-R., & Dehaene, S. (2014, April). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210.

King, J.-R., Gwilliams, L., Holdgraf, C., Sassenhagen, J., Barachant, A., Engemann, D., … Gramfort, A. (2018). Encoding and Decoding Neuronal Dynamics: Methodological Framework to Uncover the Algorithms of Cognition. , 19.

Knauff, M., & May, E. (2006, January). Mental imagery, reasoning, and blindness. *Quarterly Journal of Experimental Psychology (2006)*, *59*(1), 161–177.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., … Herbst, E. (2007, June). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* (pp. 177–180). Prague, Czech Republic: Association for Computational Linguistics.

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012, July). Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*, *75*(2), 265–270.

Koumura, T., Terashima, H., & Furukawa, S. (2019). Cascaded tuning to amplitude modulation for natural sound recognition. *Journal of Neuroscience*, *39*(28), 5517–5533.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446. (_eprint: https://doi.org/10.1146/annurev-vision-082114-035447)

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. (Publisher: Frontiers)

Krishna, K., Roy, A., & Iyyer, M. (2021, May). Hurdles to Progress in Long-form Question Answering. *arXiv:2103.06332 [cs]*. (arXiv: 2103.06332)

Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T., & Pruitt, J. (2005). Early speech perception and later language development: Implications for the" critical period". *Language learning and development*, *1*(3-4), 237–264.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual review of psychology*, *62*, 621–647.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Lake, B., & Baroni, M. (2018, July). Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2873–2882). PMLR. (ISSN: 2640-3498)

Lake, B. M., & Murphy, G. L. (2021, April). Word meaning in minds and machines. *arXiv:2008.01766 [cs]*. (arXiv: 2008.01766)

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016, November). Building Machines That Learn and Think Like People. *arXiv:1604.00289 [cs, stat]*. (arXiv: 1604.00289)

Lakretz, Y., Dehaene, S., & King, J.-R. (2020). What limits our capacity to process nested long-range dependencies in sentence comprehension? *Entropy*, *22*(4), 446.

Lakretz, Y., Desbordes, T., King, J.-R., Crabbé, B., Oquab, M., & Dehaene, S. (2021, January). Can RNNs learn Recursive Nested Subject-Verb Agreements? *arXiv:2101.02258 [cs]*. (arXiv: 2101.02258)

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019, June). The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 11–20). Minneapolis, Minnesota: Association for Computational Linguistics.

Lample, G., & Conneau, A. (2019, January). Cross-lingual Language Model Pretraining. *arXiv:1901.07291 [cs]*. (arXiv: 1901.07291)

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020, February). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942 [cs]*. (arXiv: 1909.11942)

LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27.

LeCun, Y., Bengio, Y., & Hinton, G. (2015, May). Deep learning. *Nature*, *521*(7553), 436–444. (Number: 7553 Publisher: Nature Publishing Group)

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989, December). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551.

Lee, C. S., Aly, M., & Baldassano, C. (2021, April). Anticipation of temporally structured events in the brain. *eLife*, *10*, e64972.

Legg, S., & Hutter, M. (2007, December). *Universal Intelligence: A Definition of Machine Intelligence.* arXiv. (arXiv:0712.3329 [cs])

Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006, August). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, *442*(7102), 572–575. Retrieved 2023-03-16, from https://www.nature.com/articles/nature04951 (Number: 7102 Publisher: Nature Publishing Group) doi: 10.1038/nature04951

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011, February). Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *Journal of Neuroscience*, *31*(8), 2906–2915.

Levesque, H., Davis, E., & Morgenstern, L. (2012). The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., … Zettlemoyer, L. (2019, October). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv:1910.13461 [cs, stat]*. (arXiv: 1910.13461)

Li, J., Bhattasali, S., Zhang, S., Franzluebbers, B., Luh, W.-M., Spreng, R. N., … Hale, J. T. (2021). Le petit prince: A multilingual fmri corpus using ecological stimuli. *bioRxiv*.

Lichteim, L. (1885, January). On Aphasia1. *Brain*, *7*(4), 433–484.

Lin, C.-Y. (2004, July). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*, 195–212.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019, July). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. (arXiv: 1907.11692)

Lopopolo, A., Frank, S. L., Bosch, A. v. d., & Willems, R. M. (2017, May). Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PLOS ONE*, *12*(5), e0177794. (Publisher: Public Library of Science)

Lu, Q., Hasson, U., & Norman, K. A. (2022, February). A neural network model of when to retrieve and encode episodic memories. *eLife*, *11*, e74445.

Lycan, W. G. (2018). *Philosophy of language: A contemporary introduction*. Routledge.

Mahowald, K., & Fedorenko, E. (2016, October). Reliable individual-level neural markers of high-level language processing: A necessary precursor for relating neural variability to behavioral and genetic variability. *NeuroImage*, *139*, 74–93.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023, January). *Dissociating language and thought in large language models: a cognitive perspective.* arXiv. (arXiv:2301.06627 [cs])

Makuuchi, M., Bahlmann, J., Anwander, A., & Friederici, A. D. (2009, May). Segregating the core computational faculty of human language from working memory. *Proceedings of the National Academy of Sciences*, *106*(20), 8362–8367. (Publisher: Proceedings of the National Academy of Sciences)

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., … Fedorenko, E. (2022, August). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*(8), 1014–1019. (Number: 8 Publisher: Nature Publishing Group)

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020, June). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 201907367.

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Marcus, G. (2020a). Gpt-2 and the nature of intelligence. *The Gradient*. https://thegradient.pub/gpt2-and-the-nature-of-intelligence/.

Marcus, G. (2020b). *GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about*.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., … Dosenbach, N. U. F. (2022, March). Reproducible brain-wide association studies require thousands of individuals. *Nature*.

Marr, D., & Poggio, T. (1976, May). From Understanding Computation to Understanding Neural Circuitry.
(Accepted: 2004-10-01T20:36:50Z)

Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., … Mehler, J. (1993). The cortical representation of speech. *Journal of Cognitive Neuroscience*, *5*(4), 467–479.

McCarthy, J. (2007, December). From here to human-level AI. *Artificial Intelligence*, *171*(18), 1174–1182.

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020, September). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*. (Publisher: National Academy of Sciences Section: Perspective)

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375–407. (Place: US Publisher: American Psychological Association)

McCulloch, W. S., & Pitts, W. (1943, December). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115–133.

Mesgarani, N., & Chang, E. F. (2012, May). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236. (Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7397 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Auditory system;Neuronal physiology;Perception Subject_term_id: auditory-system;neuronal-physiology;perception)

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014, February). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, *343*(6174), 1006–1010. (Publisher: American Association for the Advancement of Science Section: Report)

Michelmann, S., Kumar, M., Norman, K. A., & Toneva, M. (2023, January). *Large language models can segment narrative events similarly to humans.* arXiv. (arXiv:2301.10297 [cs, q-bio])

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, September). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. (arXiv: 1301.3781)

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates, Inc.

Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., … King, J.-R. (2022, October). Toward a realistic model of speech processing in the brain with self-supervised learning..

Millet, J., Chitoran, I., & Dunbar, E. (2021, November). Predicting non-native speech perception using the perceptual assimilation model and state-of-the-art acoustic models. In *Proceedings of the 25th conference on computational natural language learning* (pp. 661–673). Online: Association for Computational Linguistics.

Millet, J., & Dunbar, E. (2022, May). *Do self-supervised speech models develop human-like perception biases?* arXiv. (arXiv:2205.15819 [cs, eess])

Millet, J., & King, J.-R. (2021). *Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech*.

Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA, USA: MIT Press.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008, May). Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, *320*(5880), 1191–1195.

Mohsenvand, M. N., Izadi, M. R., & Maes, P. (2020, November). Contrastive Representation Learning for Electroencephalogram Classification. In *Proceedings of the Machine Learning for Health NeurIPS Workshop* (pp. 238–253). PMLR. (ISSN: 2640-3498)

Mollica, F., Diachek, E., Mineroff, Z., Kean, H., Siegelman, M., Piantadosi, S. T., … Fedorenko, E. (2019, December). Composition is the core driver of the language-selective network. *bioRxiv*, 436204. (Publisher: Cold Spring Harbor Laboratory Section: New Results)

Mousavi, Z., Kiani, M. M., & Aghajan, H. (2020, January). *Brain signatures of surprise in EEG and MEG data* (Tech. Rep.). (Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article)

Mugler, E. M., Patton, J. L., Flint, R. D., Wright, Z. A., Schuele, S. U., Rosenow, J., … Slutzky, M. W. (2014). Direct classification of all american english phonemes using signals from functional speech motor cortex. *Journal of neural engineering*, *11*(3), 035015.

Narayan, S., Cohen, S. B., & Lapata, M. (2018, August). *Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization.* arXiv. (arXiv:1808.08745 [cs])

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011, May). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009, September). Bayesian reconstruction of natural images from human brain activity. *Neuron*, *63*(6), 902–915.

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., … Hasson, U. (2020, December). *Narratives: fMRI data for evaluating models of naturalistic language comprehension* (preprint). Neuroscience.

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., … Dehaene, S. (2017, May). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, *114*(18), E3669–E3678.

Newman, S. D., Ikuta, T., & Burns, T. (2010, May). The effect of semantic relatedness on syntactic analysis: An fMRI study. *Brain and Language*, *113*(2), 51–58.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2020, May). Adversarial NLI: A New Benchmark for Natural Language Understanding. *arXiv:1910.14599 [cs].* (arXiv: 1910.14599)

Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., … Xiong, C. (2022, September). *CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis.* arXiv. (arXiv:2203.13474 [cs])

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

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

Obleser, J., Zimmermann, J., Van Meter, J., & Rauschecker, J. P. (2007, October). Multiple stages of auditory speech perception reflected in event-related FMRI. *Cerebral Cortex (New York, N.Y.: 1991)*, *17*(10), 2251–2257.

Okada, K., Matchin, W., & Hickok, G. (2018, February). Neural evidence for predictive coding in auditory cortex during speech production. *Psychonomic Bulletin & Review*, *25*(1), 423–430.

Omelianchuk, K., Atrasevych, V., Chernodub, A., & Skurzhanskyi, O. (2020, May). GECToR – Grammatical Error Correction: Tag, Not Rewrite. *arXiv:2005.12592 [cs]*. (arXiv: 2005.12592)

Oord, A. v. d., Vinyals, O., & Kavukcuoglu, K. (2018, May). *Neural Discrete Representation Learning.* arXiv. (arXiv:1711.00937 [cs])

Oota, S. R., Manwani, N., & S, B. R. (2018, June). fMRI Semantic Category Decoding using Linguistic Encoding of Word Embeddings. *arXiv:1806.05177 [cs, q-bio]*. (arXiv: 1806.05177)

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., … Lowe, R. (2022, March). *Training language models to follow instructions with human feedback.* arXiv. (arXiv:2203.02155 [cs])

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011, February). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, *108*(6), 2522–2527.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5206–5210).

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, July). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Pasad, A., Chou, J.-C., & Livescu, K. (2021). Layer-wise analysis of a self-supervised speech representation model. *arXiv preprint arXiv:2107.04734*.

Pasquiou, A., Lakretz, Y., Hale, J. T., Thirion, B., & Pallier, C. (2022, June). Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 17499–17516). PMLR. (ISSN: 2640-3498)

Pasquiou, A., Lakretz, Y., Thirion, B., & Pallier, C. (2023, 02). Information-restricted neural language models reveal different brain regions' sensitivity to semantics, syntax and context.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007, December). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987. (Bandiera_abtest: a Cg_type: Nature Research Journals Number: 12 Primary_atype: Reviews Publisher: Nature Publishing Group)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Pelletier, F. J. (1994). The principle of semantic compositionality. *Topoi*, *13*(1), 11–24.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., ... Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, *9*(1), 1–13.

Petkov, C. I., Kikuchi, Y., Milne, A. E., Mishkin, M., Rauschecker, J. P., & Logothetis, N. K. (2015, January). Different forms of effective connectivity in primate frontotemporal pathways. *Nature Communications*, *6*(1), 6000. (Number: 1 Publisher: Nature Publishing Group)

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

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of Functional MRI Data Analysis*. Cambridge: Cambridge University Press.

Price, C. J. (2010). The anatomy of language: a review of 100 fmri studies published in 2009. *Annals of the new York Academy of Sciences*, *1191*(1), 62–88.

Qian, P., Qiu, X., & Huang, X. (2016, April). Bridging LSTM Architecture and the Neural Dynamics during Reading. *arXiv:1604.06635 [cs]*. (arXiv: 1604.06635)

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021, February). Learning Transferable Visual Models From Natural Language Supervision. *arXiv:2103.00020 [cs]*. (arXiv: 2103.00020)

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. , 24.

Rae, J. W., Potapenko, A., Jayakumar, S. M., & Lillicrap, T. P. (2019, November). Compressive Transformers for Long-Range Sequence Modelling. *arXiv:1911.05507 [cs, stat]*. (arXiv: 1911.05507)

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … Liu, P. J. (2020, July). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683 [cs, stat]*. (arXiv: 1910.10683)

Raikote, P. (2021, June). *Expire-Span: Not All Memories are Created Equal explained.*

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016, November). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383–2392). Austin, Texas: Association for Computational Linguistics.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022, April). *Hierarchical Text-Conditional Image Generation with CLIP Latents.* arXiv. (arXiv:2204.06125 [cs])

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., … Sutskever, I. (2021, February). Zero-Shot Text-to-Image Generation. *arXiv:2102.12092 [cs]*. (arXiv: 2102.12092)

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., … Hochreiter, S. (2021, April). Hopfield Networks is All You Need. *arXiv:2008.02217 [cs, stat]*. (arXiv: 2008.02217)

Rao, R. P. N., & Ballard, D. H. (1999, January). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. (Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group)

Reddy, A. J., & Wehbe, L. (2020, June). *Syntactic representations in the human brain: beyond effort-based metrics* (preprint). Neuroscience.

Richard, H., Martin, L., Pinho, A. L., Pillow, J., & Thirion, B. (2019, December). *Fast shared response model for fMRI data.* arXiv. (arXiv:1909.12537 [cs, eess, q-bio])

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., … Kording, K. P. (2019, November). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770.

Rolfe, J. T. (2017, April). *Discrete Variational Autoencoders.* arXiv. (arXiv:1609.02200 [cs, stat])

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022, April). *High-Resolution Image Synthesis with Latent Diffusion Models.* arXiv. (arXiv:2112.10752 [cs])

Ruan, Y.-P., Ling, Z.-H., & Hu, Y. (2016). Exploring Semantic Representation in Brain Activity Using Word Embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 669–679). Austin, Texas: Association for Computational Linguistics.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986, October). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. (Number: 6088 Publisher: Nature Publishing Group)

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*(1), 60–94. (Place: US Publisher: American Psychological Association)

Sabri, M., Binder, J. R., Desai, R., Medler, D. A., Leitl, M. D., & Liebenthal, E. (2008, February). Attentional and linguistic interactions in speech perception. *NeuroImage*, *39*(3), 1444–1456.

Santi, A., & Grodzinsky, Y. (2010, July). fMRI adaptation dissociates syntactic complexity dimensions. *NeuroImage*, *51*(4), 1285–1293.

Sassenhagen, J., & Fiebach, C. J. (2019, April). *Traces of Meaning Itself: Encoding distributional word vectors in brain activity* (preprint). Neuroscience.

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., … Batra, D. (2019, November). *Habitat: A Platform for Embodied AI Research.* arXiv. (arXiv:1904.01201 [cs])

Saxe, A., Nelli, S., & Summerfield, C. (2021, January). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, *22*(1), 55–67. (Number: 1 Publisher: Nature Publishing Group)

Schatz, T. (2016). *Abx-discriminability measures and applications* (Unpublished doctoral dissertation). Université Paris 6 (UPMC).

Schmidhuber, J. (2015, January). Deep Learning in Neural Networks: An Overview. *Neural Networks*, *61*, 85–117. (arXiv:1404.7828 [cs])

Schoffelen, J.-M., Oostenveld, R., Lam, N. H. L., Uddén, J., Hultén, A., & Hagoort, P. (2019, April). A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data*, *6*(1), 17. (Number: 1 Publisher: Nature Publishing Group)

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., … Fedorenko, E. (2021, November). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. (Publisher: Proceedings of the National Academy of Sciences)

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., … DiCarlo, J. J. (2018, September). *Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?* (preprint). Neuroscience.

Schwartz, D., Toneva, M., & Wehbe, L. (2019). Inducing brain-relevant bias in natural language processing models. , 11.

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th python in science conference.*

Searle, J. (2009, August). Chinese room argument. *Scholarpedia*, *4*(8), 3100.

Seghier, M. L., & Price, C. J. (2018). Interpreting and utilising intersubject variability in brain function. *Trends in Cognitive Sciences*, *22*, 517–530. (Place: Netherlands Publisher: Elsevier Science)

Sennrich, R., Haddow, B., & Birch, A. (2016, June). Neural Machine Translation of Rare Words with Subword Units. *arXiv:1508.07909 [cs]*. (arXiv: 1508.07909)

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2014, February). *OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks.* arXiv. (arXiv:1312.6229 [cs])

Seydell-Greenwald, A., Wang, X., Newport, E., Bi, Y., & Striem-Amit, E. (2020). Spoken language comprehension activates the primary visual cortex. *bioRxiv*.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020, February). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307.

Shain, C., Kean, H., Lipkin, B., Affourtit, J., Siegelman, M., Mollica, F., & Fedorenko, E. (2021). 'constituent length' effects in fmri do not provide evidence for abstract syntactic processing. *bioRxiv*.

Shallice, T., & Burgess, P. (1991, May). Deficits in strategy application following frontal lobe damage in man. *Brain : a journal of neurology*, *114 (Pt 2)*, 727–41.

Shamma, S., Patel, P., Mukherjee, S., Marion, G., Khalighinejad, B., Han, C., … Mesgarani, N. (2021). Learning speech production and perception through sensorimotor interactions. *Cerebral cortex communications*, *2*(1), tgaa091.

Shen, T., Lei, T., Barzilay, R., & Jaakkola, T. (2017, November). *Style Transfer from Non-Parallel Text by Cross-Alignment.* arXiv. (arXiv:1705.09655 [cs])

Sinha, K., Sodhani, S., Dong, J., Pineau, J., & Hamilton, W. L. (2019, November). CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text. 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)* (pp. 4506–4515). Hong Kong, China: Association for Computational Linguistics.

Smolensky, P. (1990, November). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1), 159–216.

Solaiman, I., & Dennison, C. (2021). Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., … Wu, Z. (2022, June). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.* arXiv. (arXiv:2206.04615 [cs, stat])

Stehwien, S., Henke, L., Hale, J., Brennan, J., & Meyer, L. (2020). The little prince in 26 languages: Towards a multilingual neuro-cognitive corpus. In *Proceedings of the second workshop on linguistic and neurocognitive resources* (pp. 43–49).

Stephenson, C., Feather, J., Padhy, S., Elibol, O., Tang, H., McDermott, J., & Chung, S. (2019). Untangling in invariant speech recognition. *Advances in neural information processing systems*, *32*.

Sun, J., Wang, S., Zhang, J., & Zong, C. (2021, February). Neural Encoding and Decoding With Distributed Sentence Representations. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 589–603.

Szabó, Z. G. (2004). Compositionality.

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015, June). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–9). Boston, MA, USA: IEEE.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., … Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, *115*(35), 8835–8840.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D. d. L., … Riedmiller, M. (2018, January). *DeepMind Control Suite.* arXiv. (arXiv:1801.00690 [cs])

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021, September). BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *arXiv:2104.08663 [cs]*. (arXiv: 2104.08663)

Thomas, A. W., Ré, C., & Poldrack, R. A. (2022, June). *Self-Supervised Learning Of Brain Dynamics From Broad Neuroimaging Data.* arXiv. (arXiv:2206.11417 [q-bio])

Thompson, J. A., Bengio, Y., Formisano, E., & Schönwiesner, M. (2021, January). *Training neural networks to recognize speech increased their correspondence to the human auditory pathway but did not yield a shared hierarchy of acoustic features* (preprint). Neuroscience.

Thual, A., Tran, H., Zemskova, T., Courty, N., Flamary, R., Dehaene, S., & Thirion, B. (2022, November). *Aligning individual brains with Fused Unbalanced Gromov-Wasserstein.* arXiv. (arXiv:2206.09398 [q-bio, stat])

Toneva, M., Mitchell, T. M., & Wehbe, L. (2020a, November). Combining computational controls with natural text reveals new aspects of meaning composition. *bioRxiv*, 2020.09.28.316935. (Publisher: Cold Spring Harbor Laboratory Section: New Results)

Toneva, M., Mitchell, T. M., & Wehbe, L. (2020b, September). *The meaning that emerges from combining words is robustly localizable in space but not in time* (preprint). Neuroscience.

Toneva, M., Stretcu, O., Poczos, B., Wehbe, L., & Mitchell, T. M. (2020, November). Modeling Task Effects on Meaning Representation in the Brain via Zero-Shot MEG Prediction. *arXiv:2009.08424 [cs]*. (arXiv: 2009.08424)

Toneva, M., & Wehbe, L. (2019, November). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). (arXiv: 1905.11833)

Tour, T. D. l., Eickenberg, M., & Gallant, J. (2022, May). *Feature-space selection with banded ridge regression.* bioRxiv. (Pages: 2022.05.05.490831 Section: New Results)

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., … Lample, G. (2023, February). *LLaMA: Open and Efficient Foundation Language Models.* arXiv. (arXiv:2302.13971 [cs])

Tremblay, P., & Dick, A. S. (2016, November). Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *Brain and Language*, *162*, 60–71.

Turing, A. M. (1950, October). I.—Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460. (_eprint: https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf)

Turing, A. M. (2009). Computing machinery and intelligence. In *Parsing the turing test* (pp. 23–65). Springer.

Ullman, M. T. (2001, October). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience*, *2*(10), 717–726. (Number: 10 Publisher: Nature Publishing Group)

Vaidya, A. R., Jain, S., & Huth, A. G. (2022, May). *Self-supervised models of audio effectively explain human cortical responses to speech.* arXiv. (arXiv:2205.14252 [cs])

Van Essen, D. C. (2005). A population-average, landmark-and surface-based (pals) atlas of human cerebral cortex. *Neuroimage*, *28*(3), 635–662.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.

Vidal, Y., Brusini, P., Bonfieni, M., Mehler, J., & Bekinschtein, T. A. (2019, September). Neural Signal to Violations of Abstract Rules Using Speech-Like Stimuli. *eNeuro*, *6*(5). (Publisher: Society for Neuroscience Section: New Research)

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., & Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences*, *108*(51), 20754–20759.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., … Bowman, S. R. (2020, February). *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.* arXiv. (arXiv:1905.00537 [cs])

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 353–355). Brussels, Belgium: Association for Computational Linguistics.

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., … Dupoux, E. (2021). Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Wang, L. (2021). Dynamic predictive coding across the left fronto-temporal language hierarchy: Evidence from MEG, EEG and fMRI. , 29.

Wang, S., Zhang, J., Lin, N., & Zong, C. (2020, April). Probing Brain Activation Patterns by Dissociating Semantics and Syntax in Sentences. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 9201–9208.

Wang, S., Zhang, J., Wang, H., Lin, N., & Zong, C. (2020, January). Fine-grained neural decoding with distributed word representations. *Information Sciences*, *507*, 256–272.

Warstadt, A., & Bowman, S. R. (2022, August). *What Artificial Neural Networks Can Tell Us About Human Language Acquisition.* arXiv. (arXiv:2208.07998 [cs])

Weerts, L., Rosen, S., Clopath, C., & Goodman, D. F. (2021). The psychometrics of automatic speech recognition. *bioRxiv*.

Wegelin, J. A., Packer, A., & Richardson, T. S. (2006, January). Latent models for cross-covariance. *Journal of Multivariate Analysis*, *97*(1), 79–102.

Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014, October). 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)* (pp. 233–243). Doha, Qatar: Association for Computational Linguistics.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016, June). Prediction During Natural Language Comprehension. *Cerebral Cortex*, *26*(6), 2506–2516.

Wiseman, S., Shieber, S. M., & Rush, A. M. (2017, July). Challenges in Data-to-Document Generation. *arXiv:1707.08052 [cs]*. (arXiv: 1707.08052)

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. M. (2020, July). *HuggingFace's Transformers: State-of-the-art Natural Language Processing.* arXiv. (arXiv:1910.03771 [cs])

Woolnough, O., Donos, C., Rollo, P. S., Forseth, K. J., Lakretz, Y., Crone, N. E., … Tandon, N. (2020). Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *bioRxiv*.

Xiao, T., Hong, J., & Ma, J. (2018, March). *DNA-GAN: Learning Disentangled Representations from Multi-Attribute Images.* arXiv. (arXiv:1711.05415 [cs])

Xu, Q., Baevski, A., Likhomanenko, T., Tomasello, P., Conneau, A., Collobert, R., … Auli, M. (2020, October). *Self-training and Pre-training are Complementary for Speech Recognition.* arXiv. (arXiv:2010.11430 [cs, eess])

Yamins, D. L., & DiCarlo, J. J. (2016, April). Eight open questions in the computational modeling of higher sensory cortex. *Current Opinion in Neurobiology*, *37*, 114–120.

Yamins, D. L. K., & DiCarlo, J. J. (2016, March). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014, June). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020, January). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237 [cs]*. (arXiv: 1906.08237)

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., & Manning, C. D. (2018, September). *HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering.* arXiv. (arXiv:1809.09600 [cs])

Zador, A., Richards, B., Ölveczky, B., Escola, S., Bengio, Y., Boahen, K., … Tsao, D. (2022, October). *Toward Next-Generation Artificial Intelligence: Catalyzing the NeuroAI Revolution.* arXiv. (arXiv:2210.08340 [cs, q-bio])

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019, May). HellaSwag: Can a Machine Really Finish Your Sentence? *arXiv:1905.07830 [cs]*. (arXiv: 1905.07830)

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020, February). *BERTScore: Evaluating Text Generation with BERT.* arXiv. (arXiv:1904.09675 [cs])

Zhang, Y., Li, Z., & Min, Z. (2020). Efficient second-order TreeCRF for neural dependency parsing. In *Proceedings of acl* (pp. 3295–3305). Retrieved from https://www.aclweb.org/anthology/2020.acl-main.302