Disentangling Syntax and Semantics in the Brain with Deep Networks

Charlotte Caucheteux, Alexandre Gramfort, Jean-Remi King

Disentangling Syntax and Semantics in the Brain with Deep Networks

Charlotte Caucheteux [1 2] **Alexandre Gramfort** [1] **Jean-Remi King** [2 3]

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 ~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.

1. 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).

[1]Inria, Saclay, France [2]Facebook AI Research, Paris, France [3]École normale supérieure, PSL University, CNRS, Paris, France. Correspondence to: Charlotte Caucheteux <ccaucheteux@fb.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

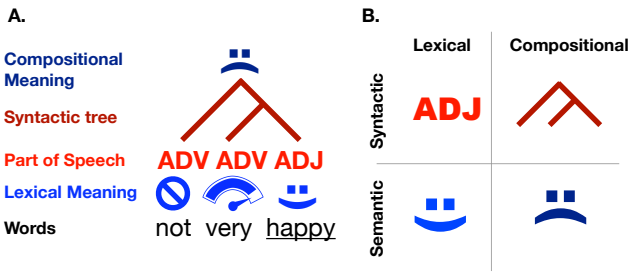

Figure 1. **Taxonomy A.** To understand the meaning of a phrase, one must combine the meaning of each word using the rules of syntax. For example, the meaning of the phrase NOT VERY HAPPY is (roughly) SAD, and can be found by recursively combining the two adverbs and the adjective. **B.** Here, we aim 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).

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., 2020; Caucheteux & King, 2020; Goldstein et al., 2021). 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, 2020; Caucheteux et al., 2021).

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.

[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.

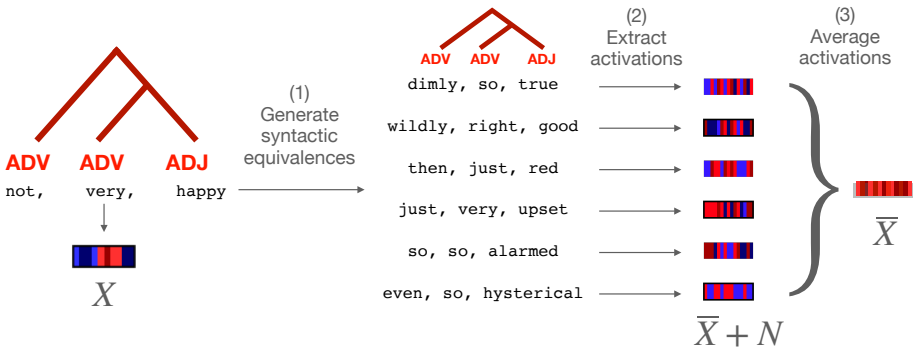

Figure 2. **Method to isolate syntactic representations in GPT-2's word and compositional embeddings.** To isolate the syntactic representations of a sequence of words e.g. w = NOT VERY HAPPY, we (1) synthesize sentences with the same syntactic structure as w (e.g DIMLY SO TRUE, *etc.*), then (2) extract 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.

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 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 et al., 2016; Nelson et al., 2017; 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., 2020; Toneva & Wehbe, 2019; Abnar et al., 2019; Reddy & Wehbe, 2020) (although see (Caucheteux & King, 2020)). 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).

2. 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 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, 2020). 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).

Finally, even though a variety of meaningful features are captured by both word embeddings (Mikolov 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 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*.

3. Methods

3.1. Isolating Syntactic Representations

We introduce below a method to isolate distributed representations of syntax in neural networks. We assume that a system Ψ ($\Psi : \mathcal{V}^M \to \mathbb{R}^{d\times M}$, $\mathcal{V}$ 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 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:

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

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

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

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 2).

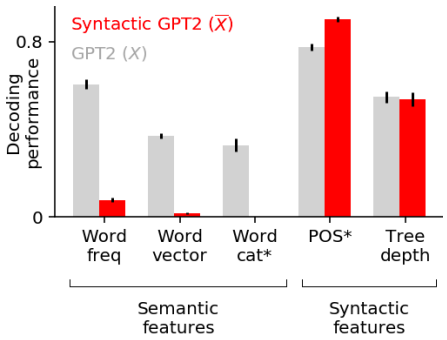

Figure 3. **Semantic and syntactic information encoded in $\overline{X}$.** To check that the syntactic embeddings $\overline{X}$ only contain syntactic information, we train a ℓ_2-regularized linear model to predict three semantic features (frequency, word embeddings and semantic category of content 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 C). 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.

3.2. 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

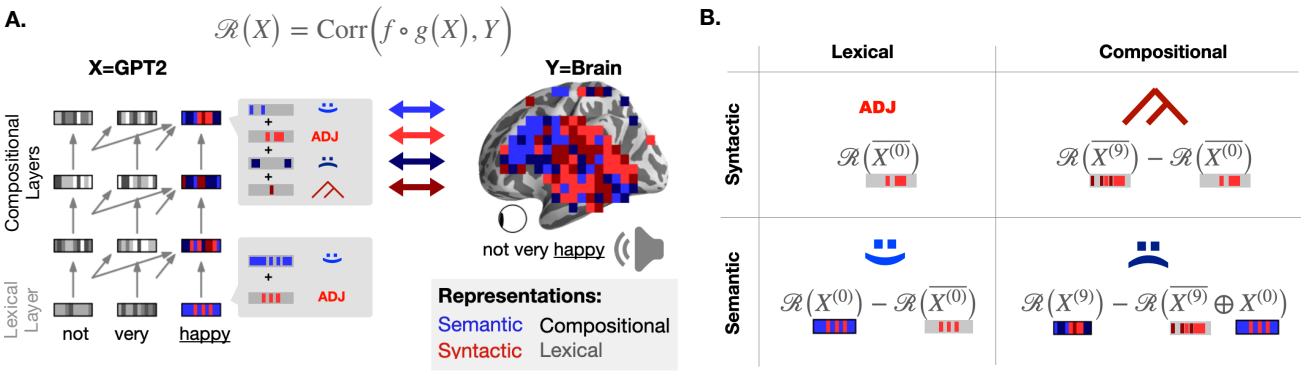

Figure 4. **Method to decompose the language representations shared between brains and deep language models A.** The human brain and modern language models like GPT-2 both generate *distributed* representations, which are thus difficult to link with the *symbolic* properties of linguistic theories. We 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 2); $\oplus$ indicates a feature concatenation, and '$-$' indicates a subtraction between scores.

$X = \Psi_1(w) \in \mathbb{R}^{M \times d}$ 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 et al., 2016; 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 $I \subset [1 \dots N]$:

$$\mathcal{R}(X) : f \mapsto \mathcal{L}\left(f \circ g(X)_{i \in I}, \overline{(Y_i)}_{i \in I}\right) \qquad (1)$$

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

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:

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

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

$$\mathcal{R} = \text{corr}\left(f \circ g(X), \overline{Y}\right) \ . \qquad (2)$$

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

3.3. Decomposing Shared Activations between Brains and Neural Language Models

Here, we use the definitions and methods introduced in Section 2, 3.1 and 3.2 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.

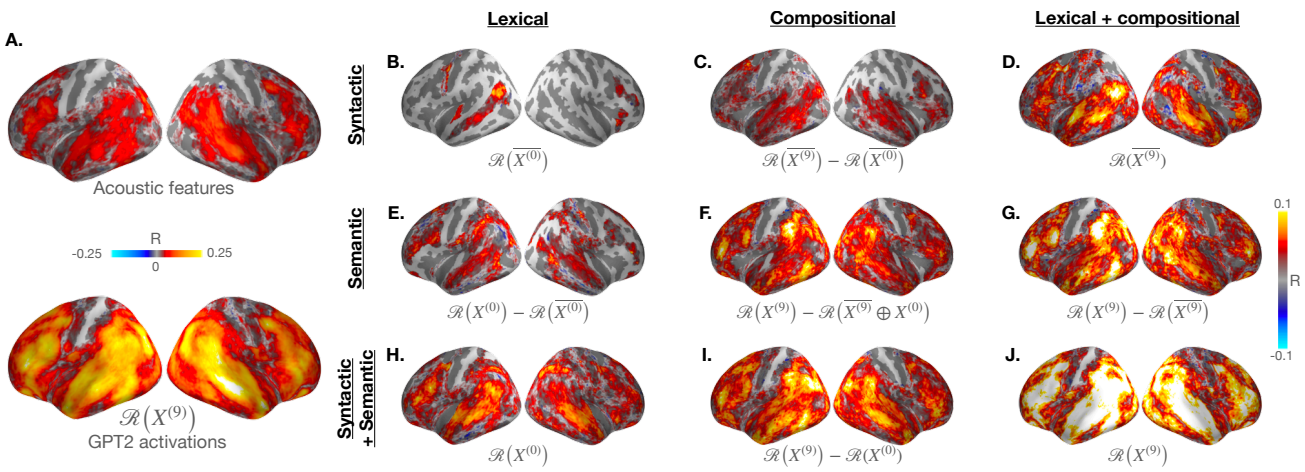

Figure 5. **Results** Decomposition of the brain scores of 345 subjects listening to narratives into their phonological (A) syntactic (B-D), semantic (E-G), lexical (B-H), compositional (C-I) components and their combinations (ten combinations in total). **A** Comparison between the brain scores of three 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 E).

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,

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

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. Following the definitions of Section 2, 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, Appendix C). 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 3.2, 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 $l \in [0 \dots L]$

4. Experiments on the Narratives fMRI Dataset

Here, we apply the general method described in Section 3.1, 3.2 and 3.3 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 $\approx$ 3 minutes to $\approx$ 56 minutes, for a total of $\approx$ 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 $\approx$ 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., 2021; Caucheteux & King, 2020; Affolter et al., 2020; Schrimpf et al., 2020; Caucheteux et al., 2021). 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 A). The activations of each GPT-2 layer are extracted, resulting in 12 vectors of 768 activations for 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 A).

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

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 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 B). 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 A).
- Then, we use Supar, a state-of-the art dependency parser (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 (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 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 5.

5. Experimental 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 5A and 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 5, 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 displayed in figures 5 and 6 (non-significant scores after correction for multiple comparisons across regions are masked).

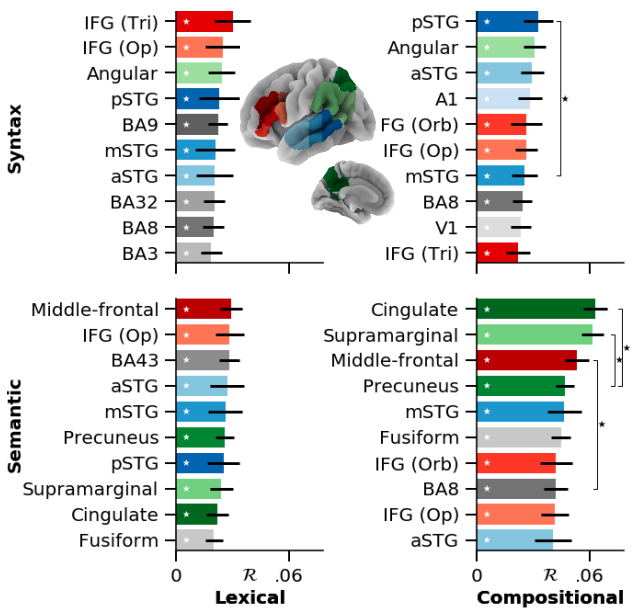

Figure 6. Same as Figure 5.BCEF, with voxel-averaged brain scores (after subtraction of phonological brain scores), for the top ten regions of interest of the left hemisphere (Appendix E). Error bars are the standard-errors of the mean across the 100 cross-validation folds. Significance ('*') is assessed with a Wilcoxon test across folds, with $p < .05$ as a threshold.

Lexical features. The lexical representations of the brain have been repeatedly investigated through the lens of a word-embedding (Mitchell et al., 2008; Huth et al., 2016; Toneva & Wehbe, 2019; Schrimpf et al., 2020; Caucheteux & King, 2020). 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 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 5B). The results reveal significant 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 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 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 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, 2020). We replicate this result with a representative contextual layer of GPT-2 (layer 9 out of 12, Figure 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 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 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 5E and 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 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

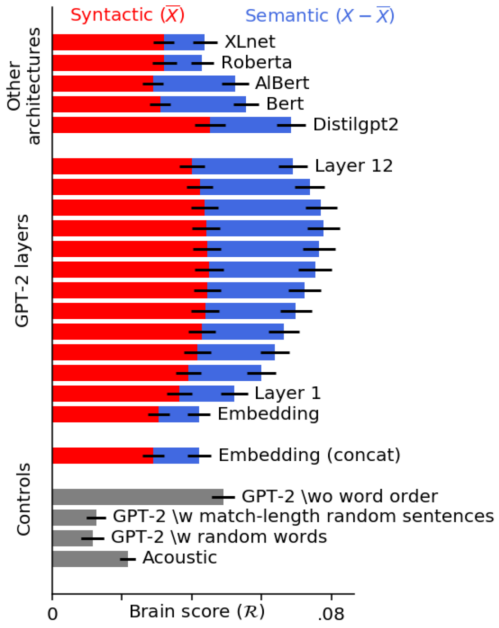

Figure 7. **Generalisation to other layers and architectures** In red, the brain scores of the syntactic embeddings ($\mathcal{R}(\bar{X})$) built out of GPT-2 layers (from the word embedding to layer 12), and the middle layer of five transformer architectures (top, cf. Appendix A, $l = 2/3 \times n_{\text{layers}}$). 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 F.)

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 (Schrimpf et al., 2020; Caucheteux et al., 2021), 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 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 A).

6. 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., 2019; 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 et al., 2016; Brennan, 2016; Brennan & Hale, 2019; Stehwien et al., 2020; Gwilliams et al., 2020). While we replicate previous neuroscientific findings regarding lexical semantics (Figure 5E) (Huth et al., 2016) and lexical *vs* compositional processing in the brain (Figure 5.H,J) (Toneva & Wehbe, 2019; Schrimpf et al., 2020; Goldstein et al., 2021), our systematic decomposition of language representations brings new light on the brain bases of syntax (Figure 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, 2020). 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, 2020; Schrimpf et al., 2020; Heilbron et al., 2020). 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; 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 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.

7. Acknowledgement

This work was supported by ANR-17-EURE-0017, the Fyssen Foundation and the Bettencourt Foundation to JRK for his work at PSL.

References

Abnar, S., Beinborn, L., Choenni, R., and Zuidema, W. Blackbox meets blackbox: Representational Similarity and Stability Analysis of Neural Language Models and

Brains. *arXiv:1906.01539 [cs, q-bio]*, June 2019. arXiv: 1906.01539.

Affolter, N., Egressy, B., Pascual, D., and Wattenhofer, R. Brain2word: Decoding brain activity for language generation. *arXiv preprint arXiv:2009.04765*, 2020.

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

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., and Desai, R. H. Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33(3-4): 130–174, May 2016. ISSN 0264-3294. doi: 10.1080/02643294.2016.1147426. Publisher: Routledge _eprint: https://doi.org/10.1080/02643294.2016.1147426.

Brennan, J. Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10(7):299–313, 2016.

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

Brennan, J. R. and Pylkkänen, L. Meg evidence for incremental sentence composition in the anterior temporal lobe. *Cognitive science*, 41:1515–1531, 2017.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*, July 2020. arXiv: 2005.14165.

Caucheteux, C. and King, J.-R. Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv*, 2020.

Caucheteux, C., Gramfort, A., and King, J.-R. Gpt-2's activations predict the degree of semantic comprehension in the human brain. *bioRxiv*, 2021.

Chomsky, N. *The minimalist program*. MIT press, 2014.

Dehaene, S. and Cohen, L. The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15(6):254–262, June 2011. ISSN 1879-307X. doi: 10.1016/j.tics.2011.04.003.

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., and Gallant, J. L. 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, September 2019. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0675-19.2019. Publisher: Society for Neuroscience Section: Research Articles.

Destrieux, C., Fischl, B., Dale, A., and Halgren, E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15, October 2010. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2010.06.010.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. arXiv: 1810.04805.

Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, 19(1):158–164, 2016.

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., and Kanwisher, N. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104(2): 1177–1194, 2010.

Fedorenko, E., Nieto-Castanon, A., and Kanwisher, N. Lexical and syntactic representations in the brain: an fmri investigation with multi-voxel pattern analyses. *Neuropsychologia*, 50(4):499–513, 2012.

Friederici, A. D. The Brain Basis of Language Processing: From Structure to Function. *Physiological Reviews*, 91 (4):1357–1392, October 2011. ISSN 0031-9333, 1522-1210. doi: 10.1152/physrev.00006.2011.

Friederici, A. D., Opitz, B., and Von Cramon, D. Y. Segregating semantic and syntactic aspects of processing in the human brain: an fmri investigation of different word types. *Cerebral cortex*, 10(7):698–705, 2000.

Givón, T. *Syntax: an introduction*, volume 1. John Benjamins Publishing, 2001.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., et al. Thinking ahead: prediction in context as a keystone of language in humans and machines. *bioRxiv*, pp. 2020–12, 2021.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. MEG and EEG data

analysis with MNE-Python. *Frontiers in Neuroscience*, 7, 2013. ISSN 1662-453X. doi: 10.3389/fnins.2013.00267. Publisher: Frontiers.

Gwilliams, L., King, J.-R., Marantz, A., and Poeppel, D. Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. *bioRxiv*, 2020.

Hale, J., Dyer, C., Kuncoro, A., and Brennan, J. R. Finding syntax in human encephalography with beam search. *arXiv preprint arXiv:1806.04127*, 2018.

Hamilton, L. S. and Huth, A. G. The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 35(5):573–582, 2020.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., and de Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv*, 2020.

Hickok, G. and Poeppel, D. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5): 393–402, May 2007. ISSN 1471-0048. doi: 10.1038/nrn2113. Number: 5 Publisher: Nature Publishing Group.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL https://doi.org/10.5281/zenodo.1212303.

Hupkes, D., Dankers, V., Mul, M., and Bruni, E. The compositionality of neural networks: integrating symbolism and connectionism. *arXiv preprint arXiv:1908.08351*, 2019.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, April 2016. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature17637.

Jackendoff, R. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, 2002. ISBN 978-0-19-171325-5. Publication Title: Foundations of Language.

Jain, S. and Huth, A. G. Incorporating context into language encoding models for fmri. *BioRxiv*, pp. 327601, 2018.

Jat, S., Tang, H., Talukdar, P., and Mitchell, T. Relating simple sentence representations in deep neural networks and the brain. *arXiv preprint arXiv:1906.11861*, 2019.

Jawahar, G., Sagot, B., and Seddah, D. 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, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1356.

King, J.-R., Gwilliams, L., Holdgraf, C., Sassenhagen, J., Barachant, A., Engemann, D., Larson, E., and Gramfort, A. Encoding and decoding neuronal dynamics: Methodological framework to uncover the algorithms of cognition. *The Cognitive Neurosciences, Sixth Edition*, 2018.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008. Publisher: Frontiers.

Kutas, M. and Hillyard, S. A. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, 1980.

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., and Baroni, M. The emergence of number and syntax units in LSTM language models. *arXiv:1903.07435 [cs]*, April 2019. arXiv: 1903.07435.

Lakretz, Y., Dehaene, S., and King, J.-R. What limits our capacity to process nested long-range dependencies in sentence comprehension? *Entropy*, 22(4):446, 2020.

Lample, G. and Conneau, A. Cross-lingual Language Model Pretraining. *arXiv:1901.07291 [cs]*, January 2019. arXiv: 1901.07291.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942 [cs]*, February 2020. arXiv: 1909.11942.

Linzen, T. and Baroni, M. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7, 2020.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*, July 2019. arXiv: 1907.11692.

Lycan, W. G. *Philosophy of language: A contemporary introduction*. Routledge, 2018.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., and Levy, O. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, pp. 201907367, June 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1907367117.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008.

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C. J., Yeshurun, Y., Regev, M., Nguyen, M., Chang, C. H. C., Baldassano, C., Lositsky, O., Simony, E., Chow, M. A., Leong, Y. C., Brooks, P. P., Micciche, E., Choe, G., Goldstein, A., Vanderwal, T., Halchenko, Y. O., Norman, K. A., and Hasson, U. Narratives: fMRI data for evaluating models of naturalistic language comprehension. preprint, Neuroscience, December 2020.

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C., and Dehaene, S. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114 (18):E3669–E3678, May 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1701590114.

Omelianchuk, K., Atrasevych, V., Chernodub, A., and Skurzhanskyi, O. GECToR – Grammatical Error Correction: Tag, Not Rewrite. *arXiv:2005.12592 [cs]*, May 2020. arXiv: 2005.12592.

Pallier, C., Devauchelle, A.-D., and Dehaene, S. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108 (6):2522–2527, 2011.

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Reddy, A. J. and Wehbe, L. Syntactic representations in the human brain: beyond effort-based metrics. *bioRxiv*, 2020.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N. G., Tenenbaum, J. B., and Fedorenko, E. Artificial Neural Networks Accurately Predict Language Processing in the Brain. *bioRxiv*, pp. 2020.06.26.174482, June 2020. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Sennrich, R., Haddow, B., and Birch, A. Neural Machine Translation of Rare Words with Subword Units. *arXiv:1508.07909 [cs]*, June 2016. arXiv: 1508.07909.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., and Fedorenko, E. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307, February 2020. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2019.107307.

Smolensky, P. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1):159–216, November 1990. ISSN 0004-3702. doi: 10.1016/0004-3702(90)90007-M.

Stehwien, S., Henke, L., Hale, J., Brennan, J., and Meyer, L. 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, 2020.

Szabó, Z. G. Compositionality. 2004.

Toneva, M. and Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *arXiv preprint arXiv:1905.11833*, 2019.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

Yamins, D. L. and DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237 [cs]*, January 2020. arXiv: 1906.08237.

Zhang, Y., Li, Z., and Min, Z. Efficient second-order TreeCRF for neural dependency parsing. In *Proceedings of ACL*, pp. 3295–3305, 2020. URL https://www.aclweb.org/anthology/2020.acl-main.302.

Appendix

A. Deep Neural Networks' Activations

Pre-trained tansformers In Section 4, 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 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, 2020) 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.

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

In Section 3.1 and 3.3, 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 8 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.

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

In Section 3.3 and Figure 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 $\overline{X}$.

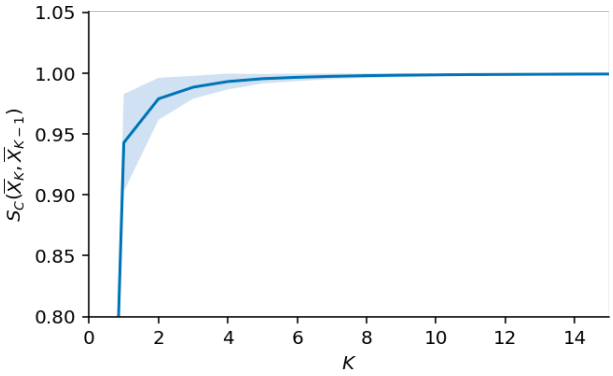

Figure 8. **Convergence of the method to build syntactic embeddings.** Cosine similarity S_C between the syntactic component $\overline{X}$ of GPT-2 activations induced by a sequence w, when computed with K and $K-1$ syntactically equivalent sequences. The syntactic embeddings $\overline{X}_K$ and $\overline{X}_{K-1}$ 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.

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 (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, continuous feature). We use the 'zipf_frequency' from the wordfreq[4] python library.
- 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[5]. Categories are not available

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

[5] 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, miscel-

for all the 2,800 Wikipedia words studied here. Thus, we first train a linear model (scikit-learn 'RidgeCV-Classifier') 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, 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 (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 (grey bars).

D. Temporal Alignment g between X and Y

In Section 3.2, 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. 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.

E. Brain Parcellation

In Figure 6, brain scores are averaged across voxels within regions of interest using the Brodmann's areas from the PALS parcellation of freesurfer[6]. To gain in precision, we split the superior temporal gyrus (BA22) into its anterior, middle and posterior parts. In Figure 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

laneous, multimodal, object, part, perceptual, period, physical, place, plant, property, social, somatosensory, sound, spatial, state, temporal, time, tool, vehicle, visual, weather

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

F. Control for Low-level Linguistic Features

In Section 5 and Figure 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 4. R scores are evaluated for each subject and reported in Figure 7.