Disentangling Syntax and Semantics in the Brain with Deep Networks

Charlotte Caucheteux Alexandre Gramfort Jean-Remi King

ICML 2021

Motivation

- The activations of language transformers like GPT-2 have been shown to ***linearly*** map onto brain activity during speech comprehension.
- The nature of these activations remains largely unknown and presumably conflate distinct linguistic classes.
  - Symbols vs Vectors
  - Why ***linearly*** mapping?
  - Distractions
- The paper propose a taxonomy to factorize the high-dimensional activations of language models into four combinatorial classes: lexical, compositional, syntactic, and semantic representations.

What's linear mapping?

- When presented with the same sentences, the activations of language transformers tend to ***linearly*** map onto those of the human brain.
  $Y=f(X)$

Five definitions to decompose representations

- ***Representation***

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

In this view, a system *Ψ1* is said to share the representation of a system *Ψ2* if there exists a linear mapping from *X* to *Y* , where *X = Ψ1(w)* and *Y = Ψ2(w)* are the activations elicited by the words w in each system.

Five definitions to decompose representations

- ***Lexical***

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.

Five definitions to decompose representations

- ***Compositional***

"Contextualized" representations generated by a system combining multiples words: $\Psi(w_1 \dots w_M)$

We restrict the term "compositional" to its strict sense: to the set of representations that cannot be accounted for by lexical representations, and thus by a linear combination of word-embeddings.

Five definitions to decompose representations

- ***Syntactic***

The set of representations associated with the structure of sentences independently of their meaning (e.g. part-of-speech, dependency and constituency trees).

We extract these syntactic representations from the average activations elicited by a set of synthetic sentences that share the same syntactic properties.

Syntactic representation

- ***Isolating Syntactic Representation***

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.

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.).

Syntactic representation

- ***Isolating Syntactic Representation***

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.

Syntactic representation

- ***Isolating Syntactic Representation***

Thus, $\overline{\psi}$ (the syntactic representation of w) can be approximated through:

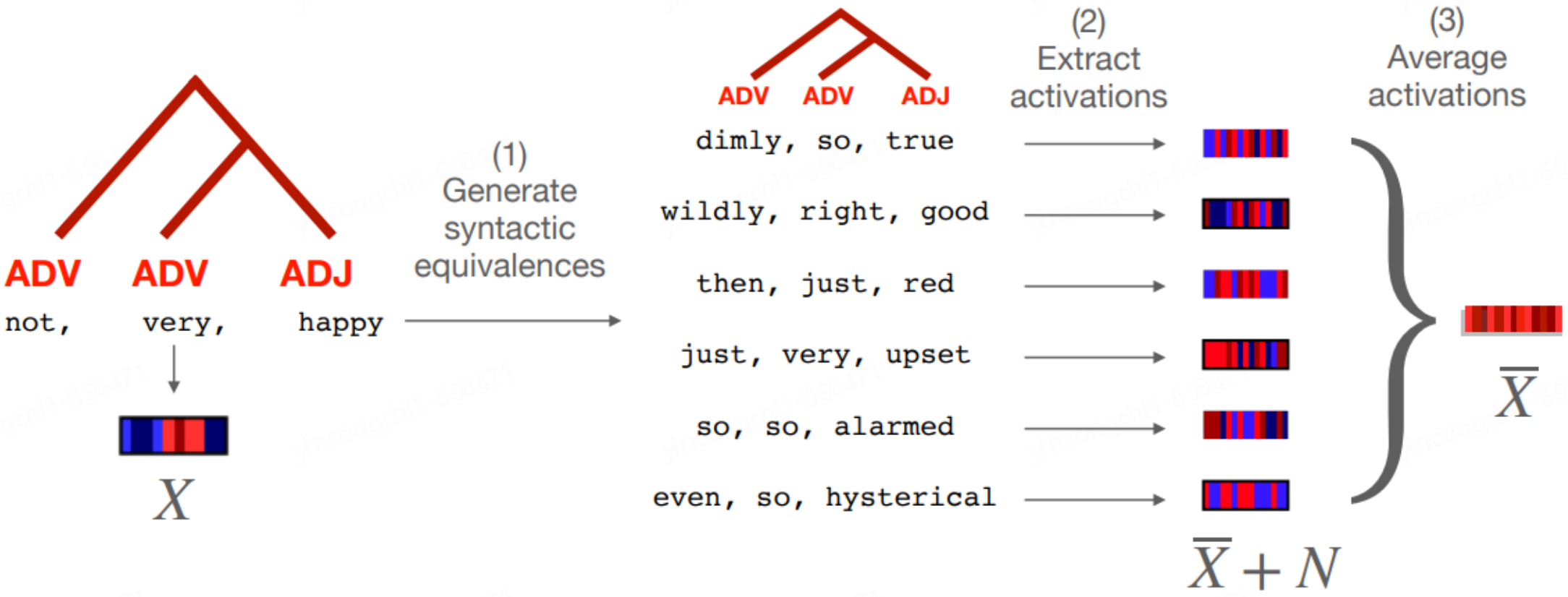
$$\overline{\Psi}_k = \frac{1}{k}\sum_{i=1}^{k}\left(\overline{\psi} + z_{w_i}\right) \xrightarrow[k\to\infty]{l.l.n} \overline{\psi}$$

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.

Syntactic representation

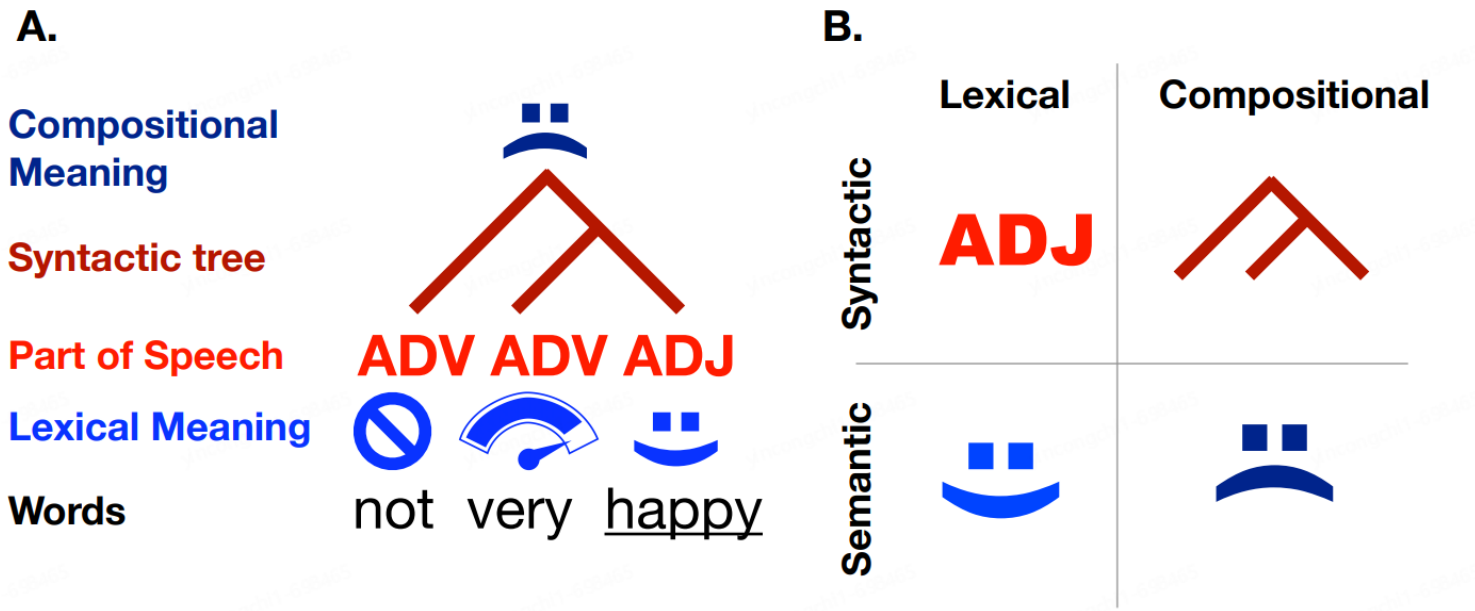
- *Isolating Syntactic Representation*



Five definitions to decompose representations

- ***Semantic***

To decompose syntax and semantics in distributed representations, the lexical or supra-lexical representations of a language system that are not syntactic.



Mapping representations to fMRI

- 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).

Mapping representations to fMRI

- To assess the mapping between *X* and *Y* , we use a linear spatio-temporal encoding model trained to predict the i^{th} fMRI volume given the network's activations *X*, on a given interval $I \subset [1 \dots N]$:

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

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

$$g_i : \mathbb{R}^{M \times d} \to \mathbb{R}^{5d}$$

$$u \mapsto \left[\widetilde{u_i}, \widetilde{u_{i-1}}, \dots, \widetilde{u_{i-4}} \right]$$

$$\widetilde{u_i} = \sum_{\substack{m \in [\![1 \dots M]\!] \\ \mathcal{T}(m) = i}} u_m$$

$$\mathcal{T} : [\![1 \dots M]\!] \to [\![1 \dots N]\!]$$

$$m \mapsto i \quad / \quad |t_{y_i} - t_{x_j}| = \min_{k \in [\![1 \dots N]\!]} |t_{y_k} - t_{x_m}|$$

Mapping representations to fMRI

- 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:

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

- We summarize the mapping with a Pearson correlation score evaluated on left out data:

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

Decomposing shared activations

- we use the definitions and methods to decompose the shared representations of two systems: a deep neural network, 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
  IV. decompose the brain activations by evaluating this mapping

Decomposing shared activations

- Language transformers are composed of multiple layers, each layer can be written as a non-linear system $\Psi^{(l)}$ that transforms a sequence of words into a vectorial representation.

$$\Psi^{(l)} : \mathcal{V}^M \rightarrow \mathbb{R}^{M \times d}$$

$$w \mapsto \Psi^{(l)}(w) = [\Psi^{(l)}(w)_1, \ldots, \Psi^{(l)}(w)_M]$$

- 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 before.

Decomposing shared activations

- Following the four definitions, we can decompose the activations X of Ψ into their:
  - lexical representations: the word embedding of the network $X^{(0)}$
  - compositional representations: $X^{(l)}, l > 0$
  - syntactic representations: $\overline{X^{(l)}}$, that can be extracted for any layer $l \in [0 \dots L]$
  - semantic representations: $X^{(l)} - \overline{X^{(l)}}$, as the residuals of syntactic representations
- Brain scores to decompose brain activity into:
  - lexical representations: $\mathcal{R}(X^{(0)})$
  - compositional representations: $\mathcal{R}(X^{(l)})$
  - syntactic representations: $\mathcal{R}(\overline{X^{(l)}}), l \in [0 \dots L]$
  - semantic representations: $\mathcal{R}(X^{(l)}) - \mathcal{R}(\overline{X^{(l)}})$

Review

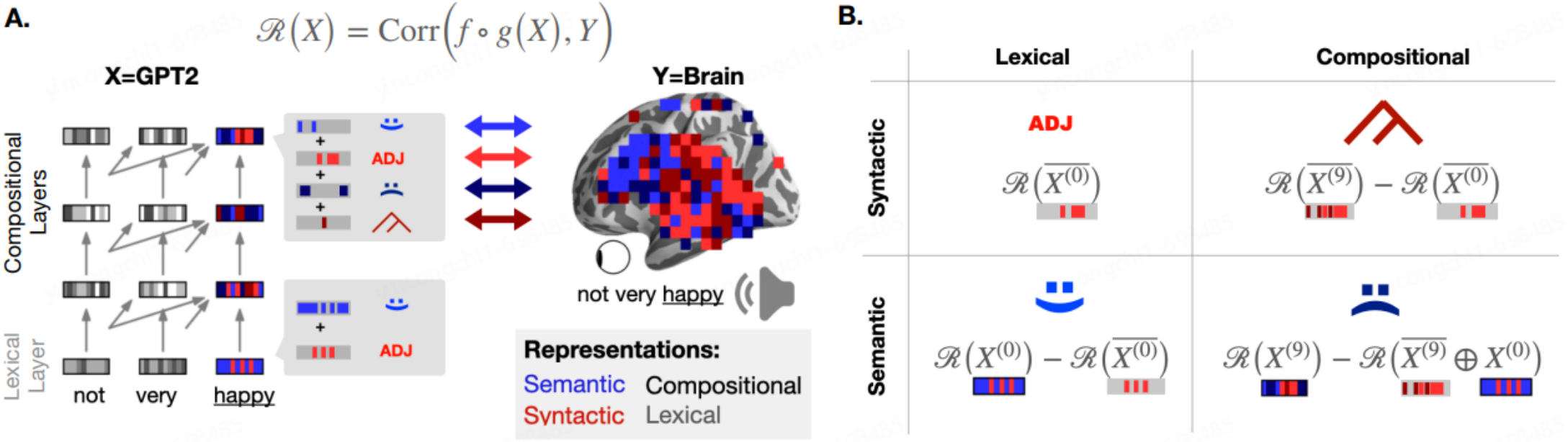


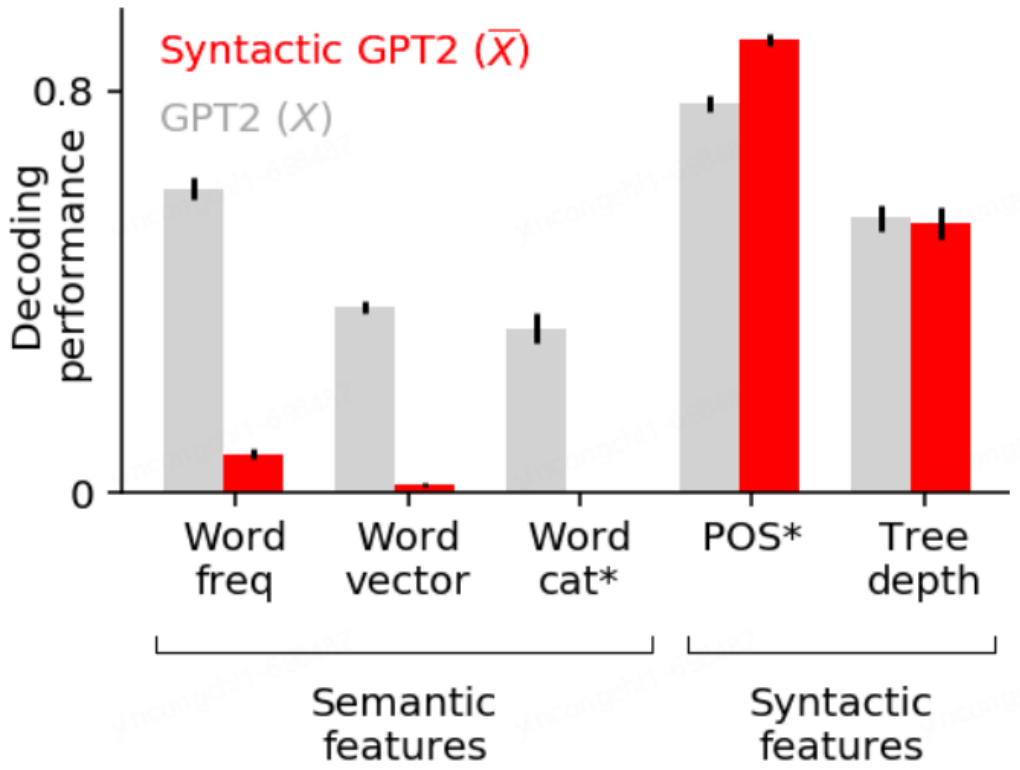
Figure 4. **Method to decompose the language representations shared between brains and deep language models A.** The human brain and modern language models like GPT-2 both generate *distributed* representations, which are thus difficult to link with the *symbolic* properties of linguistic theories. We 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.

Experiments

- Datasets
  - Narratives
- Phonological features
  - Phone rate
  - Word rate
  - Phoneme, stress and tone of words
- Language model
  - GPT-2 with 12 Transformer layers
  - BERT
  - XLNet
  - RoBERTa
  - ALBERT

Experiments

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.



Experiments

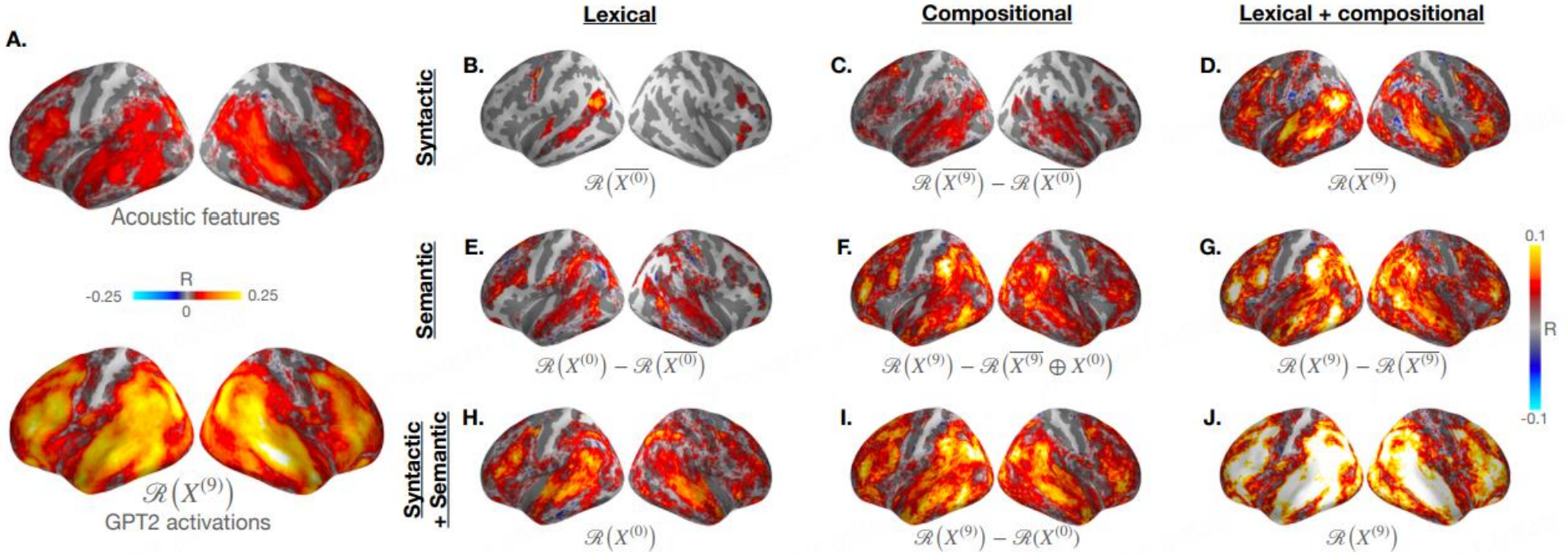
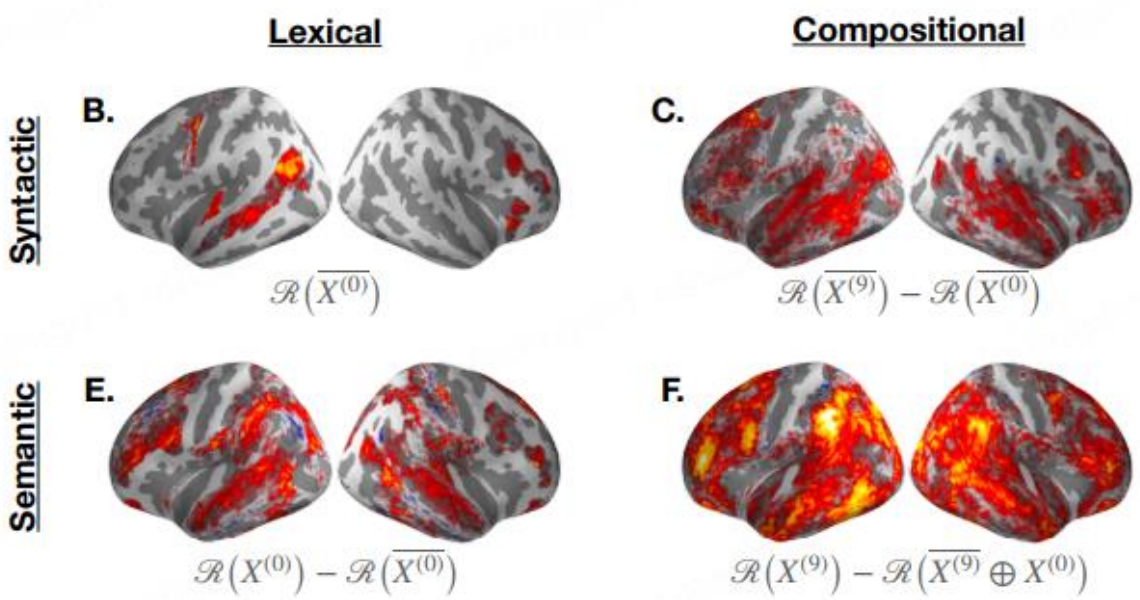
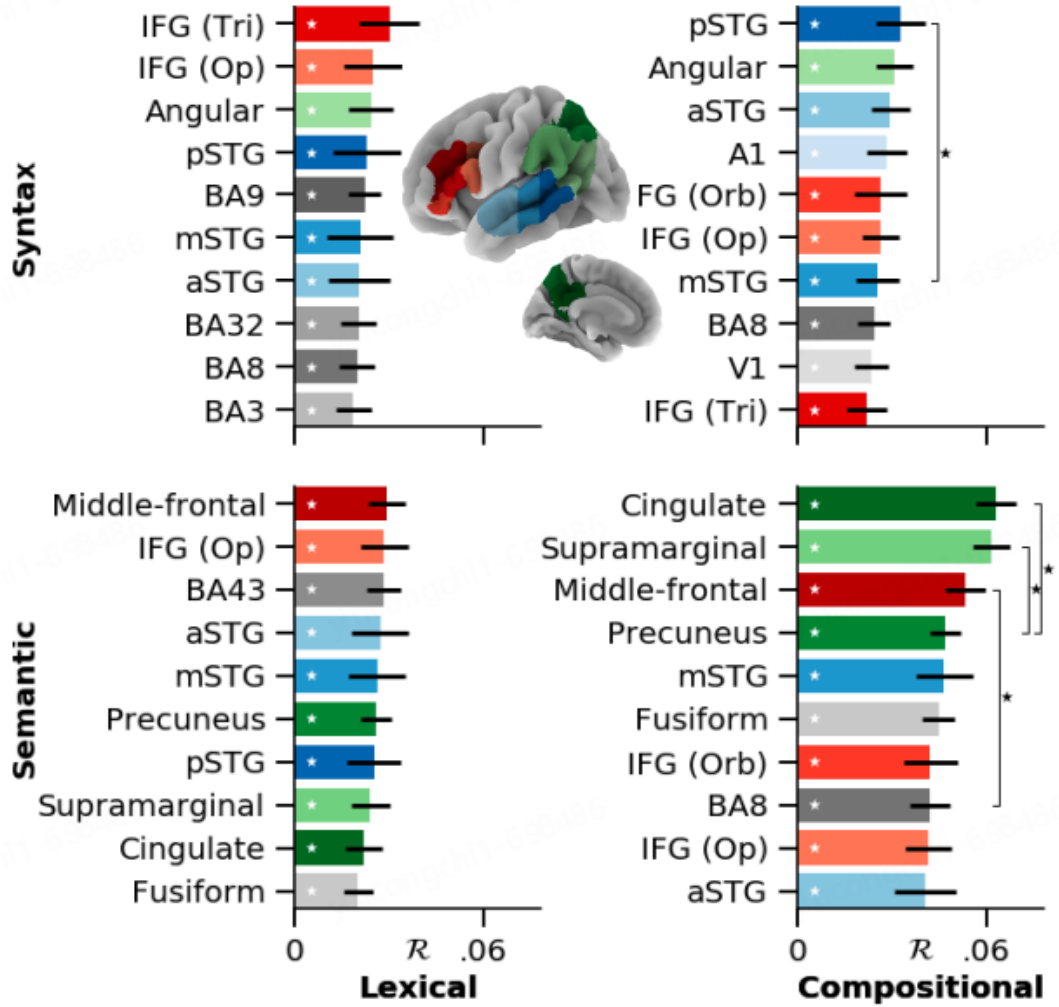


Figure 5. **Results** Decomposition of the brain scores of 345 subjects listening to narratives into their phonological (A) syntactic (B-D), semantic (E-G), lexical (B-H), compositional (C-I) components and their combinations (ten combinations in total). **A** Comparison between the brain scores of three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

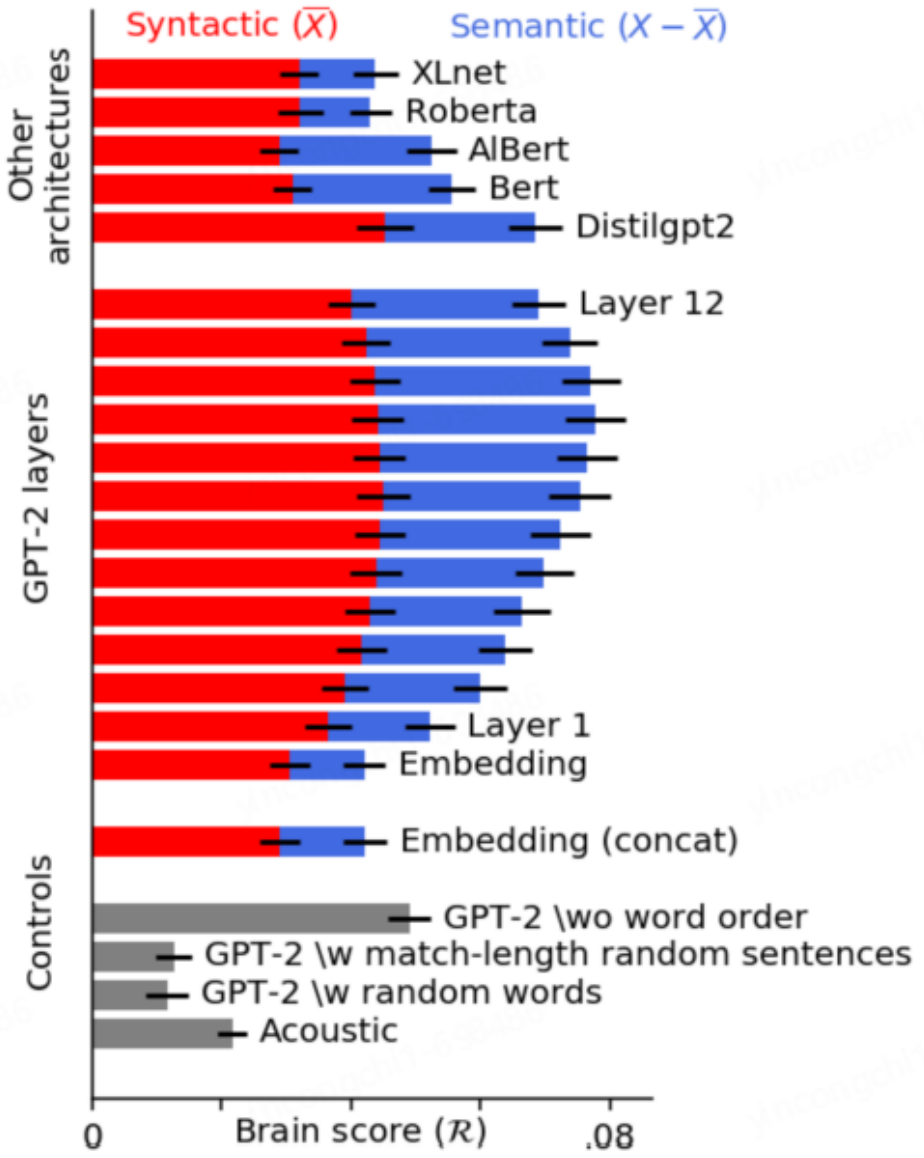
Experiments





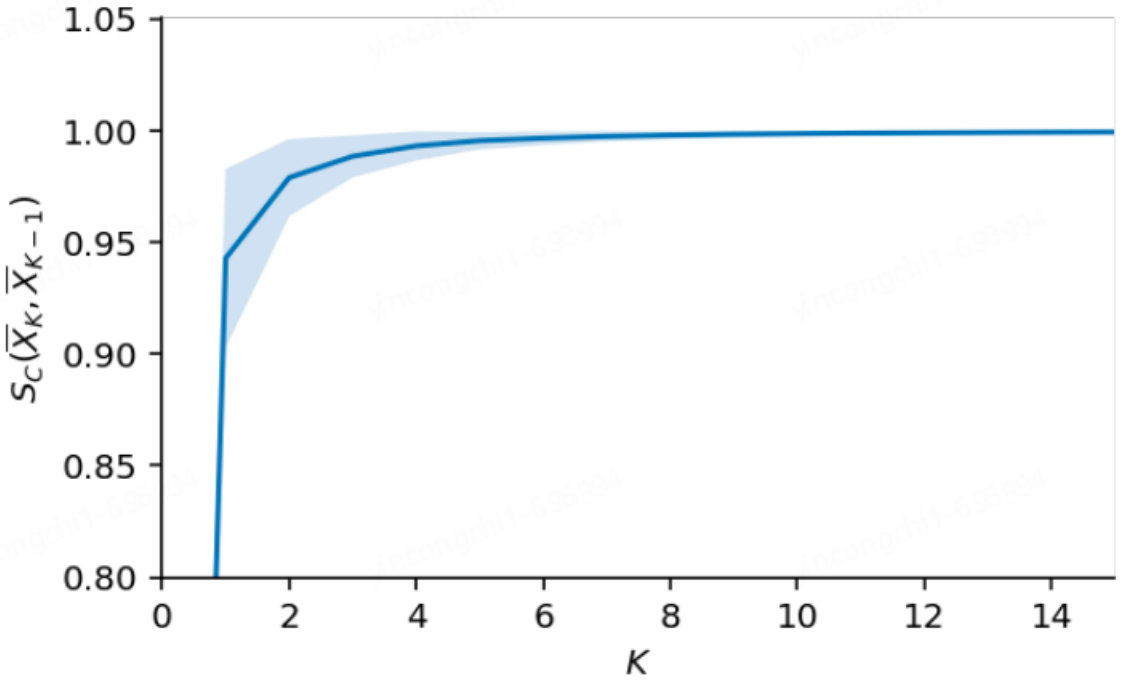
Experiments

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.)



Experiments

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.



Thanks for listening