Disentangling Syntax and Semantics in the Brain with Deep Networks

facebook
Artificial Intelligence Research

Charlotte Caucheteux (INRIA/FAIR), Alexandre Gramfort (INRIA), Jean-Remi King (FAIR/ENS)

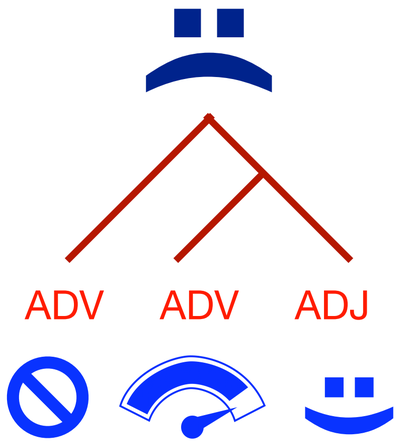
What are the neural bases of syntax and semantics?

Compositional **meaning**

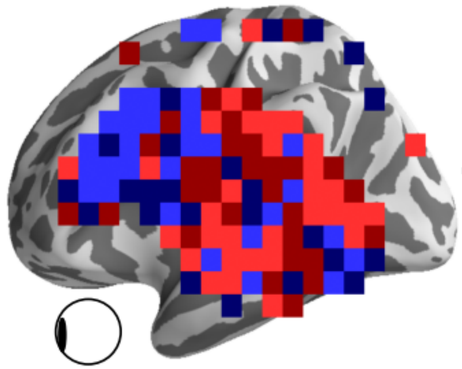
Compositional **syntax**

Lexical **syntax**

Lexical **meaning**



not very happy



Semantic representations

Syntactic representations

ADJ

ADV

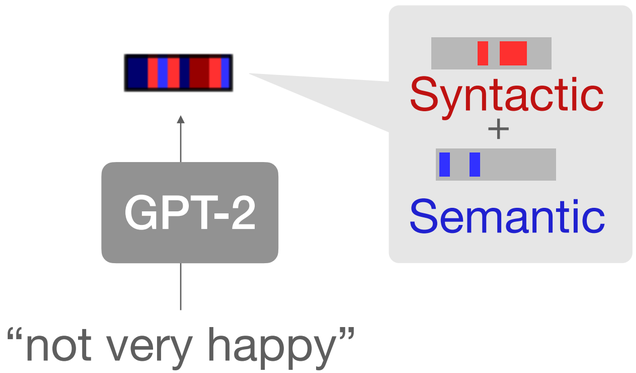
not very happy

What are the neural bases of syntax and semantics?

1. We introduce a method to disentangle syntax and semantics in deep nets' activations

2. We use the disentangled activations to decompose language in the brain

3. We apply our method to the fMRI recordings of 345 subjects listening to stories*

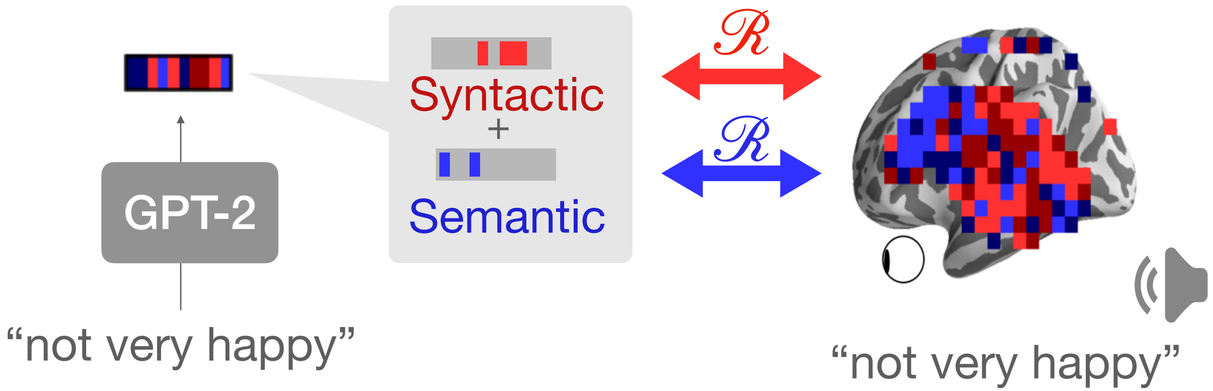


What are the neural bases of syntax and semantics?

1. We introduce a method to disentangle syntax and semantics in deep nets' activations

2. We use the disentangled activations to decompose language in the brain

3. We apply our method to the fMRI recordings of 345 subjects listening to stories*

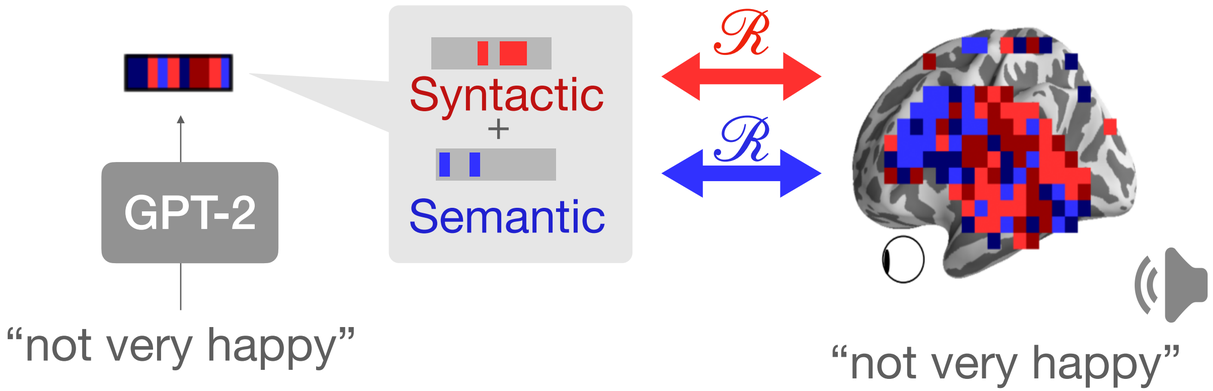


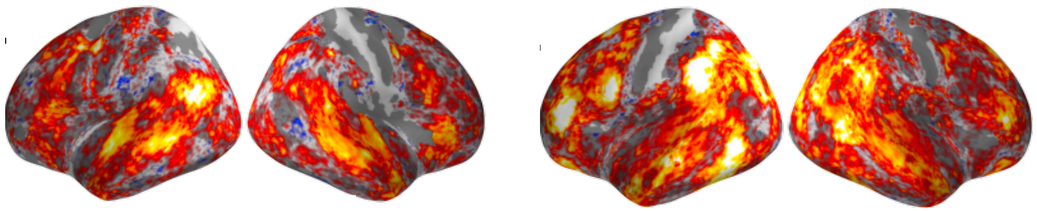
What are the neural bases of syntax and semantics?

1. We introduce a method to disentangle syntax and semantics in deep nets' activations

2. We use the disentangled activations to decompose language in the brain

3. We apply our method to the fMRI recordings of 345 subjects listening to stories*





Syntax
$\mathscr{R}$

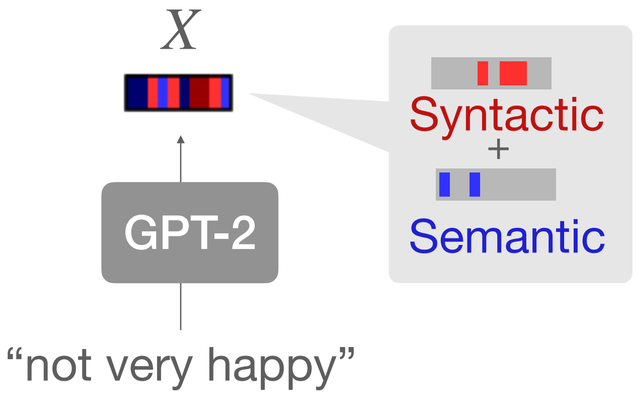
Semantics
$\mathscr{R}$

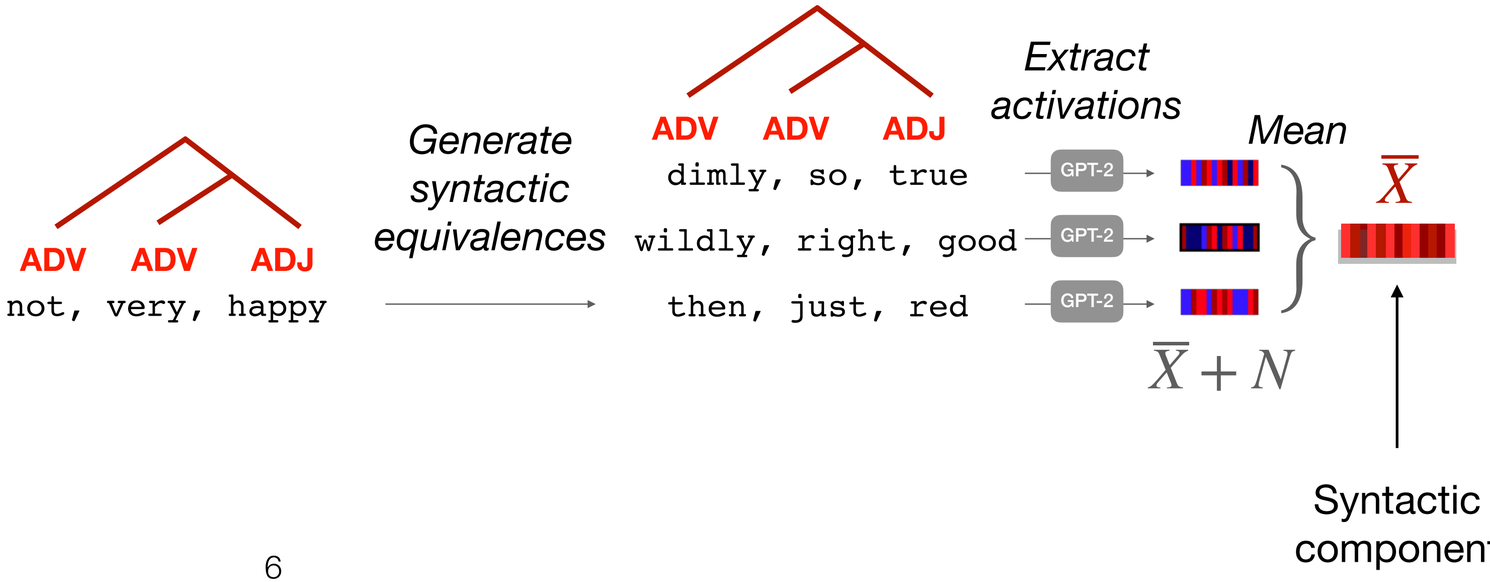
What are the neural bases of syntax and semantics?

1. We introduce a method to disentangle **syntax** and **semantics** in **deep nets' activations**

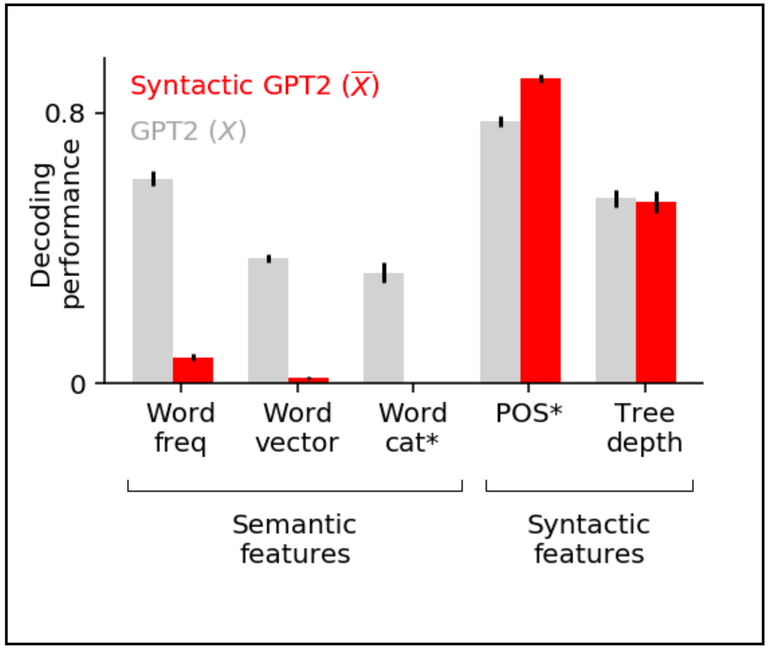
2. We use the disentangled activations to decompose language in the brain

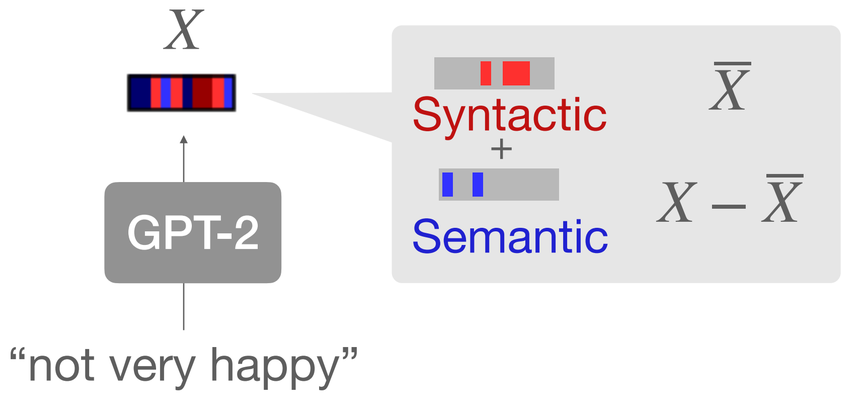
3. We apply our method to the fMRI recordings of 345 subjects listening to stories*

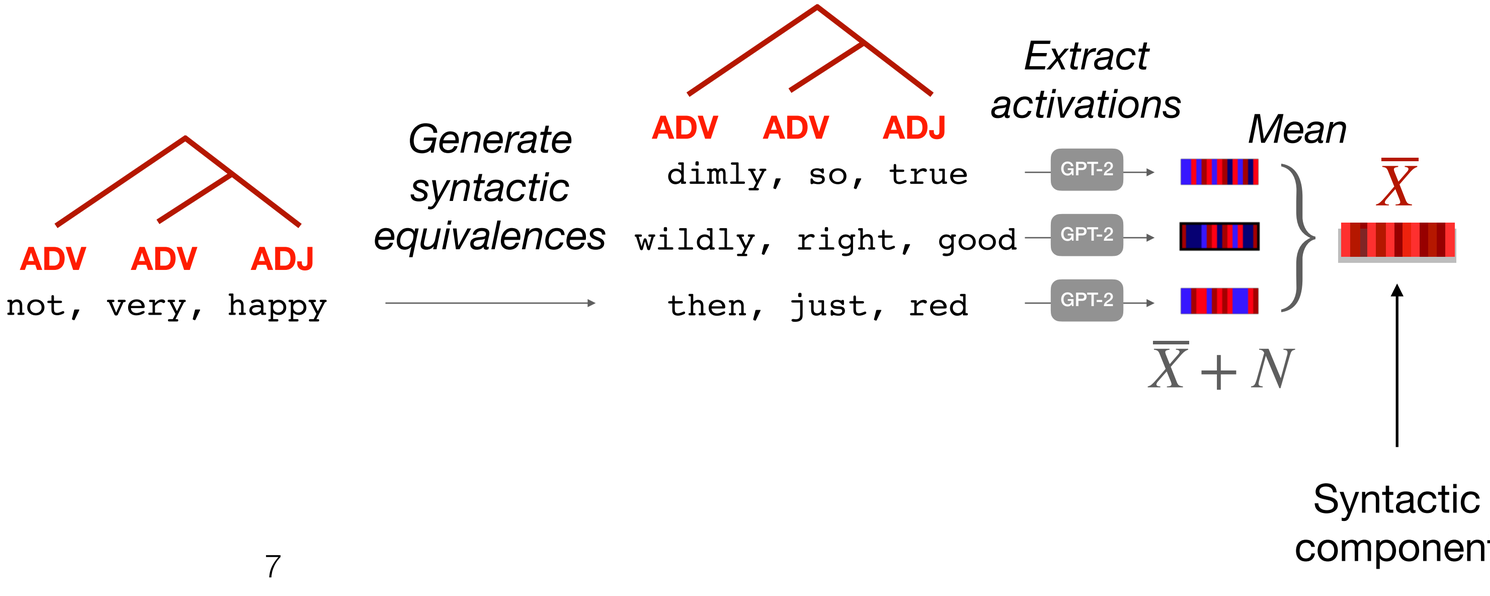




What are the neural bases of syntax and semantics?





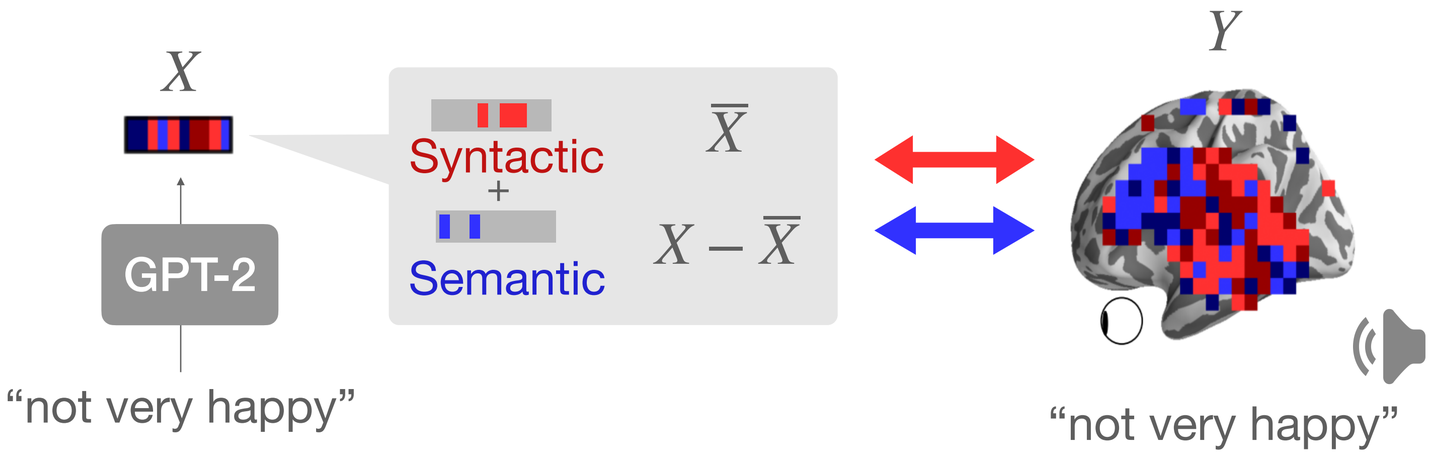


What are the neural bases of syntax and semantics?

1. We introduce a method to disentangle syntax and semantics in deep nets' activations

2. We use the disentangled activations to **decompose language in the brain**

3. We apply our method to the fMRI recordings of 345 subjects listening to stories*

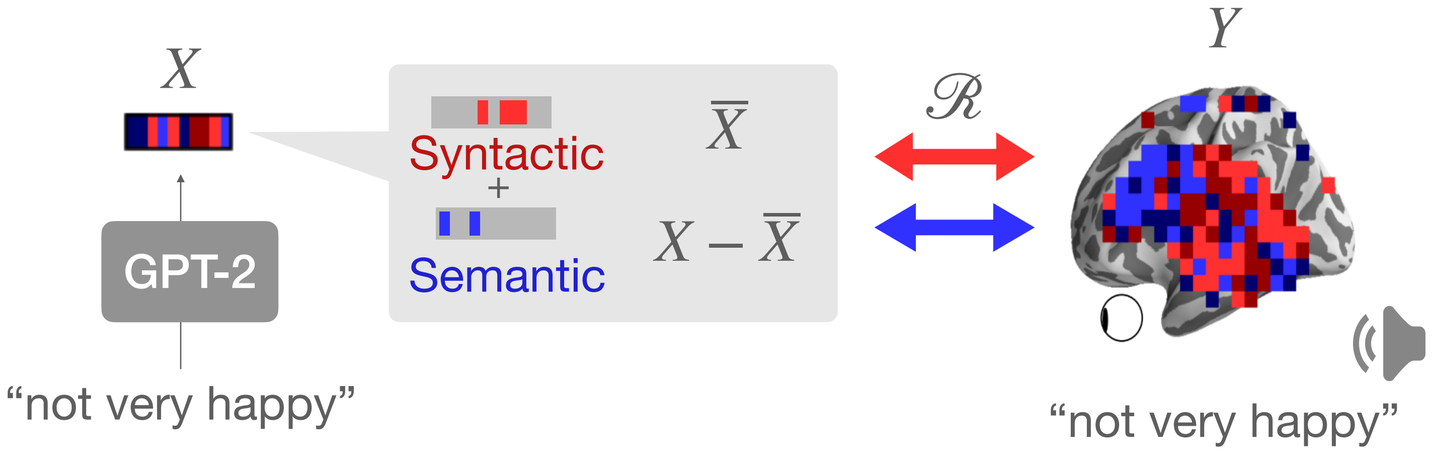


What are the neural bases of syntax and semantics?

1. We introduce a method to disentangle syntax and semantics in deep nets' activations

2. We use the disentangled activations to **decompose language in the brain**

3. We apply our method to the fMRI recordings of 345 subjects listening to stories*



$$\text{Brain score}: \mathscr{R}\left(\overline{X}\right) = \text{Corr}\left(W^T\overline{X}, Y\right)$$

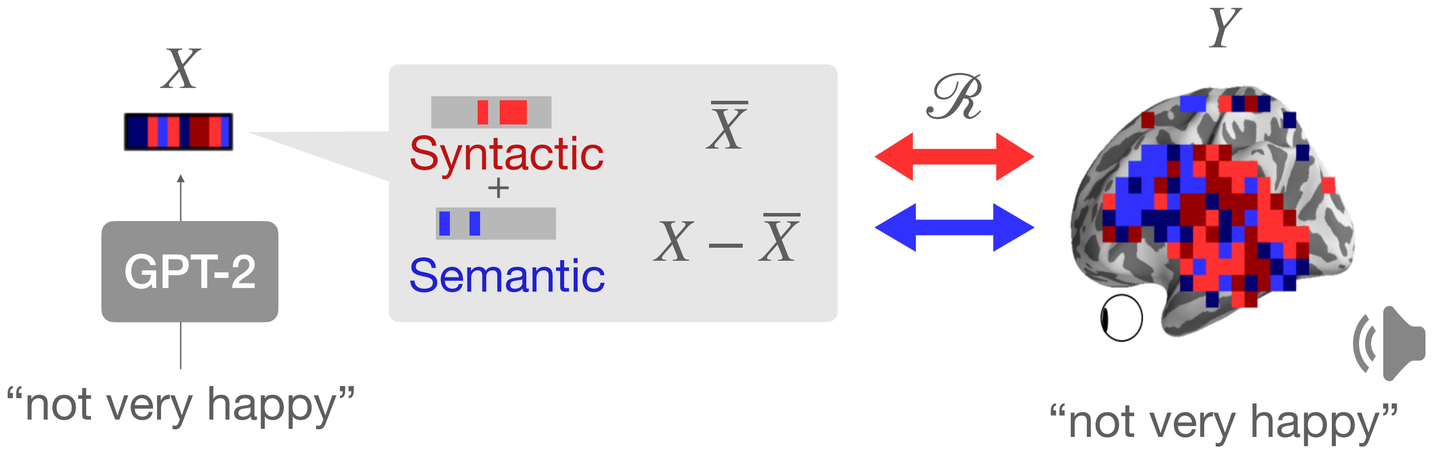
Estimated by Ridge regression on training data

What are the neural bases of syntax and semantics?

1. We introduce a method to disentangle syntax and semantics in deep nets' activations

2. We use the disentangled activations to **decompose language in the brain**

3. We apply our method to the fMRI recordings of 345 subjects listening to stories*



$$\text{Brain score}: \mathscr{R}\left(\overline{X}\right) = \text{Corr}\left(W^T\overline{X}, Y\right)$$

Measures of brain representations

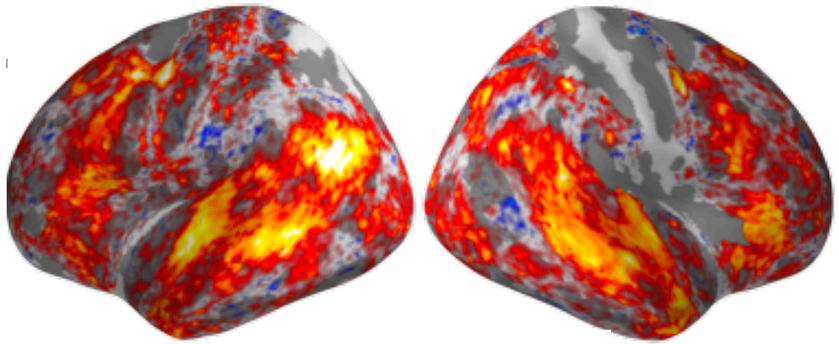
Syntactic score	$\mathscr{R}\left(\overline{X}\right)$
Semantic score	$\mathscr{R}\left(X\right) - \mathscr{R}\left(\overline{X}\right)$

What are the neural bases of syntax and semantics?

1. We introduce a method to disentangle syntax and semantics in deep nets' activations

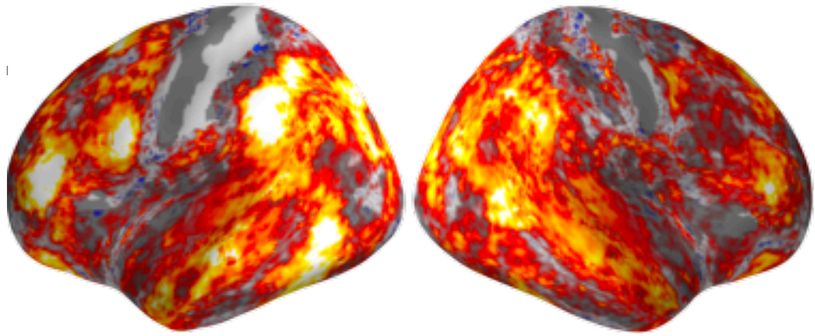
2. We use the disentangled activations to decompose language in the brain

3. We apply our method to the **fMRI** recordings of **345 subjects** listening to stories*



Syntax

$\mathscr{R}(\overline{X})$



Semantics

$\mathscr{R}(X) - \mathscr{R}(\overline{X})$

Measures of brain representations

Syntactic score	$\mathscr{R}(\overline{X})$
Semantic score	$\mathscr{R}(X) - \mathscr{R}(\overline{X})$

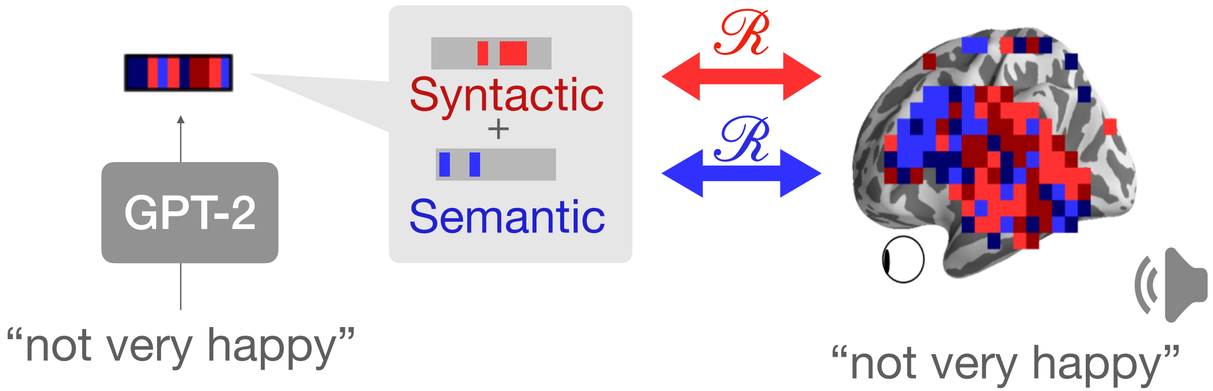
*345 subjects listening to ~4 hours of unique audio from the "Narratives" dataset (*Nastase et al. 2020*)

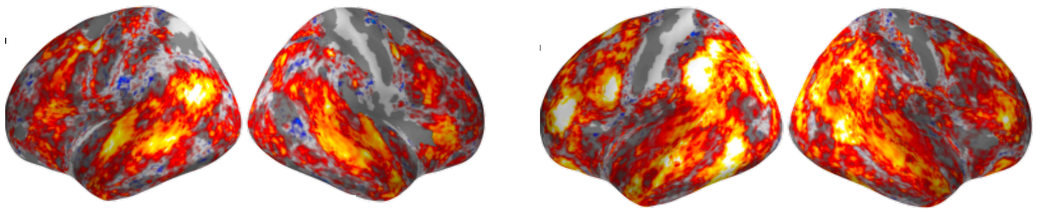
What are the neural bases of syntax and semantics?

1. We introduce a method to disentangle syntax and semantics in deep nets' activations

2. We use the disentangled activations to decompose language in the brain

3. We apply our method to the fMRI recordings of 345 subjects listening to stories*





Syntax
$\mathscr{R}$

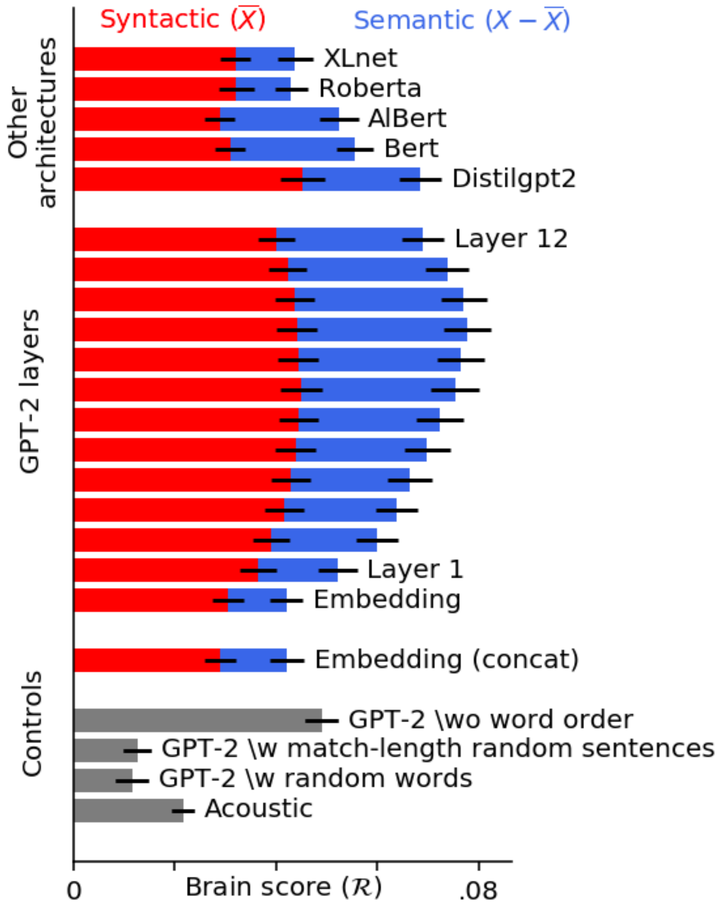
Semantics
$\mathscr{R}$

Generalisation to other architectures and layers

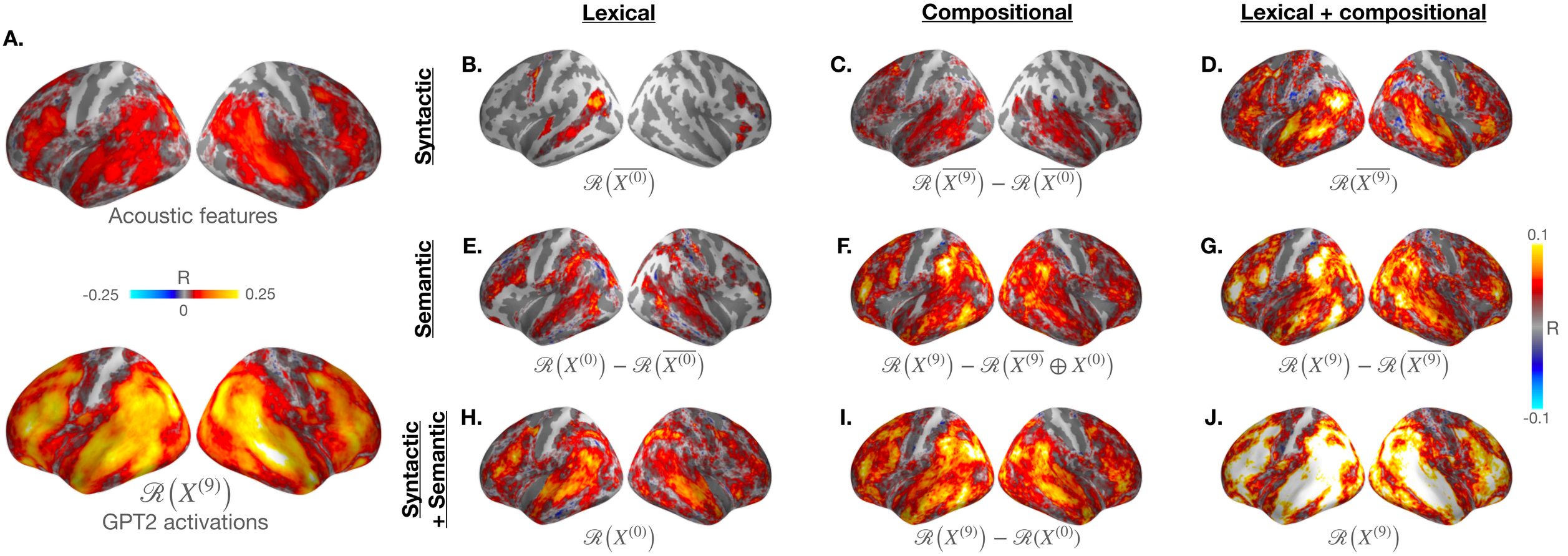
1. We introduce a method to disentangle syntax and semantics in deep nets' activations

2. We use the disentangled activations to decompose language in the brain

3. We apply our method to the fMRI recordings of 345 subjects listening to stories*



Fine grained decomposition



Thank you for your attention!

facebook
Artificial Intelligence Research

Charlotte Caucheteux (INRIA/FAIR), Alexandre Gramfort (INRIA), Jean-Remi King (FAIR/ENS)