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)

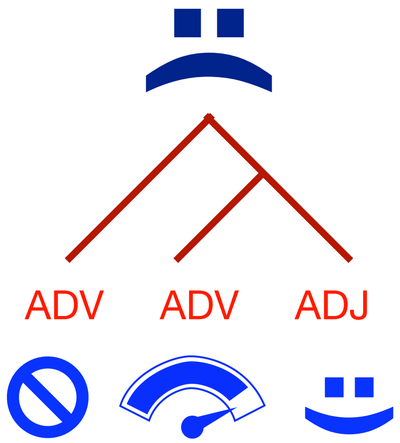
What are the neural bases of syntax and semantics?

Compositional **meaning**

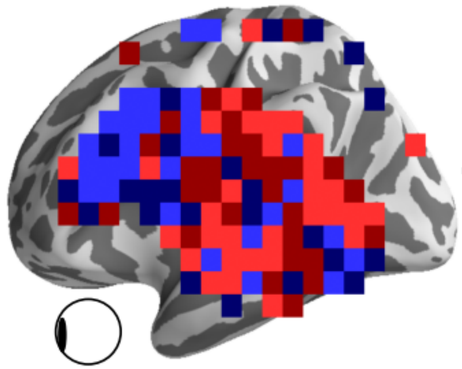
Compositional **syntax**

Lexical **syntax**

Lexical **meaning**



not very happy



Semantic representations

Syntactic representations

ADJ

ADV

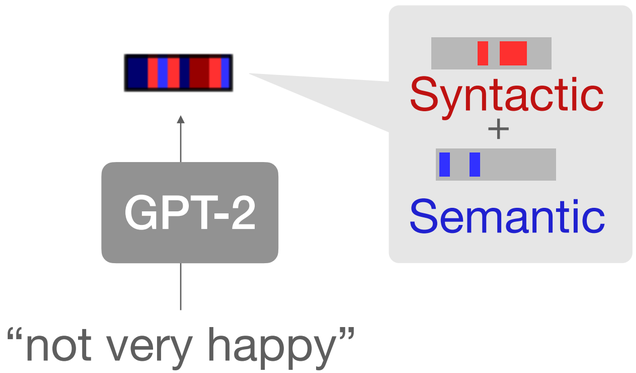
not very happy

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*

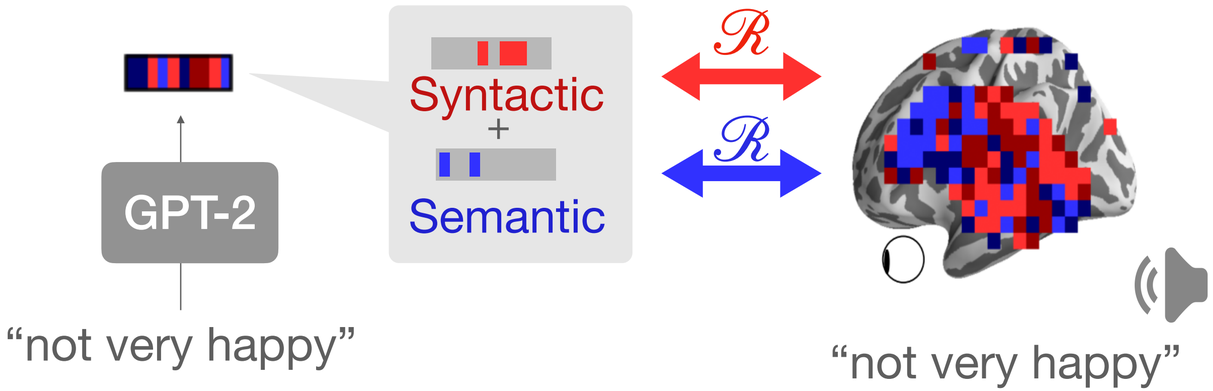


What are the neural bases of syntax and semantics?

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*

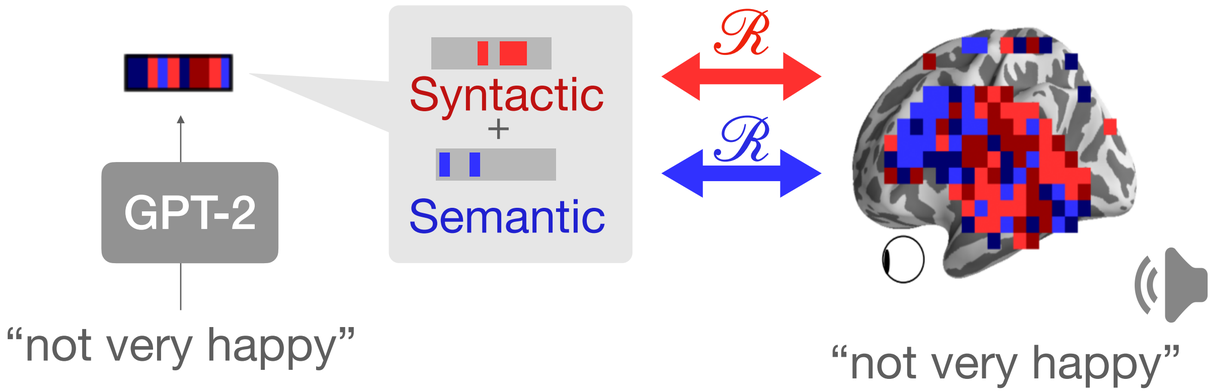


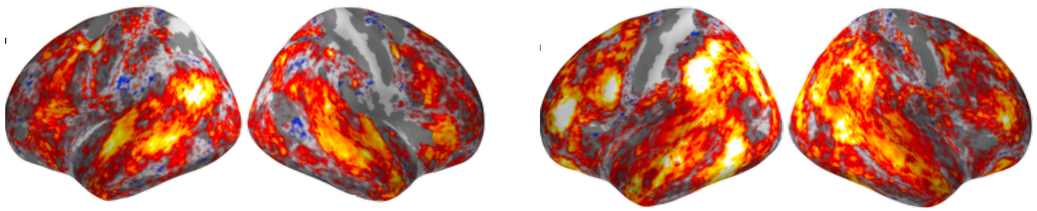
What are the neural bases of syntax and semantics?

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
$\mathscr{R}$

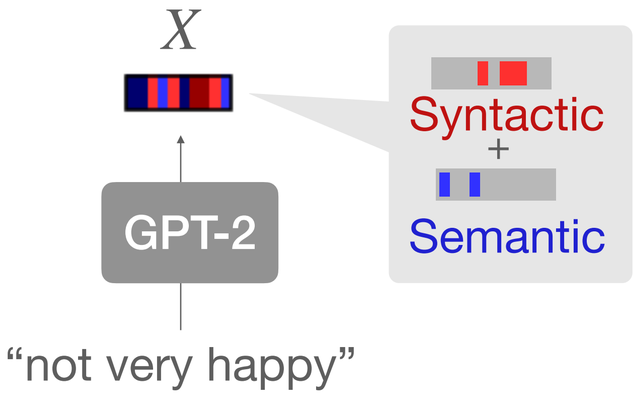
Semantics
$\mathscr{R}$

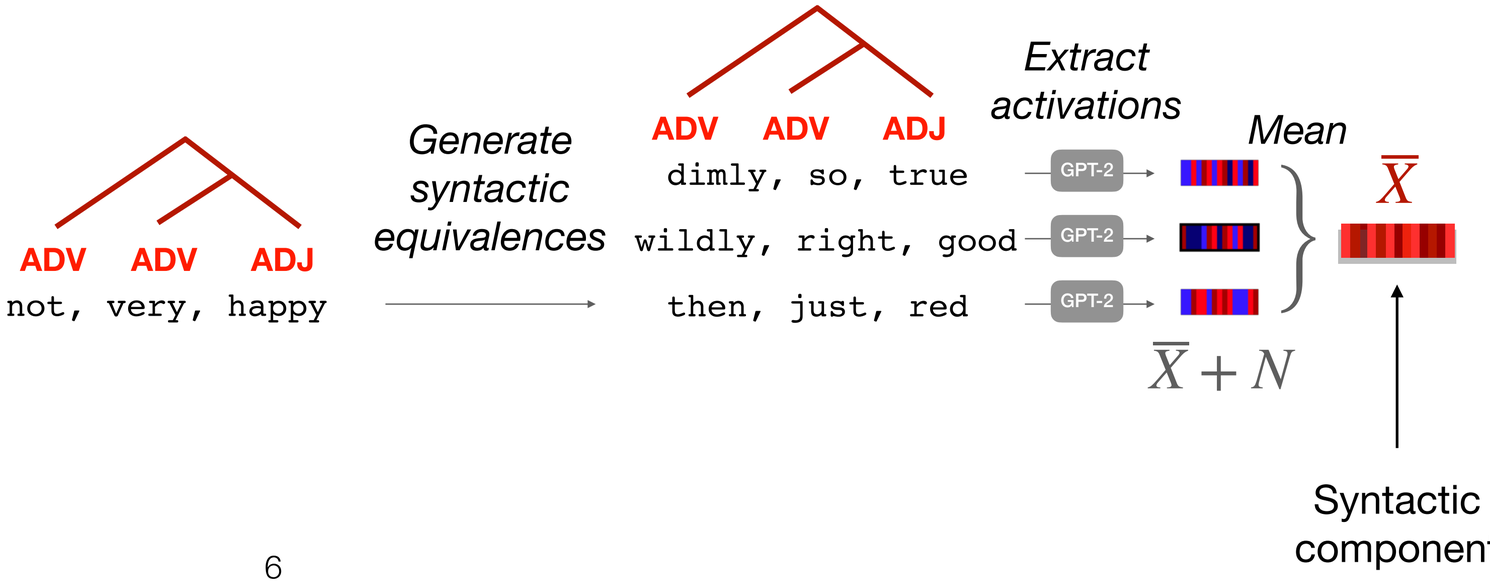
What are the neural bases of syntax and semantics?

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

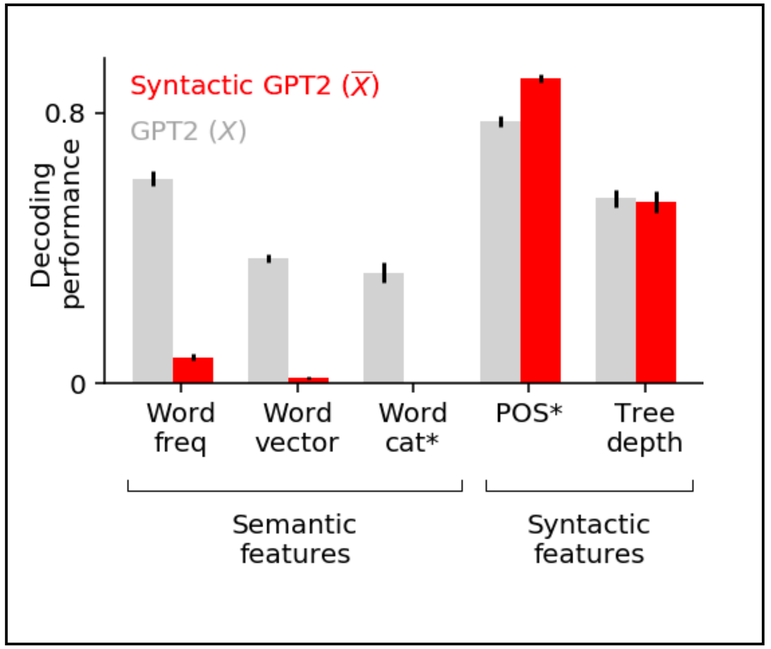
2. We use the disentangled activations to decompose language in the brain

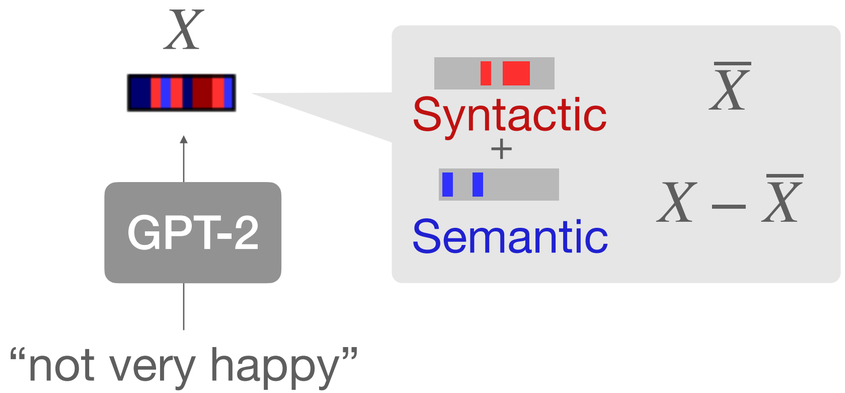
3. We apply our method to the fMRI recordings of 345 subjects listening to stories*

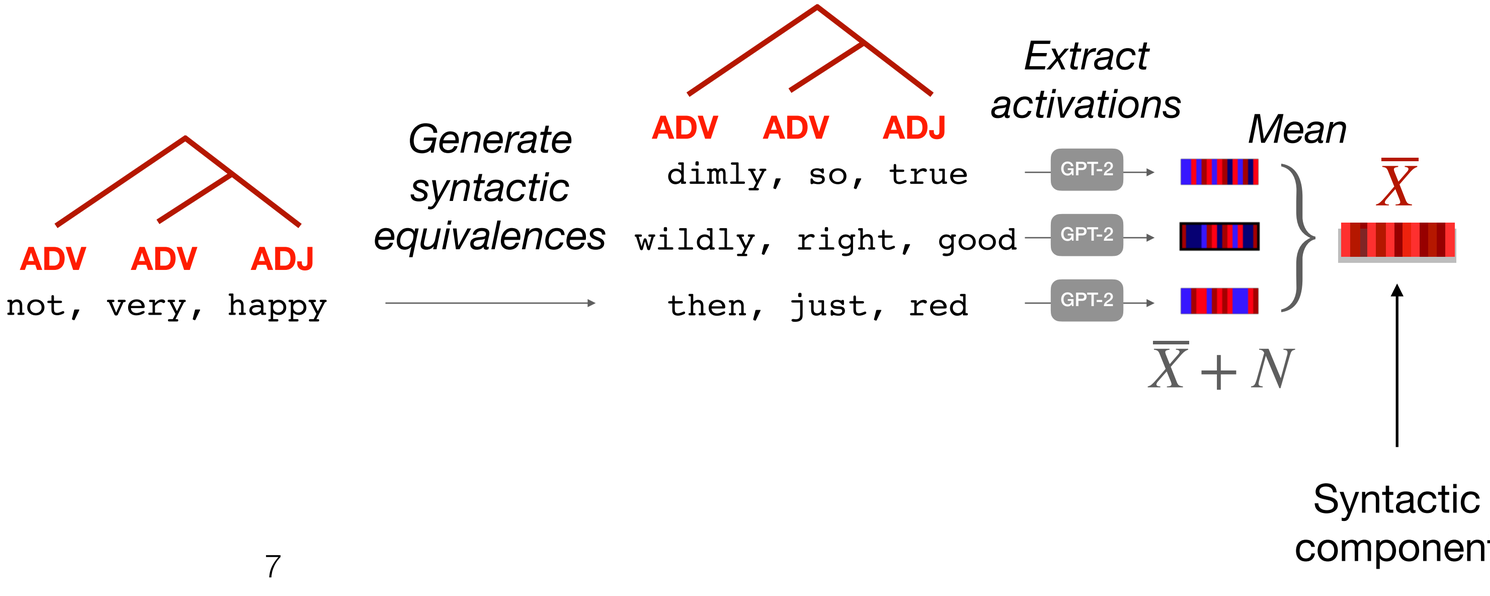




What are the neural bases of syntax and semantics?





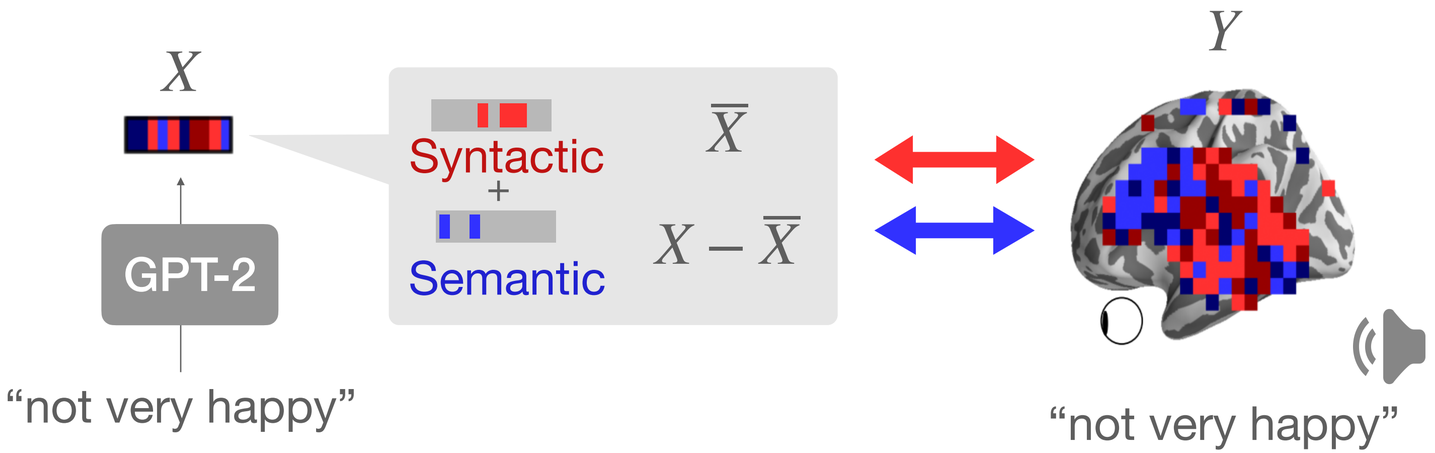


What are the neural bases of syntax and semantics?

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*

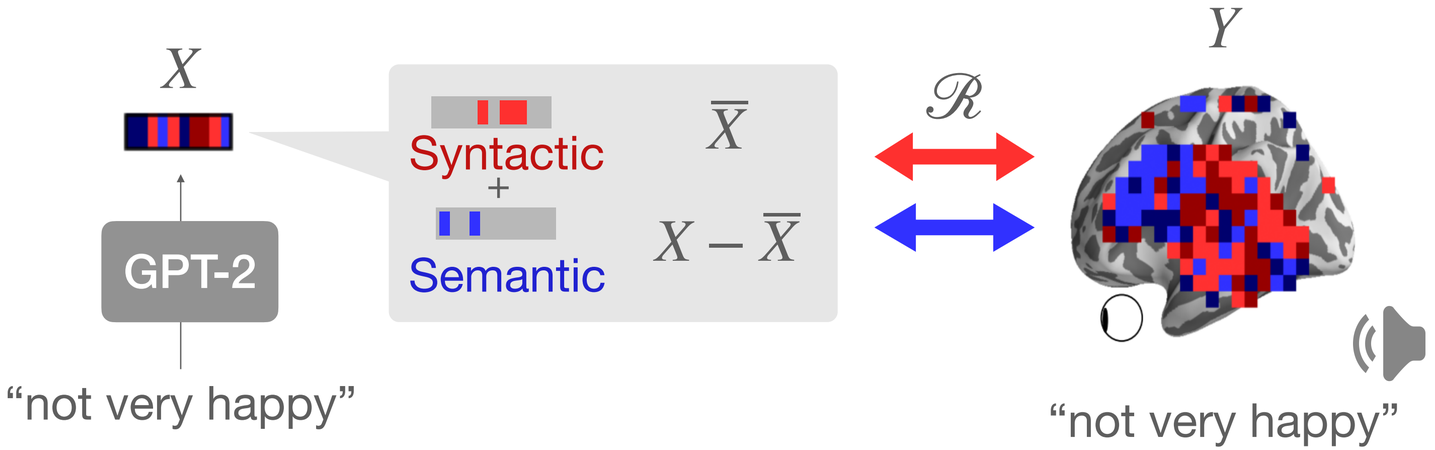


What are the neural bases of syntax and semantics?

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*



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

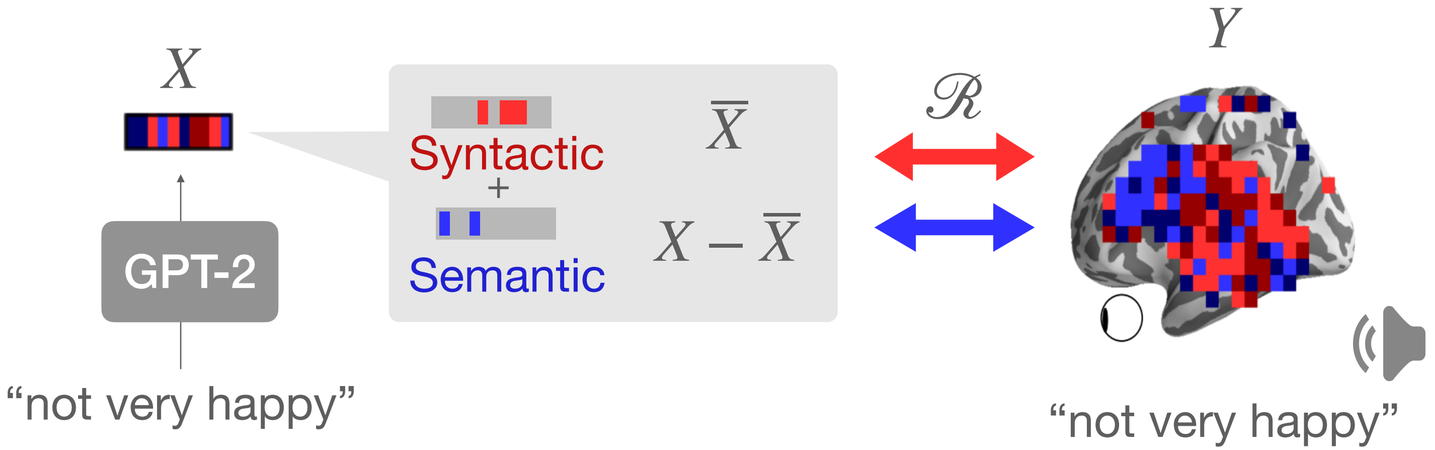
Estimated by Ridge regression on training data

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

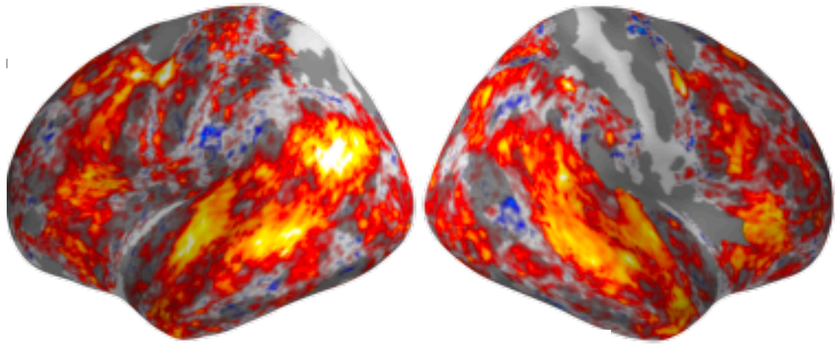
Syntactic score	$\mathscr{R}\left(\overline{X}\right)$
Semantic score	$\mathscr{R}\left(X\right) - \mathscr{R}\left(\overline{X}\right)$

What are the neural bases of syntax and semantics?

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

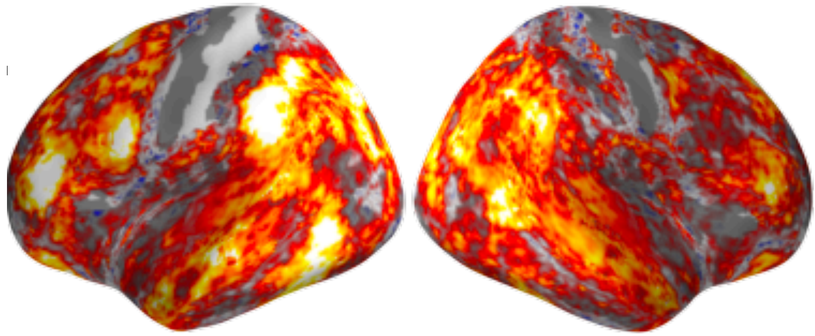
2. We use the disentangled activations to decompose language in the brain

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

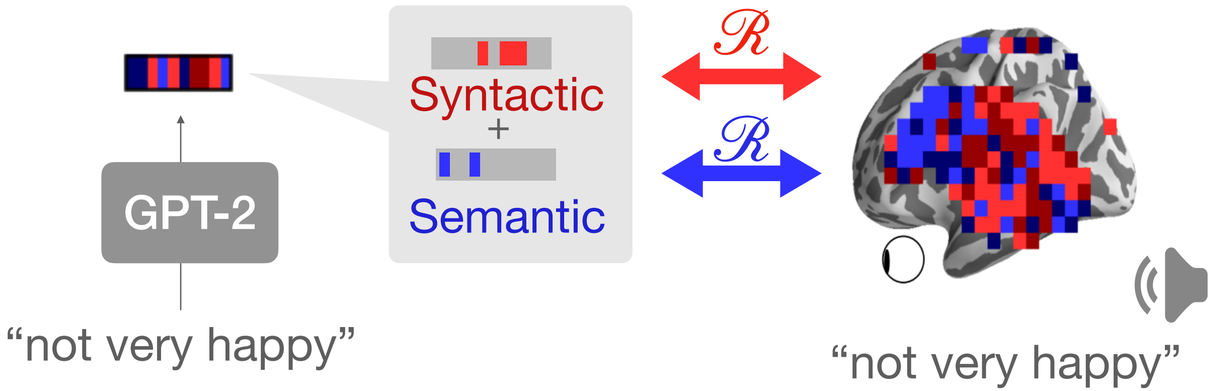
*345 subjects listening to ~4 hours of unique audio from the "Narratives" dataset (*Nastase et al. 2020*)

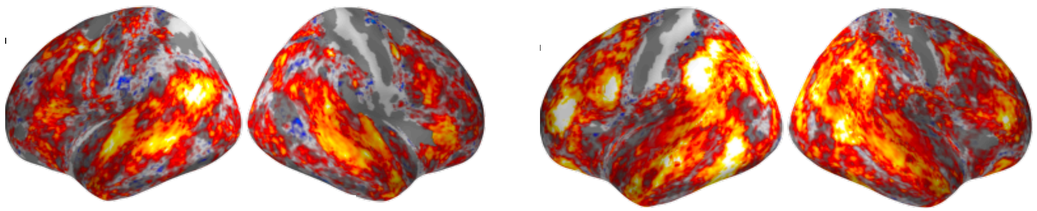
What are the neural bases of syntax and semantics?

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
$\mathscr{R}$

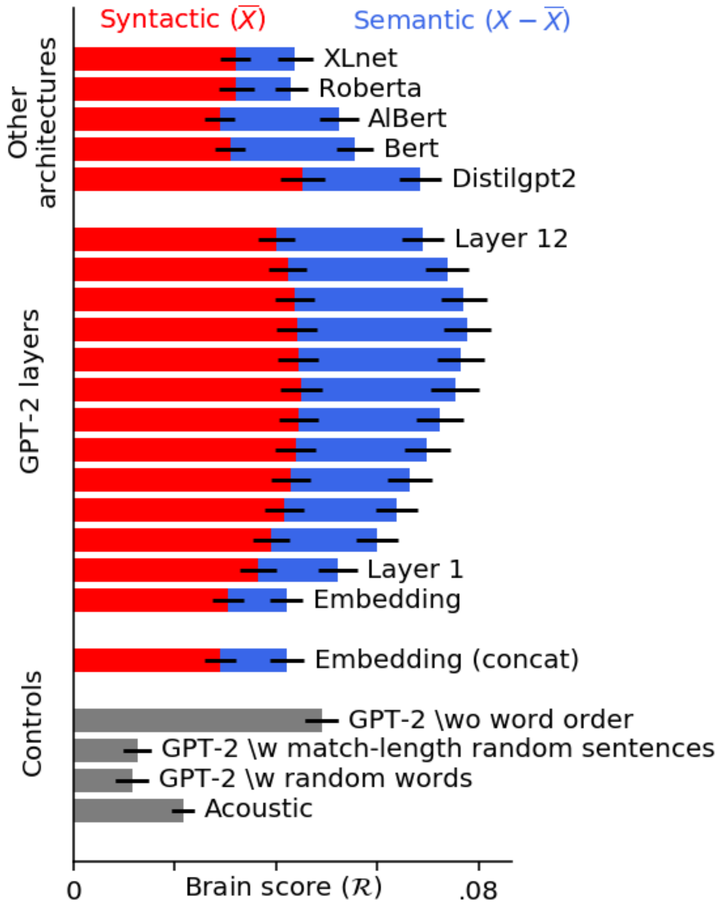
Semantics
$\mathscr{R}$

Generalisation to other architectures and layers

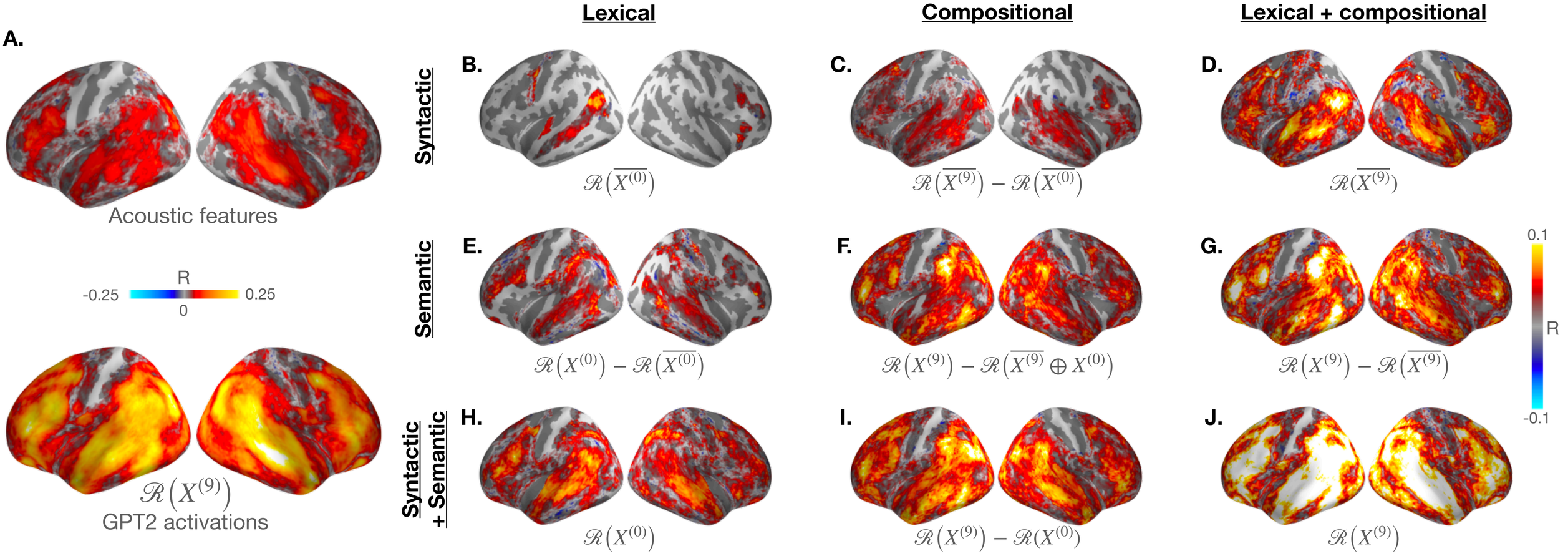
1. We introduce a method to disentangle syntax and semantics in deep nets' activations

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)

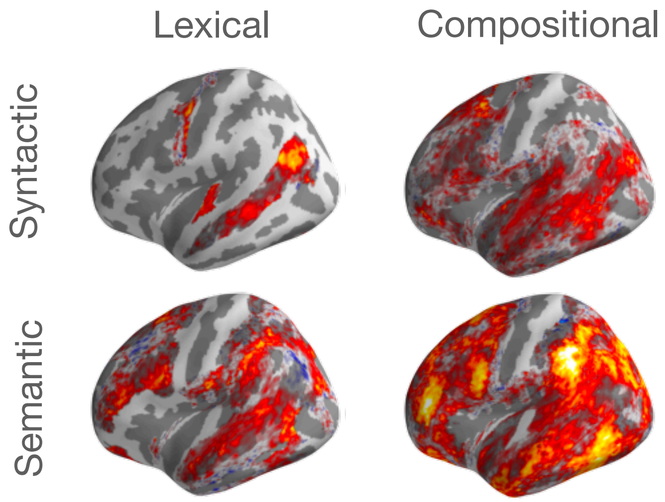
Generality of the results

A- Dataset

Narratives: fMRI data for evaluating models of naturalistic language comprehension

Samuel A. Nastase[1], Yun-Fei Liu[2], Hanna Hillman[1], Asieh Zadbood[1], Liat Hasenfratz[1], Neggin Keshavarzian[1], Janice Chen[2], Christopher J. Honey[2], Yaara Yeshurun[3], Mor Regev[4], Mai Nguyen[1], Claire H. C. Chang[1], Christopher Baldassano[5], Olga Lositsky[6], Erez Simony[7,8], Michael A. Chow[9], Yuan Chang Leong[10], Paula P. Brooks[1], Emily Micciche[11], Gina Choe[1], Ariel Goldstein[1], Tamara Vanderwal[12], Yaroslav O. Halchenko[13], Kenneth A. Norman[1], Uri Hasson[1]

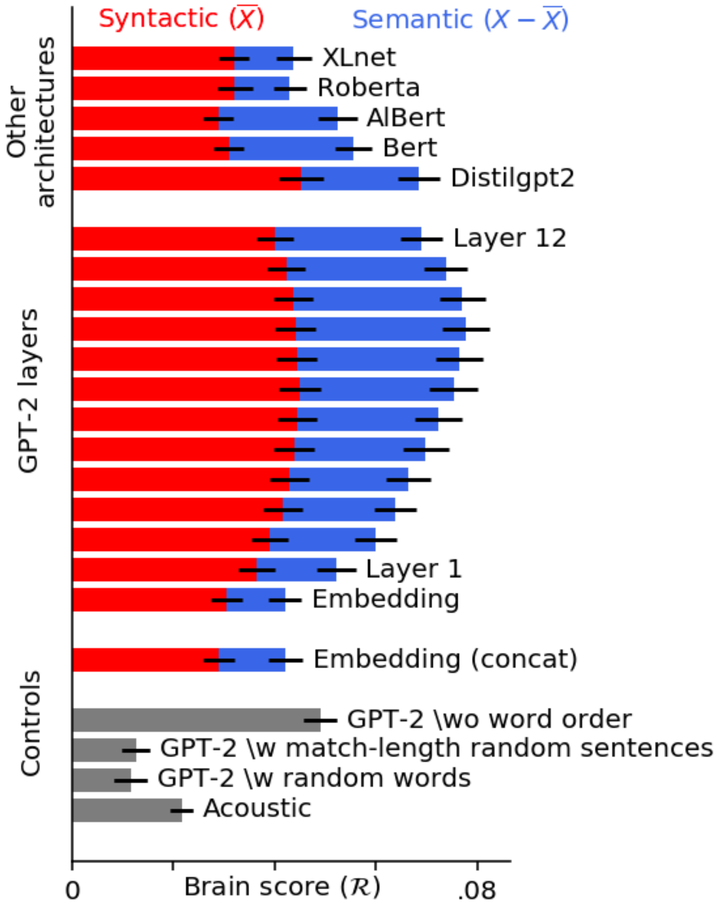
B- Results



Representations:

Syntactic	$\mathcal{R}\left(\overline{X^{(k)}}\right)$
Semantic	$\mathcal{R}\left(X^{(k)}\right) - \mathcal{R}\left(\overline{X^{(k)}}\right)$
Lexical	$\mathcal{R}\left(X^{(0)}\right)$
Compositional	$\mathcal{R}\left(X^{(9)}\right) - \mathcal{R}\left(X^{(0)}\right)$

C- Results generalise to other layers and deep networks

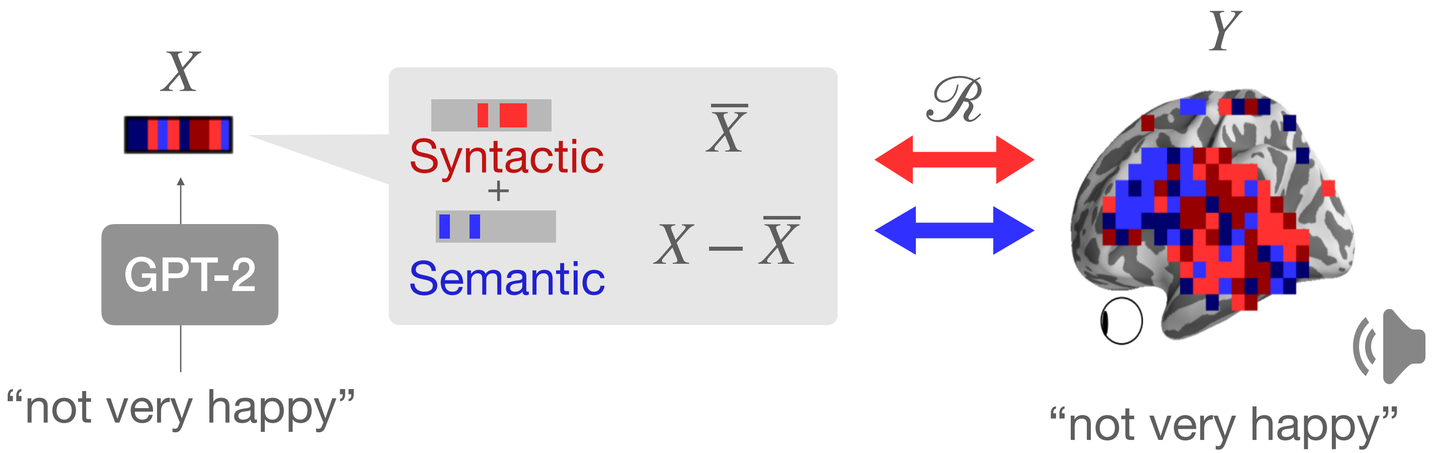


What are the neural bases of syntax and semantics?

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 brain of 345 subjects listening to stories



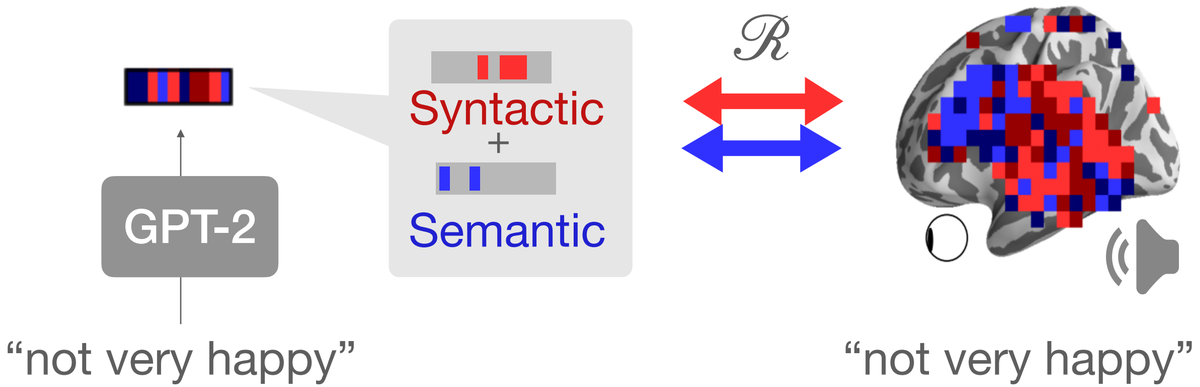
$$\text{Brain score}: \mathscr{R}(X) = \text{Corr}(W^T X, Y)$$

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 brain of 345 subjects listening to stories

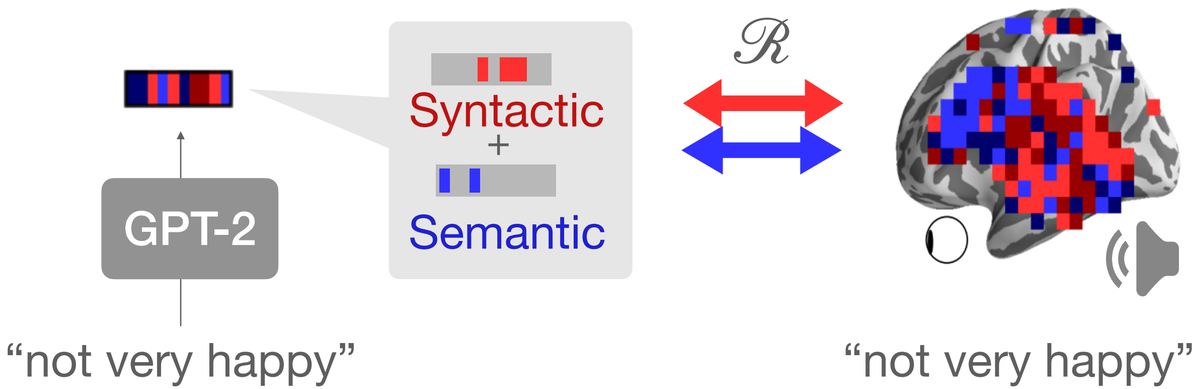


What are the neural bases of syntax and semantics?

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 brain of 345 subjects listening to stories

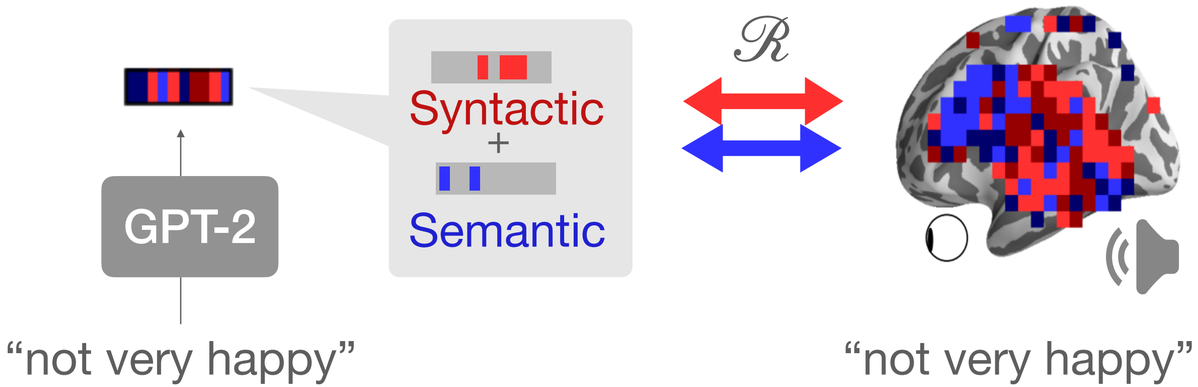


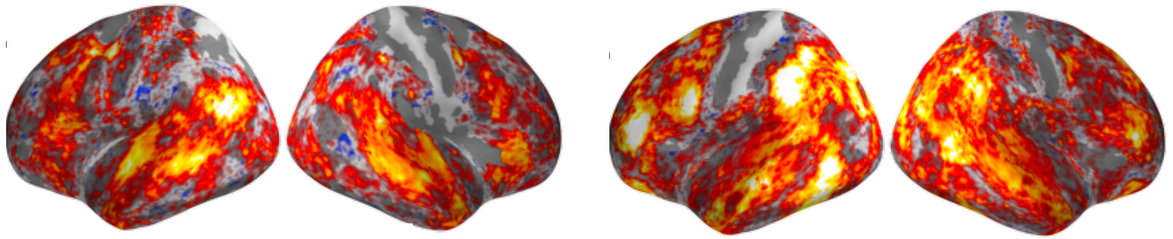
What are the neural bases of syntax and semantics?

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 brain of 345 subjects listening to stories



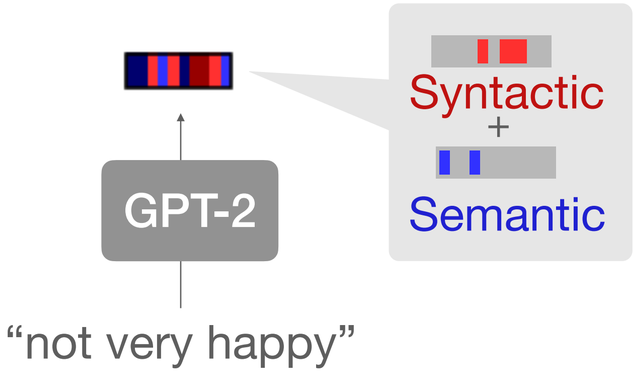


Syntax

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

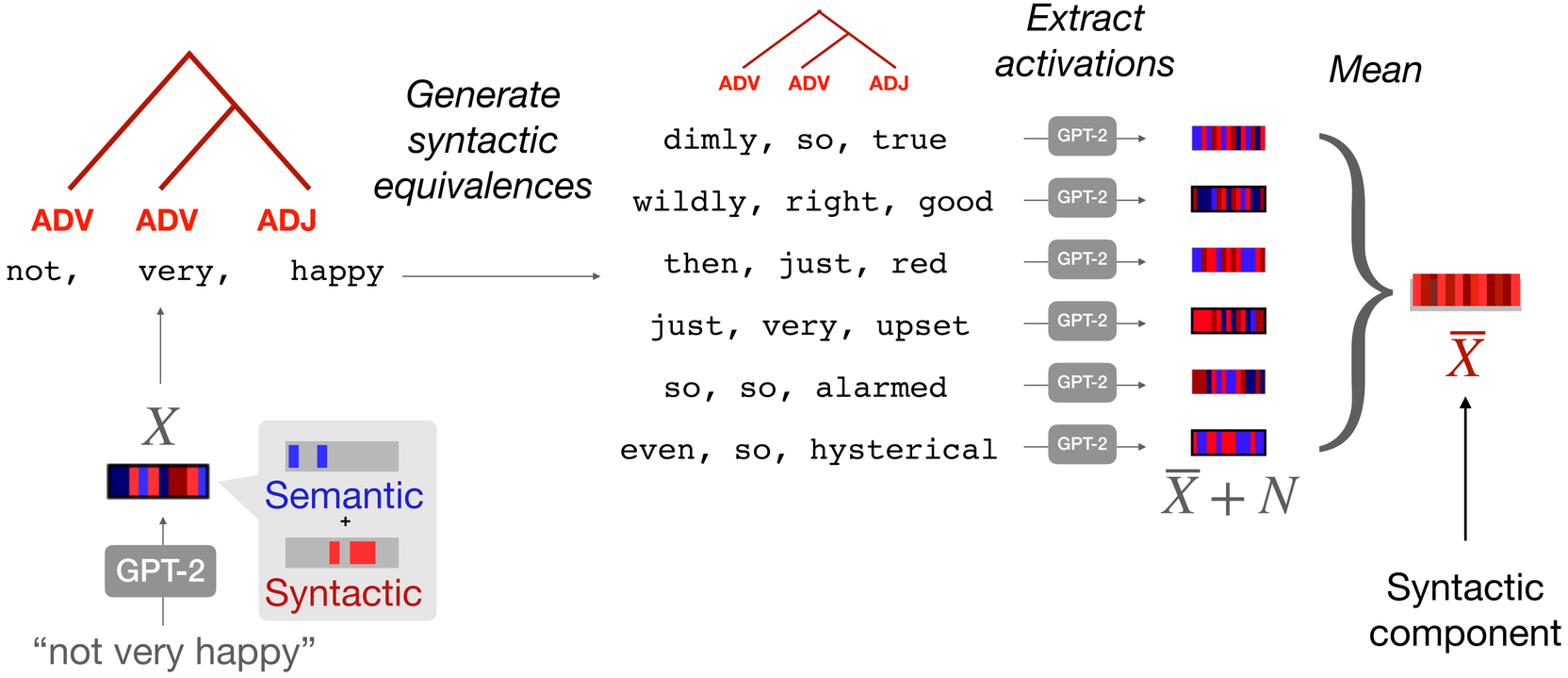
3. We apply our method to the brain of 345 subjects listening to stories

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 brain of 345 subjects listening to stories

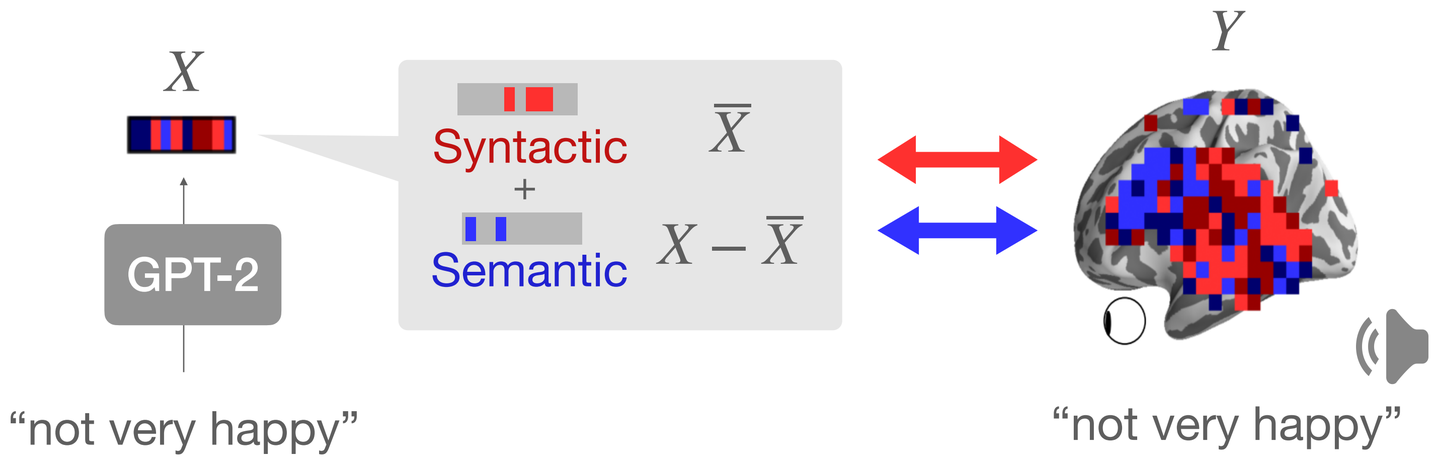


What are the neural bases of syntax and semantics?

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 brain of 345 subjects listening to stories

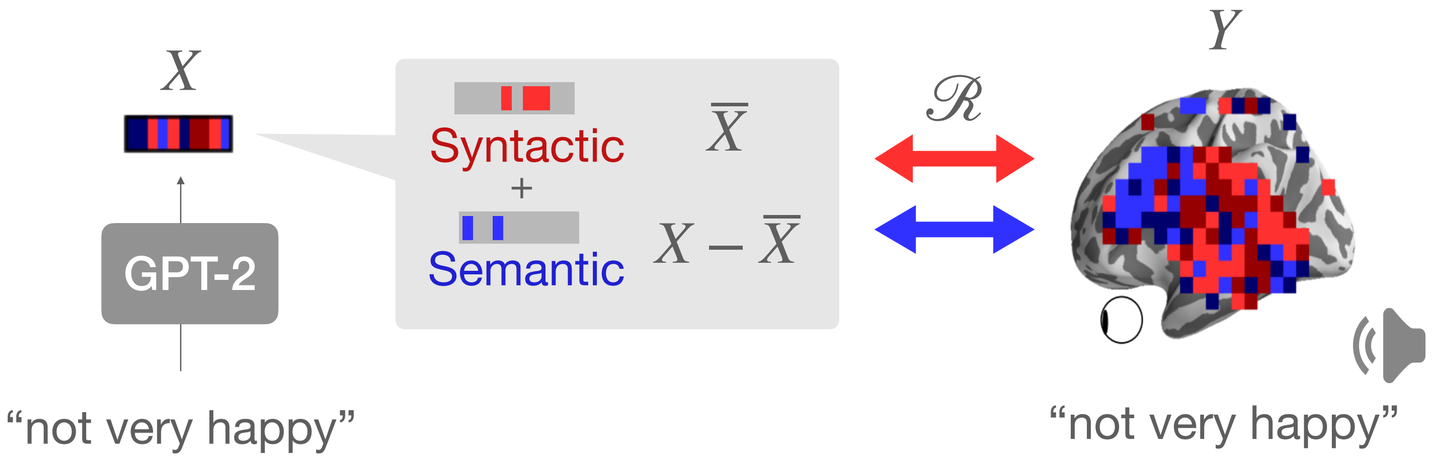


What are the neural bases of syntax and semantics?

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 brain of 345 subjects listening to stories



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

Brain representations:

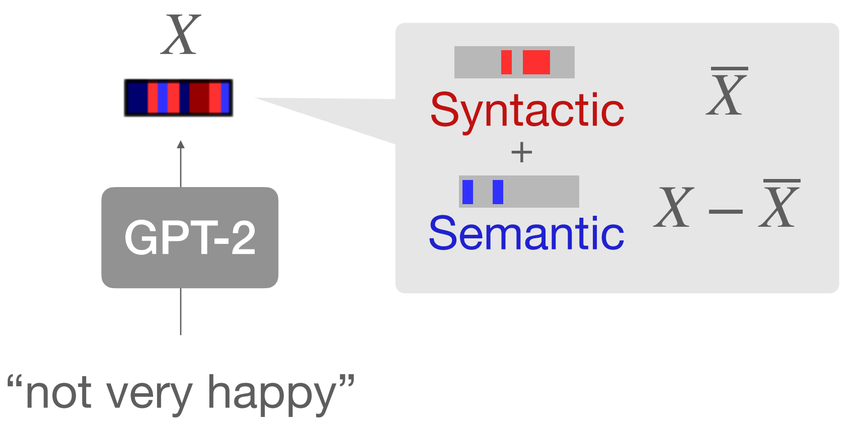
Syntactic	$\mathscr{R}(\overline{X})$
Semantic	$\mathscr{R}(X) - \mathscr{R}(\overline{X})$

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 brain of 345 subjects listening to stories

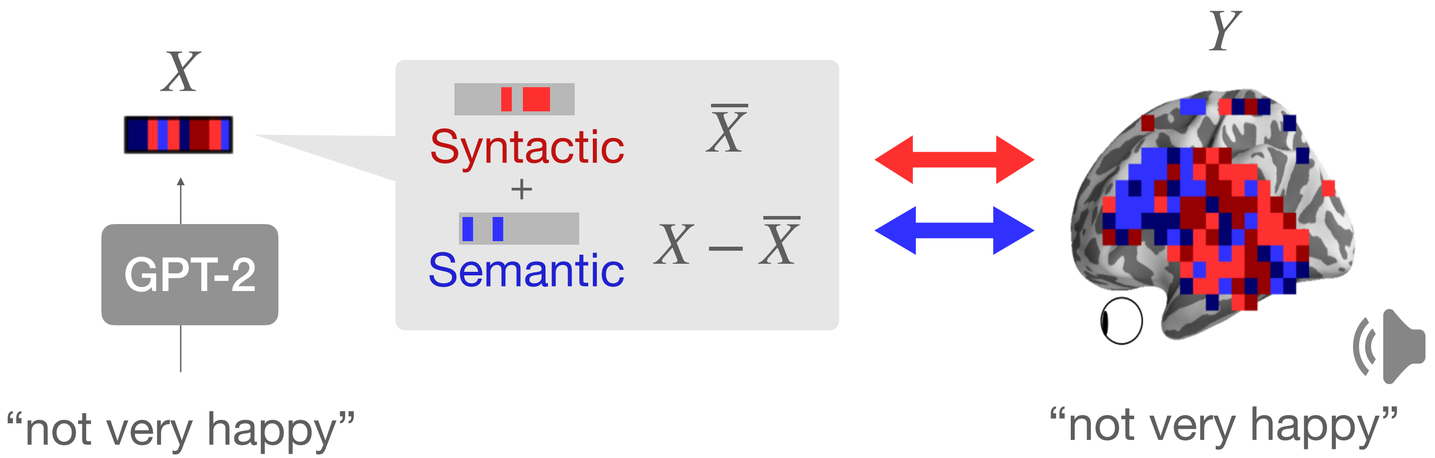


What are the neural bases of syntax and semantics?

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 brain of 345 subjects listening to stories

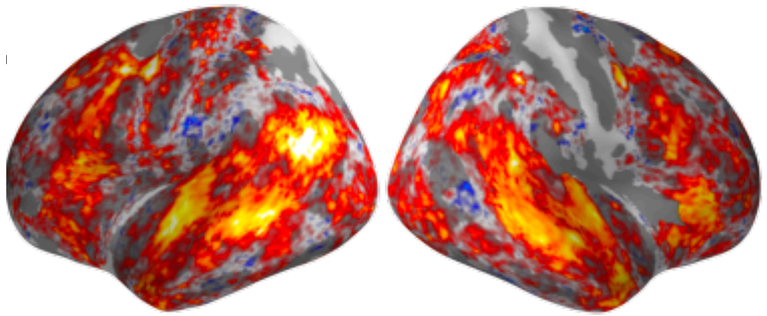


What are the neural bases of syntax and semantics?

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

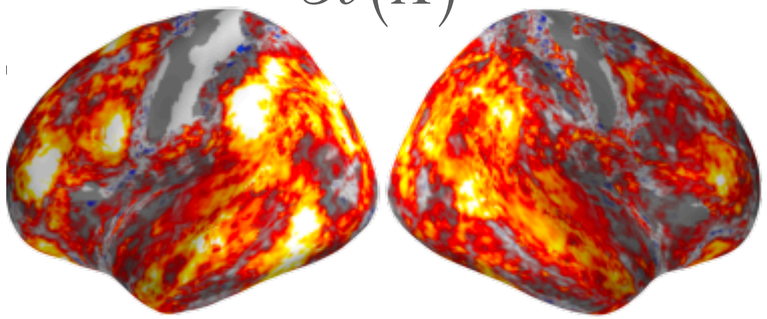
2. We use the disentangled activations to decompose language in the brain

3. We apply our method to the brain of 345 subjects listening to stories



Syntax

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

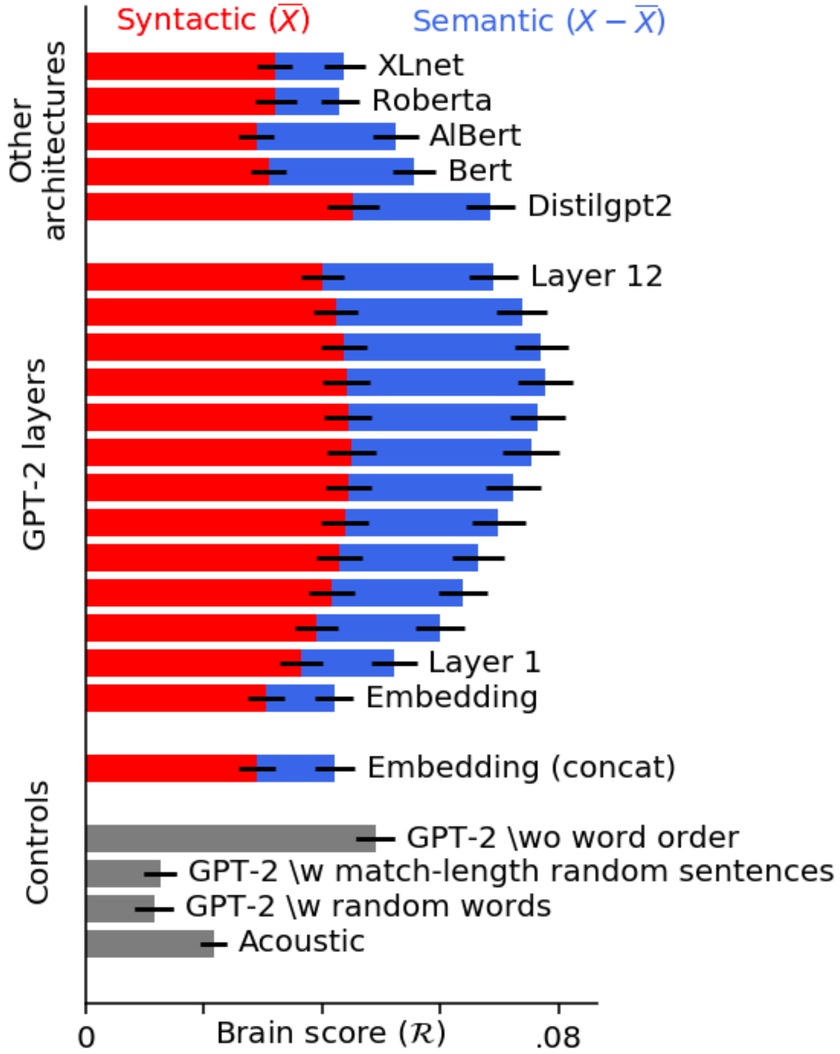


Semantics

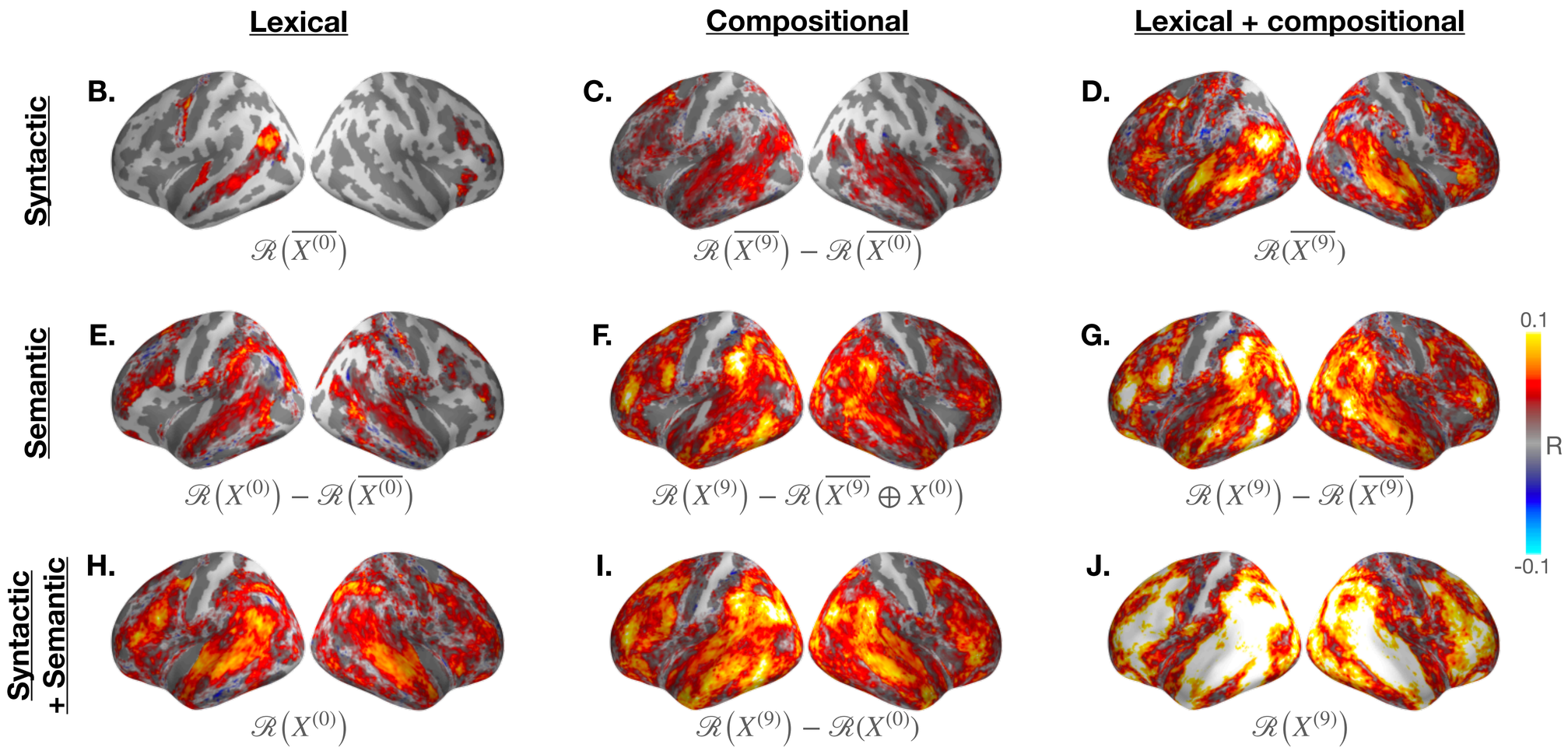
$$\mathscr{R}\left(X\right) - \mathscr{R}\left(\overline{X}\right)$$

*345 subjects listening to ~4 hours of unique audio

Generalisation to other architectures and layers



Fine grained decomposition

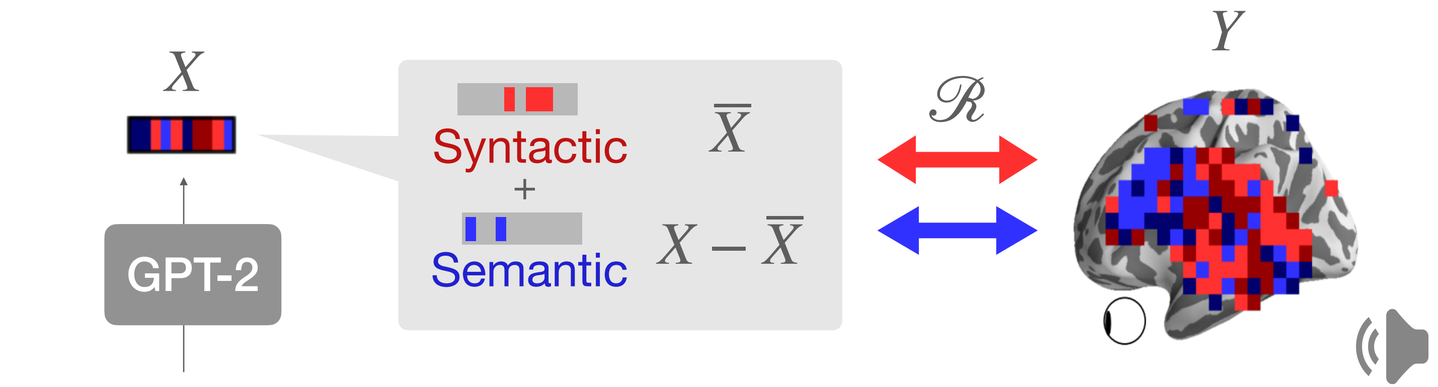


What are the neural bases of syntax and semantics?

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

2. We **linearly map** the disentangled activations **onto brain recordings** of subjects processing language

3. This allows to decompose language processing in the brain of 345 subjects



"not very happy"

"not very happy"

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

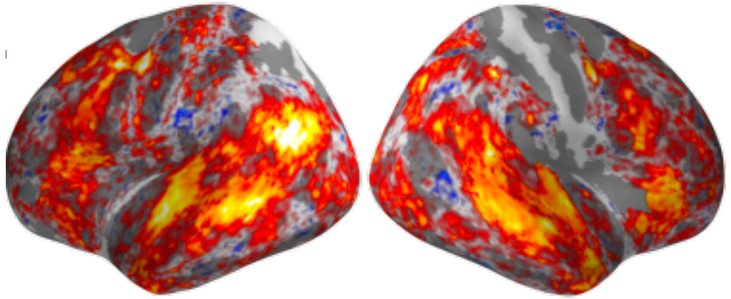
Estimated by Ridge regression on training data

What are the neural bases of syntax and semantics?

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

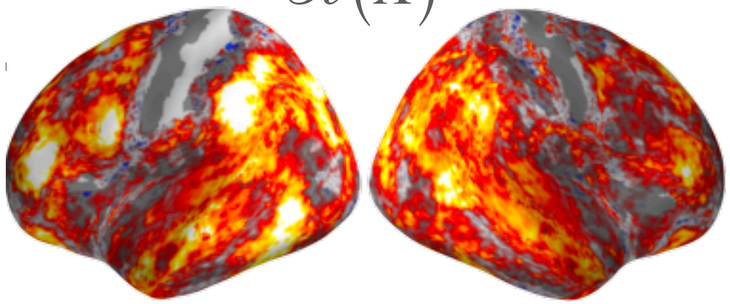
2. We linearly map the disentangled activations onto brain recordings of subjects processing language

3. This allows to decompose language processing in the brain of 345 subjects*



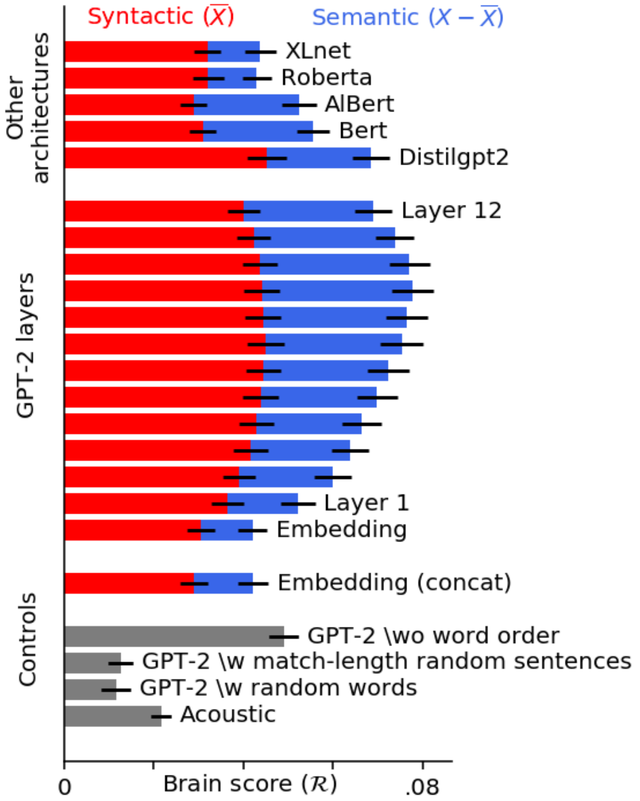
Syntax

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



Semantics

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



*345 subjects listening to ~4 hours of unique audio

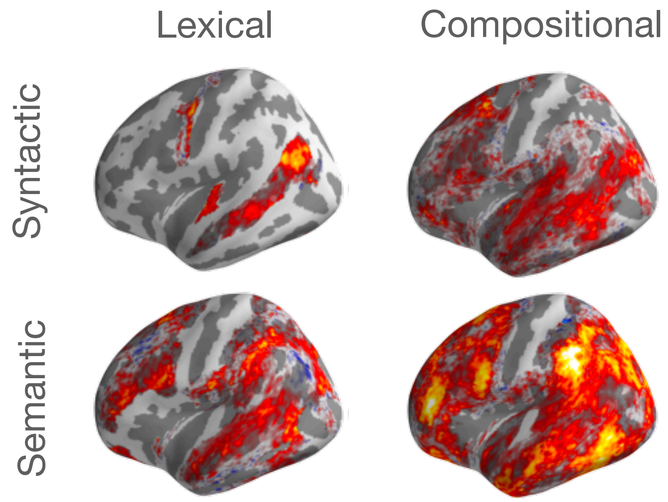
3. Results on the “Narratives” dataset

A- Dataset

Narratives: fMRI data for evaluating models of naturalistic language comprehension

Samuel A. Nastase[1], Yun-Fei Liu[2], Hanna Hillman[1], Asieh Zadbood[1], Liat Hasenfratz[1], Neggin Keshavarzian[1], Janice Chen[2], Christopher J. Honey[2], Yaara Yeshurun[3], Mor Regev[4], Mai Nguyen[1], Claire H. C. Chang[1], Christopher Baldassano[5], Olga Lositsky[6], Erez Simony[7,8], Michael A. Chow[9], Yuan Chang Leong[10], Paula P. Brooks[1], Emily Micciche[11], Gina Choe[1], Ariel Goldstein[1], Tamara Vanderwal[12], Yaroslav O. Halchenko[13], Kenneth A. Norman[1], Uri Hasson[1]

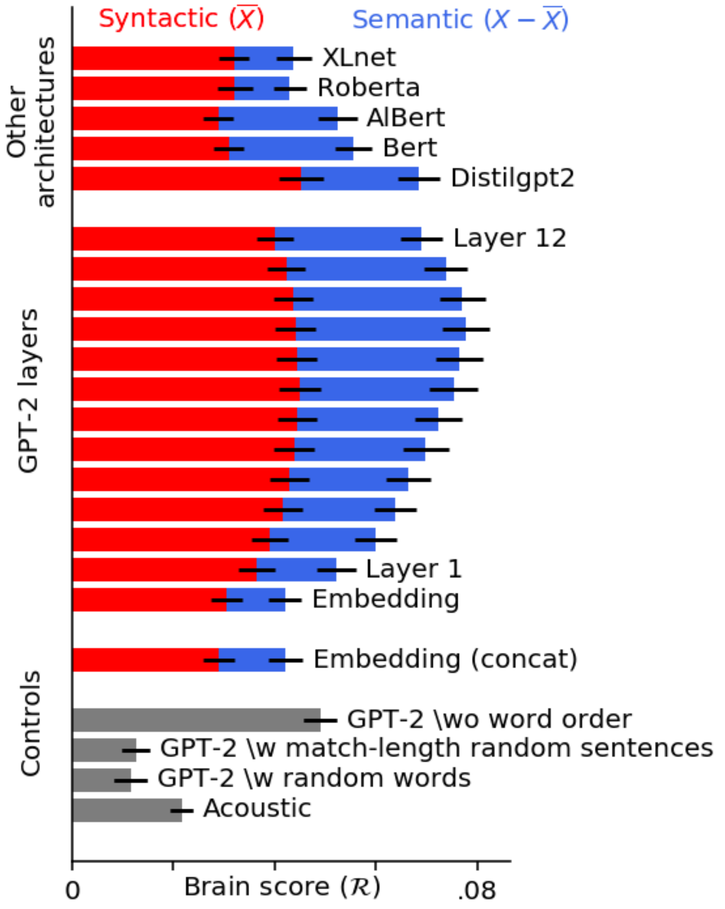
B- Results



Representations:

Syntactic	$\mathcal{R}\left(\overline{X^{(k)}}\right)$
Semantic	$\mathcal{R}\left(X^{(k)}\right) - \mathcal{R}\left(\overline{X^{(k)}}\right)$
Lexical	$\mathcal{R}\left(X^{(0)}\right)$
Compositional	$\mathcal{R}\left(X^{(9)}\right) - \mathcal{R}\left(X^{(0)}\right)$

C- Results generalise to other layers and deep networks



Brain representations:

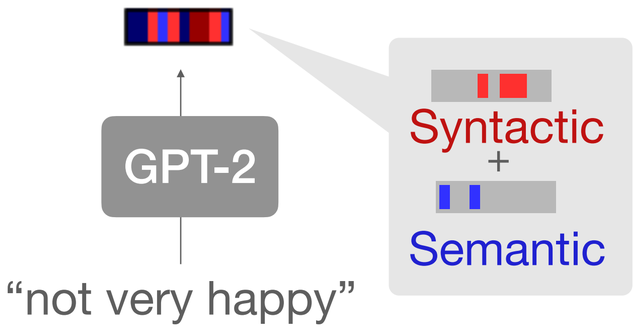
Syntactic	$\mathscr{R}(\overline{X})$
Semantic	$\mathscr{R}(X) - \mathscr{R}(\overline{X})$

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 map the disentangled activations onto brain recordings of subjects processing language

3. This allows to decompose language processing in the brain of 345 subjects

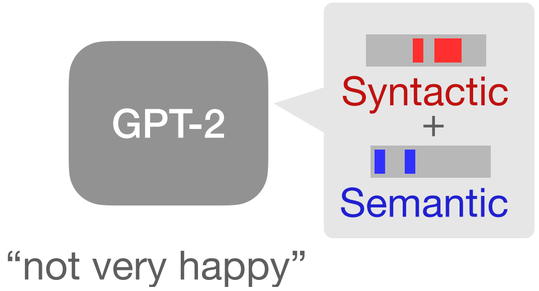


What are the neural bases of syntax and semantics?

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

2. We map the disentangled activations onto brain recordings of subjects processing language

3. This allows to decompose language processing in the brain of 345 subjects

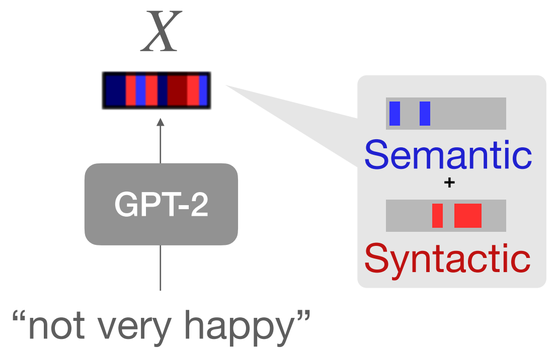


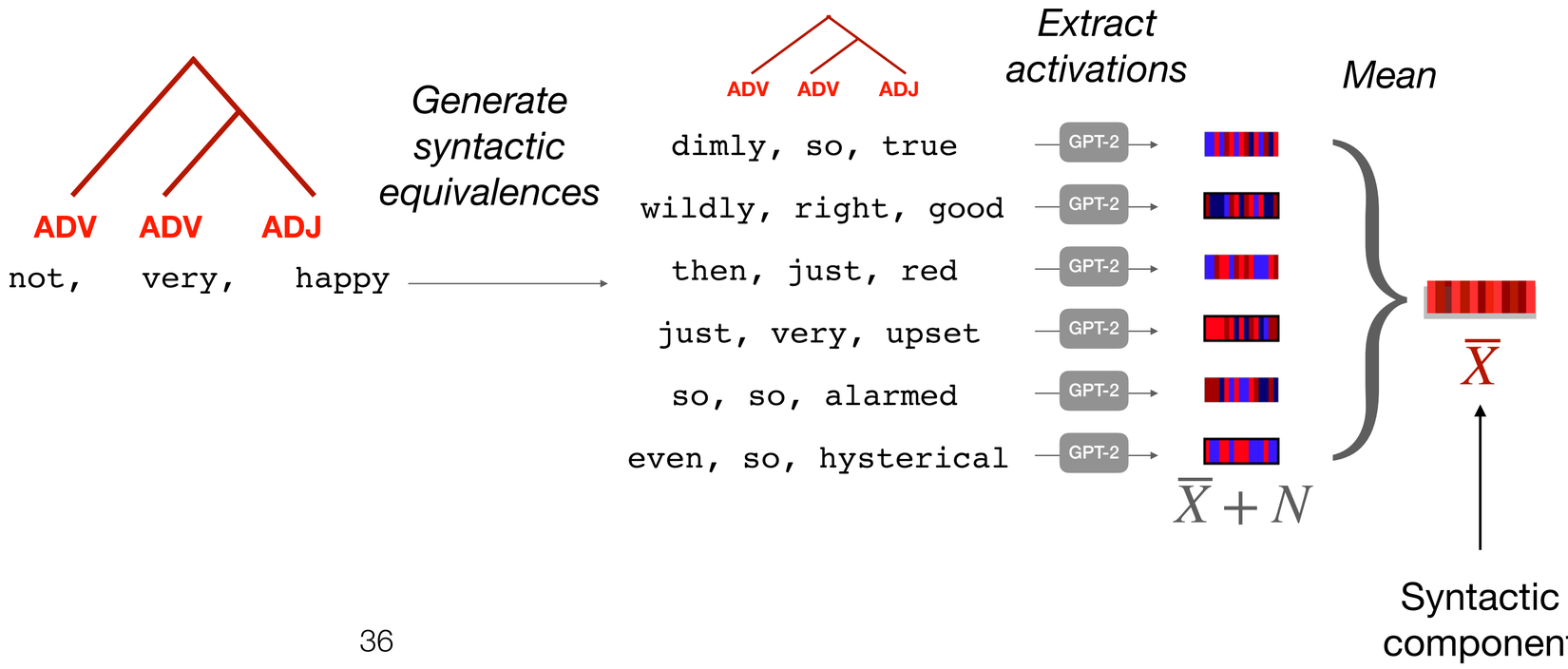
What are the neural bases of syntax and semantics?

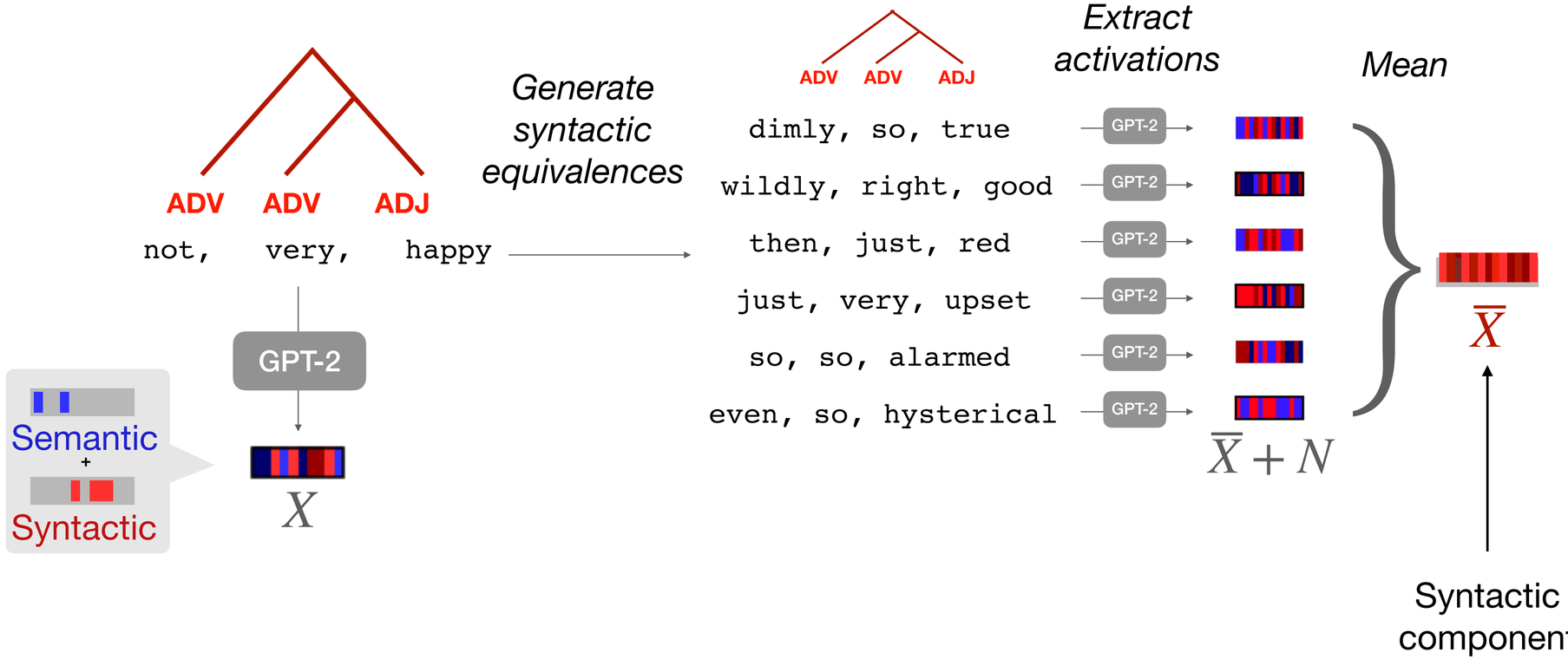
1. We introduce a method to disentangle **syntax** and **semantics** in **deep nets' activations**

2. We map the disentangled activations onto brain recordings of subjects processing language

3. This allows to decompose language processing in the brain of 345 subjects





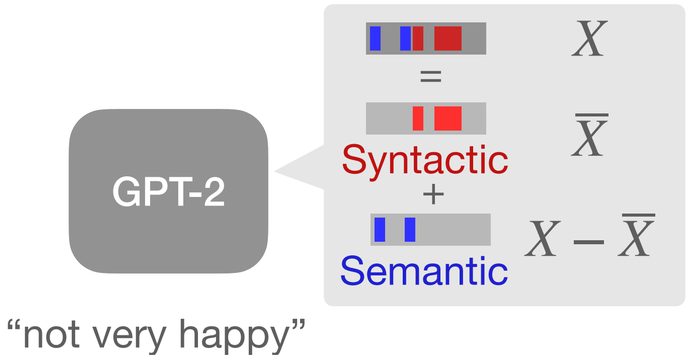

ADV ADV ADJ
not, very, happy
GPT-2
Semantic
+
Syntactic
X
Generate syntactic equivalences
ADV ADV ADJ
dimly, so, true
wildly, right, good
then, just, red
just, very, upset
so, so, alarmed
even, so, hysterical
GPT-2
Extract activations
$\overline{X} + N$
Mean
$\overline{X}$
Syntactic component

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 correlate the split activations with the brain recordings of subjects processing language

3. This allows to decompose language processing in the brain of 345 subjects

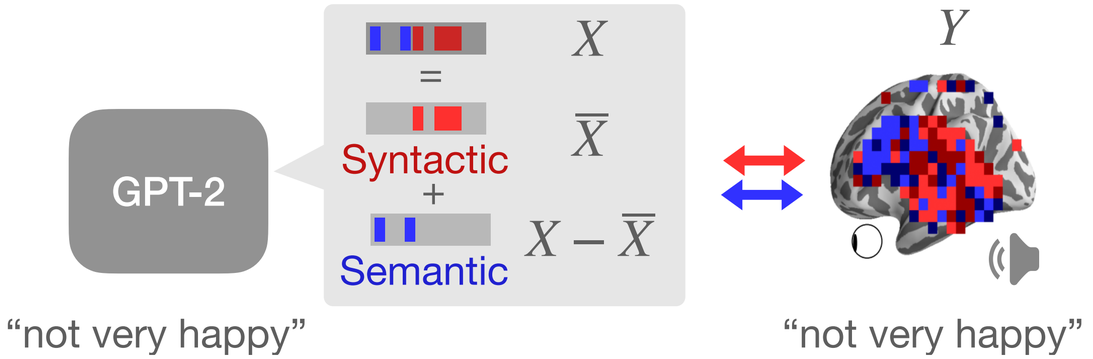


What are the neural bases of syntax and semantics?

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

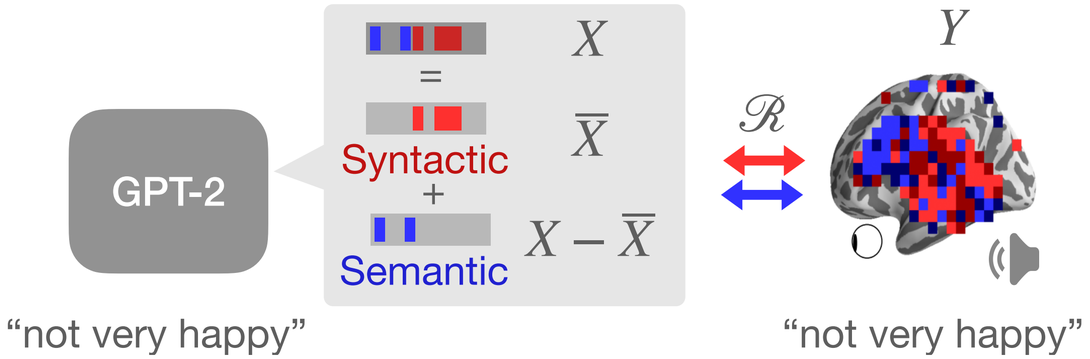
2. We **correlate** the split activations with the **brain recordings** of subjects processing language

3. This allows to decompose language processing in the brain of 345 subjects



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 **correlate** the split activations with the **brain recordings** of subjects processing language

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

Estimated by Ridge regression on training data

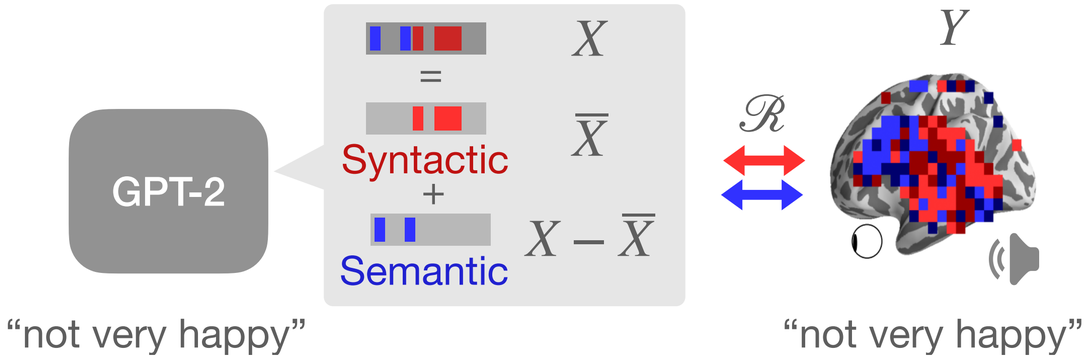
3. This allows to decompose language processing in the brain of 345 subjects

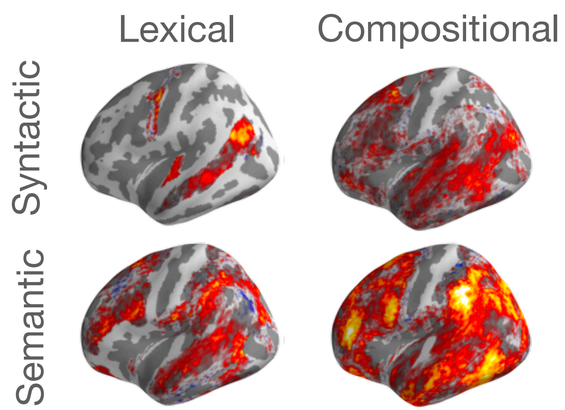
What are the neural bases of syntax and semantics?

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

2. We correlate the split activations with the brain recordings of subjects processing language

3. This allows to **decompose language processing** in the **brain of 345 subjects**



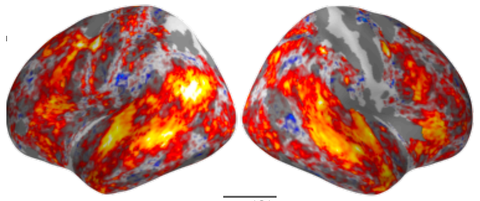


x345 subjects

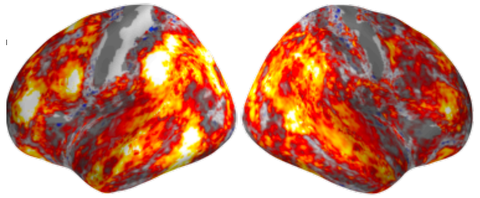
Brain representations:

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

3. Experiment on the “Narratives” dataset (n=345)



$\mathscr{R}(\overline{X^{(9)}})$



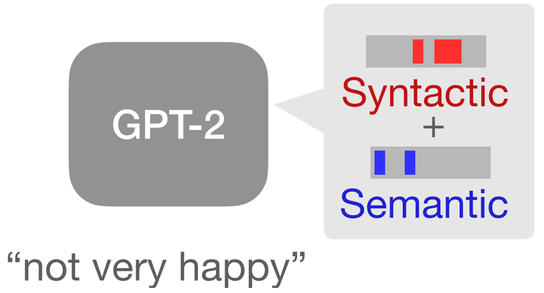
$\mathscr{R}\left(X^{(9)}\right) - \mathscr{R}\left(\overline{X^{(9)}}\right)$

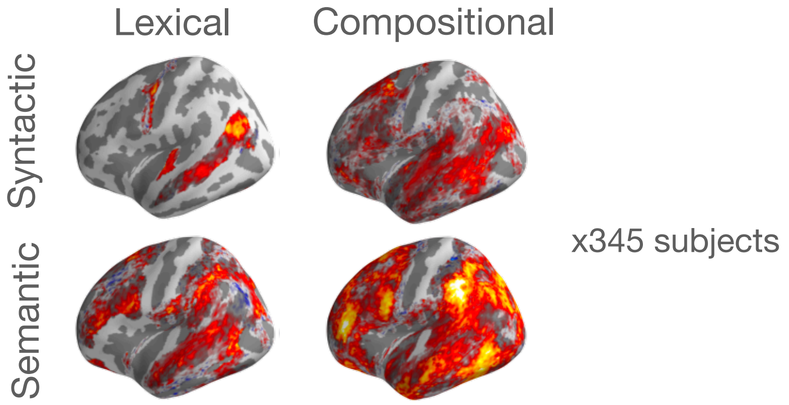
What are the neural bases of syntax and semantics?

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

2. We correlate the split activations with the brain recordings of subjects processing language

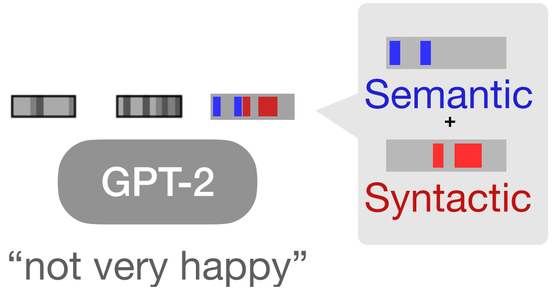
3. This allows to decompose language processing in the brain of 345 subjects





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 the distributed representations of the brain

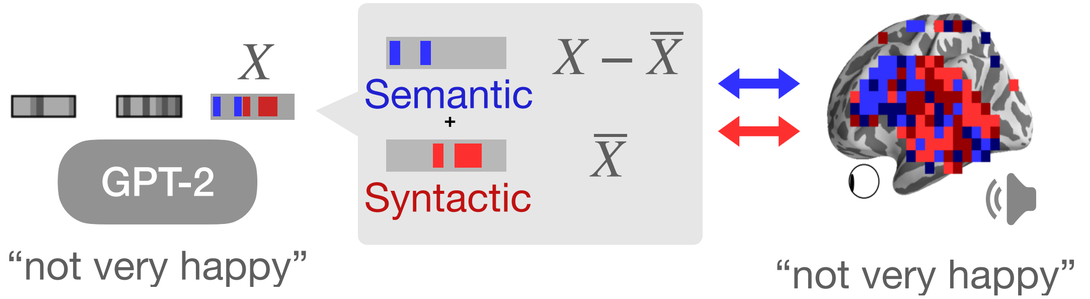
3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects

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 the distributed representations of the brain

3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects

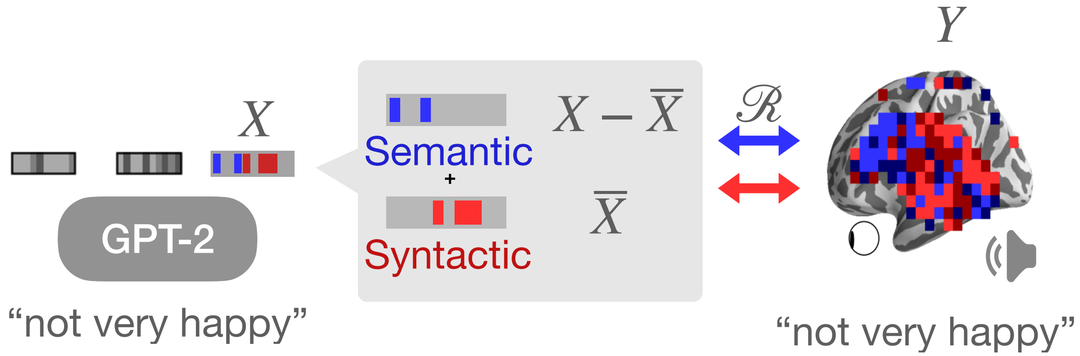


What are the neural bases of syntax and semantics?

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

2. We use the disentangled activations to decompose the distributed representations of the brain

3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects



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

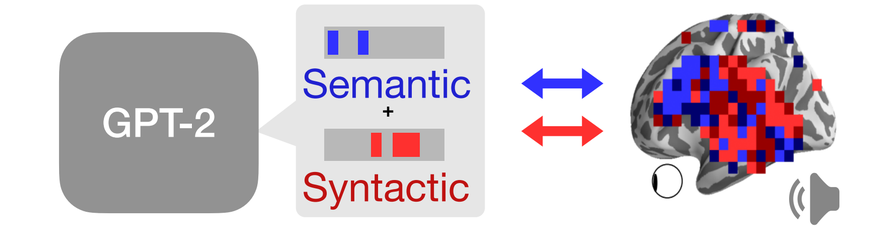
Estimated by Ridge regression on training data

What are the neural bases of syntax and semantics?

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

2. Then, we correlate these syntactic and semantic activations with brain recordings of subjects processing language

3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects



"not very happy" "not very happy"

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

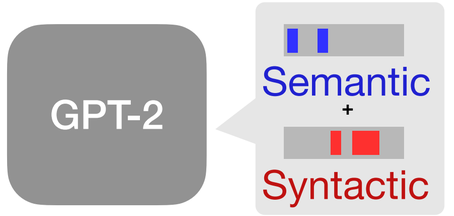
Estimated by Ridge regression on training data

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 the distributed representations of the brain

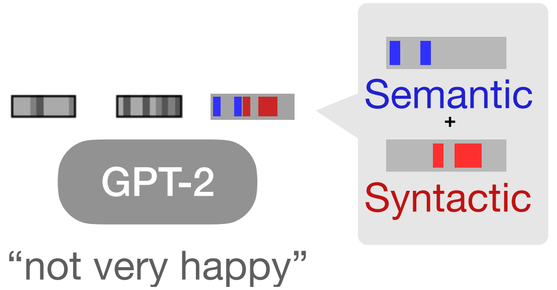
3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects



"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**

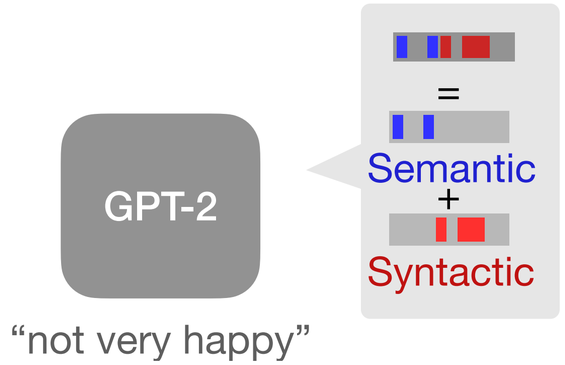


2. We use the disentangled activations to decompose the distributed representations of the brain

3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects

What are the neural bases of syntax and semantics?

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

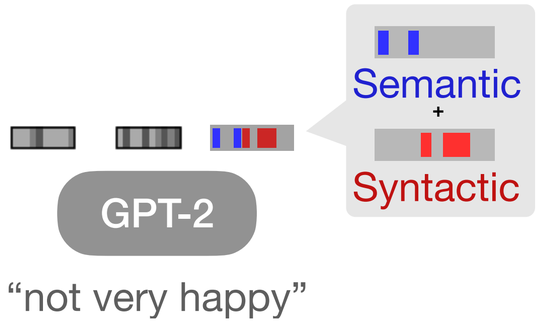


2. We use the disentangled activations to decompose the distributed representations of the brain

3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects

What are the neural bases of syntax and semantics?

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



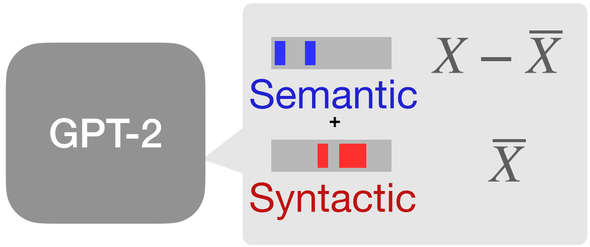
2. We use the disentangled activations to decompose the distributed representations of the brain

3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects



What are the neural bases of syntax and semantics?

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



"not very happy"

2. We use the disentangled activations to decompose the distributed representations of the brain

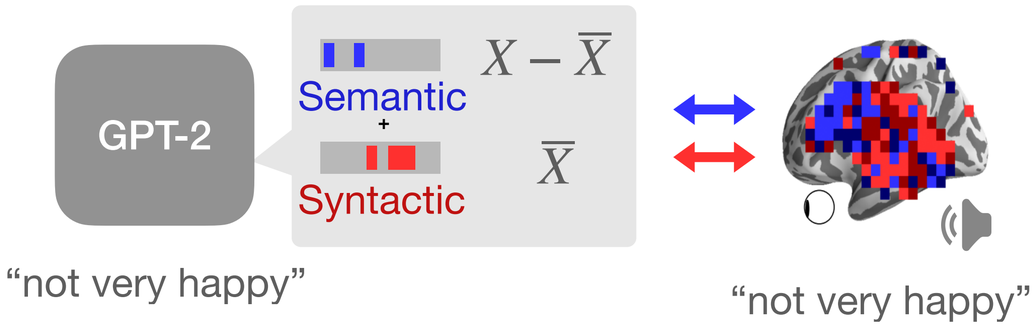
3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects

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 the distributed representations of the brain

3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects

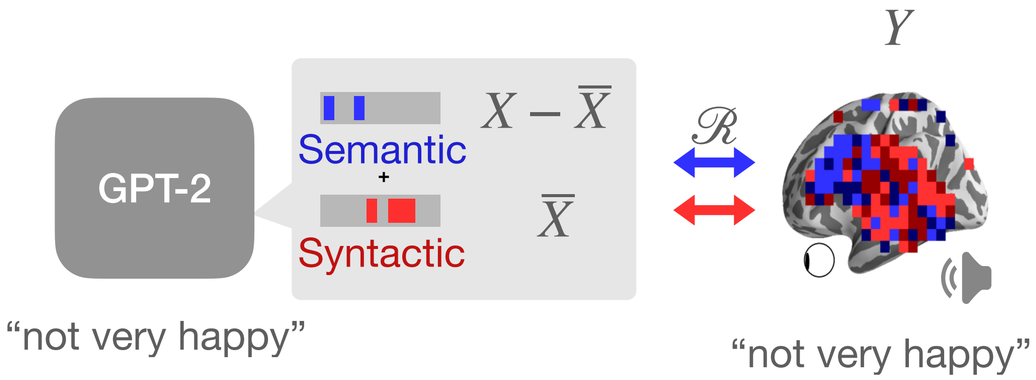


What are the neural bases of syntax and semantics?

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

2. Then, we correlate these syntactic and semantic activations with brain recordings of subjects processing language

3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects



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

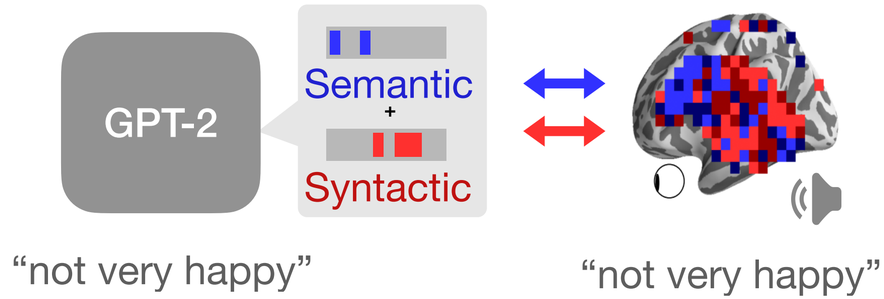
Estimated by Ridge regression on training data

What are the neural bases of syntax and semantics?

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

2. Then, we correlate these syntactic and semantic activations with brain recordings of subjects processing language

3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects

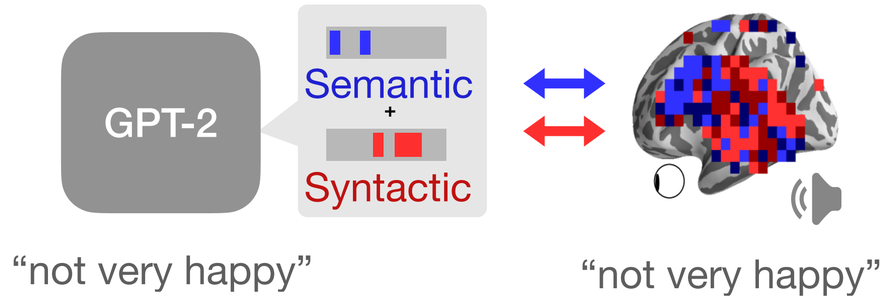


What are the neural bases of syntax and semantics?

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

2. Then, we correlate these syntactic and semantic activations with brain recordings of subjects processing language

3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects



$X^{(k)}$

GPT-2
(layer k)

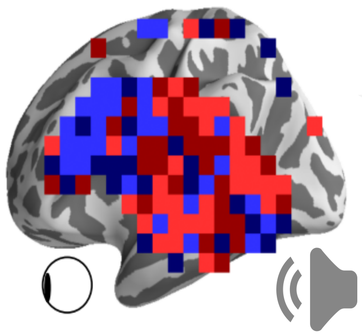
"not very happy"

Network representations:

Syntactic	$\overline{X^{(k)}}$
Semantic	$X^{(k)} - \overline{X^{(k)}}$
Lexical	$X^{(0)}$
Compositional	$X^{(9)} - V^T . X^{(0)}$

$\mathscr{R}$

Y



"not very happy"

Brain representations:

Syntactic	$\mathscr{R}\left(\overline{X^{(k)}}\right)$
Semantic	$\mathscr{R}\left(X^{(k)}\right) - \mathscr{R}\left(\overline{X^{(k)}}\right)$
Lexical	$\mathscr{R}\left(X^{(0)}\right)$
Compositional	$\mathscr{R}\left(X^{(9)}\right) - \mathscr{R}\left(X^{(0)}\right)$

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

Estimated by Ridge regression on training data

2. Method to decompose brain activity

$X^{(k)}$

GPT-2
(layer k)

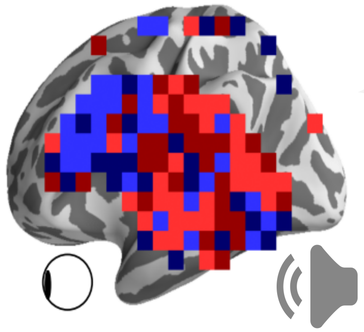
"not very happy"

Network representations:

Syntactic $\overline{X^{(k)}}$

$\mathscr{R}$

Y



"not very happy"

Brain representations:

Syntactic $\mathscr{R}\left(\overline{X^{(k)}}\right)$

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

Estimated by Ridge regression on training data

2. Method to decompose brain activity

$X^{(k)}$

GPT-2
(layer k)

"not very happy"

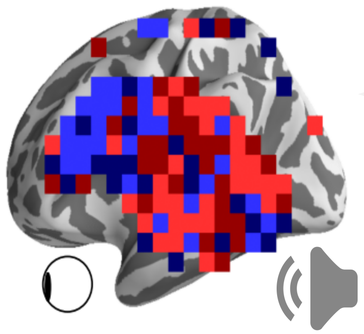
Network representations:

Syntactic $\overline{X^{(k)}}$

Semantic $X^{(k)} - \overline{X^{(k)}}$

$\mathscr{R}$

Y



"not very happy"

Brain representations:

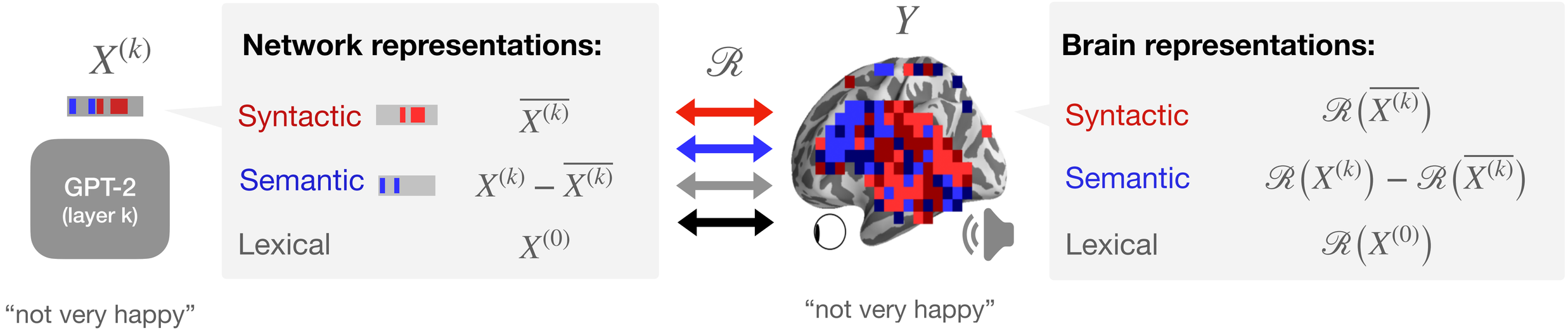
Syntactic $\mathscr{R}\left(\overline{X^{(k)}}\right)$

Semantic $\mathscr{R}\left(X^{(k)}\right) - \mathscr{R}\left(\overline{X^{(k)}}\right)$

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

Estimated by Ridge regression on training data

2. Method to decompose brain activity



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

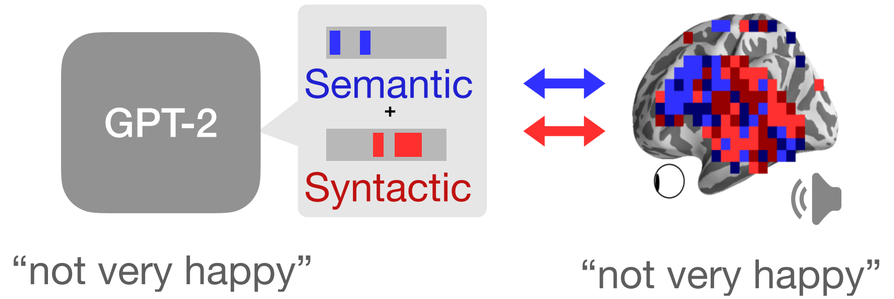
Estimated by Ridge regression on training data

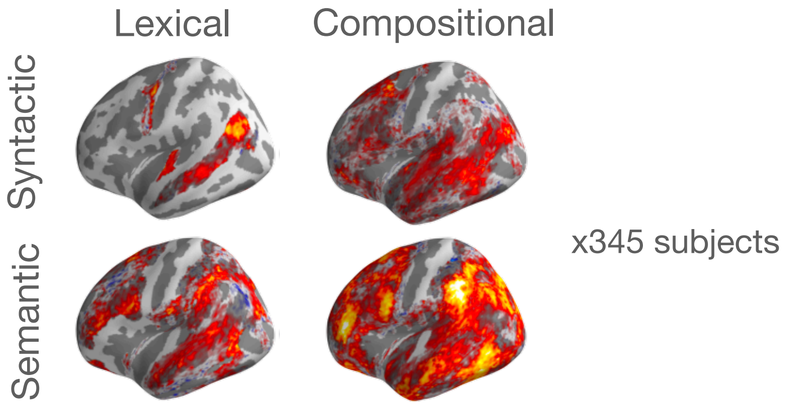
What are the neural bases of syntax and semantics?

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

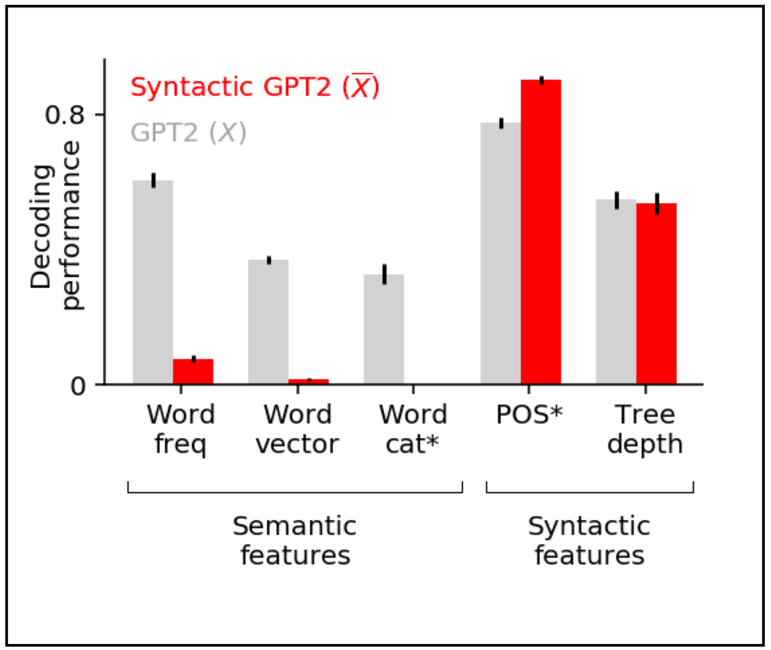
2. We use the disentangled activations to decompose the distributed representations of the brain

3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects





1. Method to extract syntactic and semantic activations

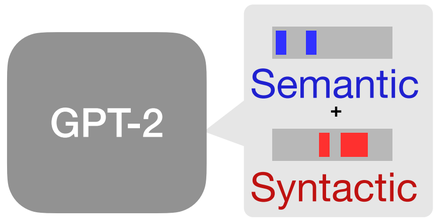




1. We propose a method to disentangle syntax and semantics in deep nets' activations

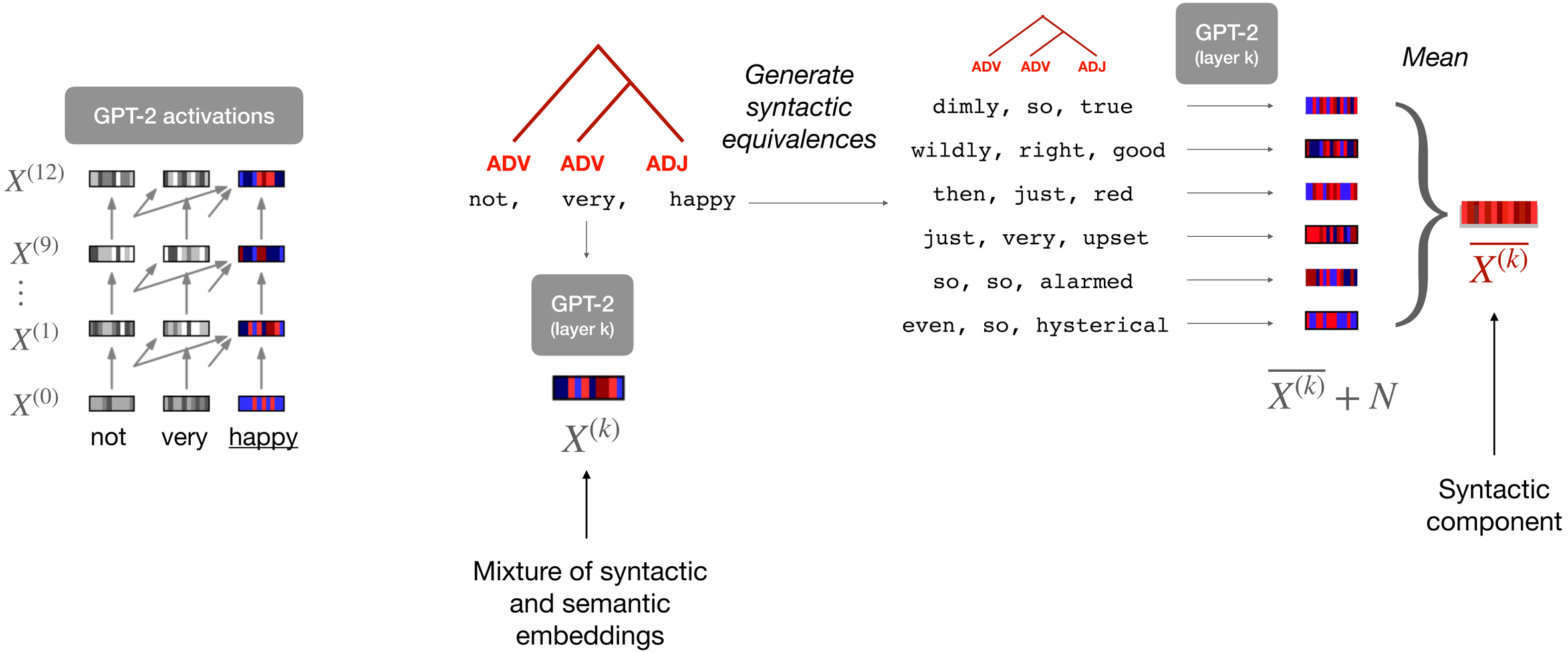
2. We use the disentangled activations to decompose the distributed representations of the brain

3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects

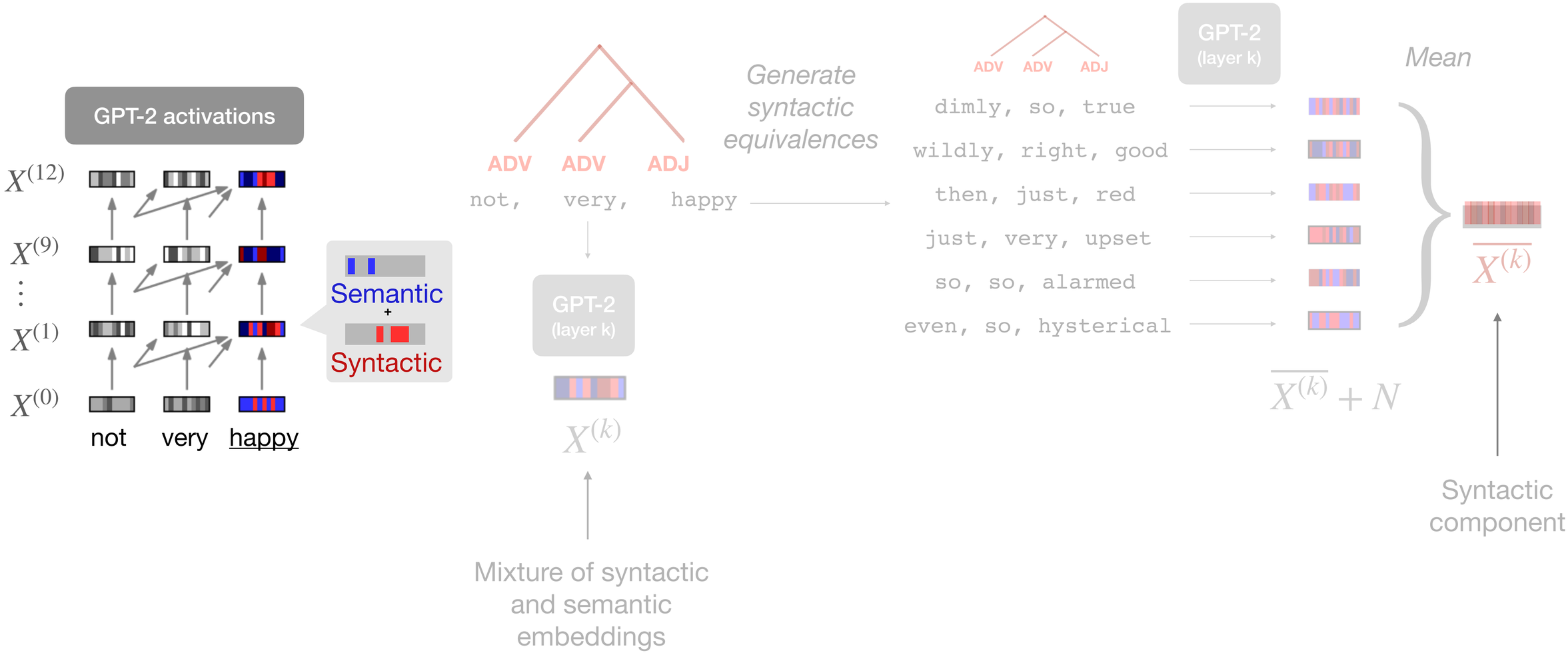


"not very happy"

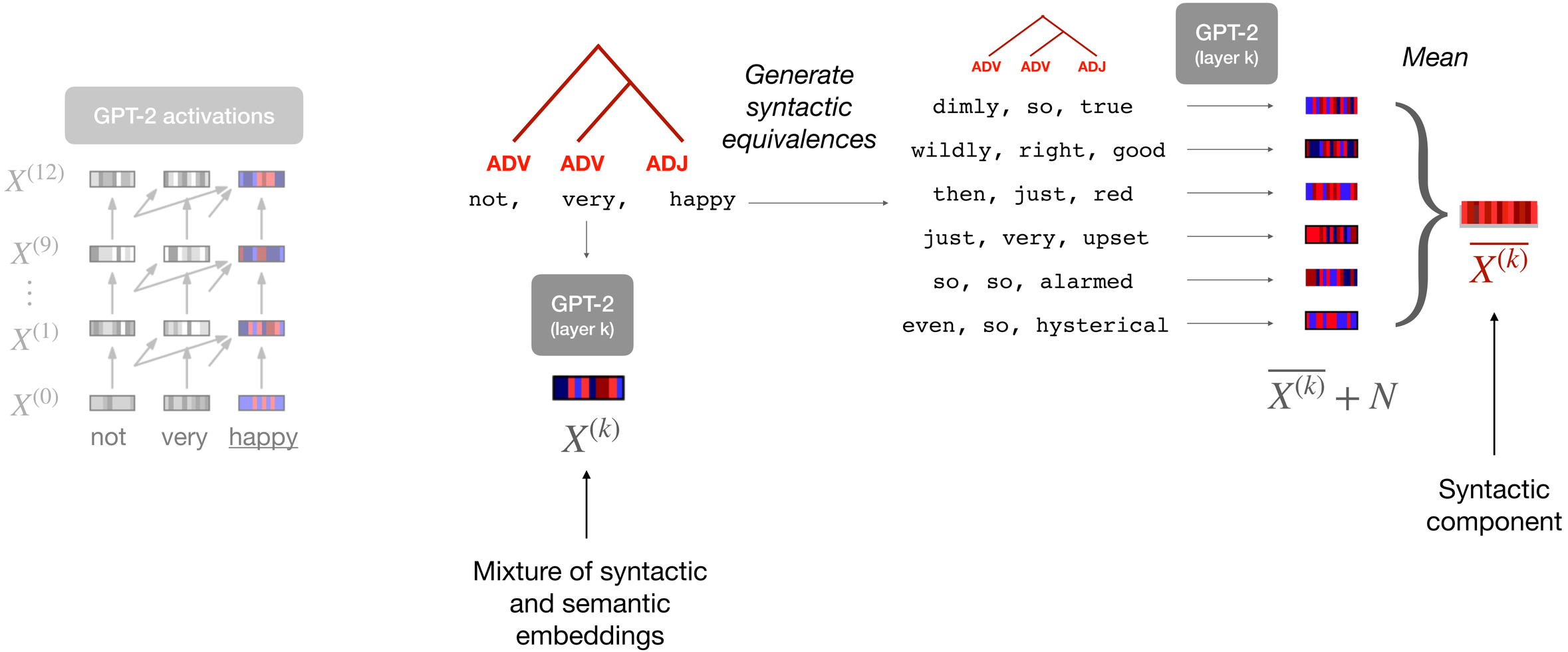
1. Method to extract syntactic and semantic activations



1. Method to extract syntactic and semantic activations



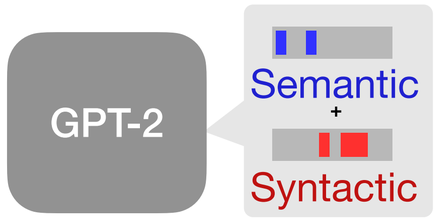
1. Method to extract syntactic and semantic activations



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

2. We use the disentangled activations to decompose the distributed representations of the brain

3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects

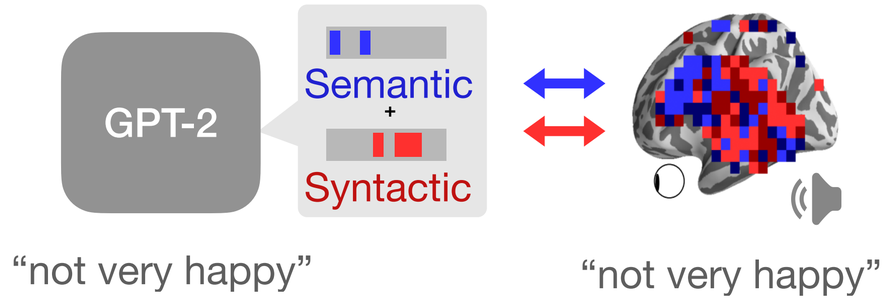


"not very happy"

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

2. We use the disentangled activations to decompose the distributed representations of the brain

3. We apply our method to disentangle syntax and semantics in the brain of 345 subjects



2. Method to decompose brain activity

$X^{(k)}$

GPT-2
(layer k)

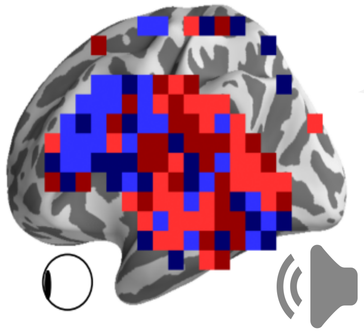
"not very happy"

Network representations:

Syntactic $\overline{X^{(k)}}$

$\mathscr{R}$

Y



"not very happy"

Brain representations:

Syntactic $\mathscr{R}\left(\overline{X^{(k)}}\right)$

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

Estimated by Ridge regression on training data

2. Method to decompose brain activity

$X^{(k)}$

GPT-2
(layer k)

"not very happy"

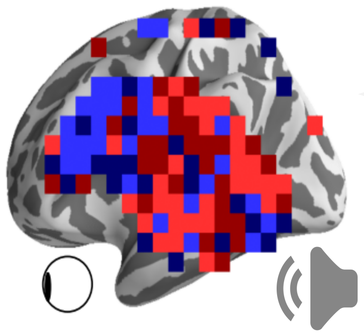
Network representations:

Syntactic $\overline{X^{(k)}}$

Semantic $X^{(k)} - \overline{X^{(k)}}$

$\mathscr{R}$

Y



"not very happy"

Brain representations:

Syntactic $\mathscr{R}\left(\overline{X^{(k)}}\right)$

Semantic $\mathscr{R}\left(X^{(k)}\right) - \mathscr{R}\left(\overline{X^{(k)}}\right)$

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

Estimated by Ridge regression on training data

2. Method to decompose brain activity

$X^{(k)}$

GPT-2
(layer k)

"not very happy"

Network representations:

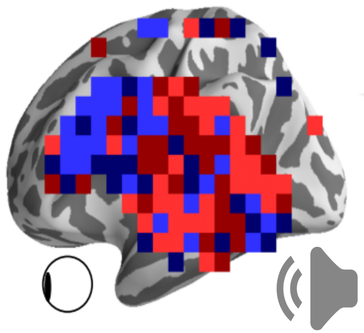
Syntactic $\overline{X^{(k)}}$

Semantic $X^{(k)} - \overline{X^{(k)}}$

Lexical $X^{(0)}$

$\mathscr{R}$

Y



"not very happy"

Brain representations:

Syntactic $\mathscr{R}\left(\overline{X^{(k)}}\right)$

Semantic $\mathscr{R}\left(X^{(k)}\right) - \mathscr{R}\left(\overline{X^{(k)}}\right)$

Lexical $\mathscr{R}\left(X^{(0)}\right)$

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

Estimated by Ridge regression on training data

2. Method to decompose brain activity

$X^{(k)}$

GPT-2
(layer k)

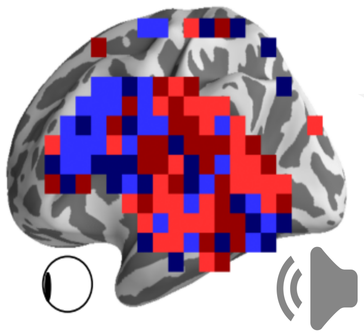
"not very happy"

Network representations:

Syntactic	$\overline{X^{(k)}}$
Semantic	$X^{(k)} - \overline{X^{(k)}}$
Lexical	$X^{(0)}$
Compositional	$X^{(9)} - V^T . X^{(0)}$

$\mathscr{R}$

Y



"not very happy"

Brain representations:

Syntactic	$\mathscr{R}\left(\overline{X^{(k)}}\right)$
Semantic	$\mathscr{R}\left(X^{(k)}\right) - \mathscr{R}\left(\overline{X^{(k)}}\right)$
Lexical	$\mathscr{R}\left(X^{(0)}\right)$
Compositional	$\mathscr{R}\left(X^{(9)}\right) - \mathscr{R}\left(X^{(0)}\right)$

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

Estimated by Ridge regression on training data