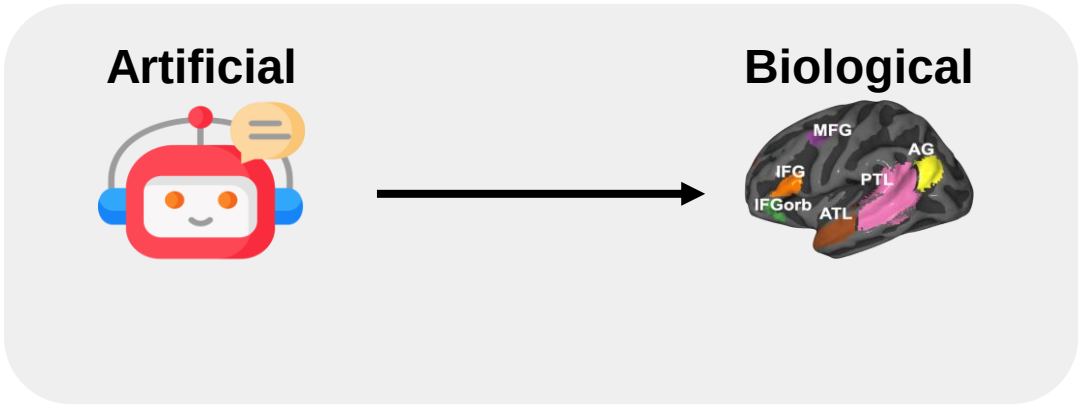
Neuro-Computational Models of Language Comprehension: characterizing similarities and differences between language processing in brains and language models

Subba Reddy Oota[1]

[1]Inria Bordeaux, France

Mnemosyne (IMN-Team7)

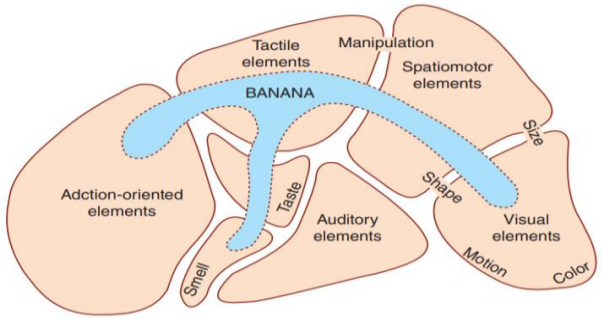


Outline

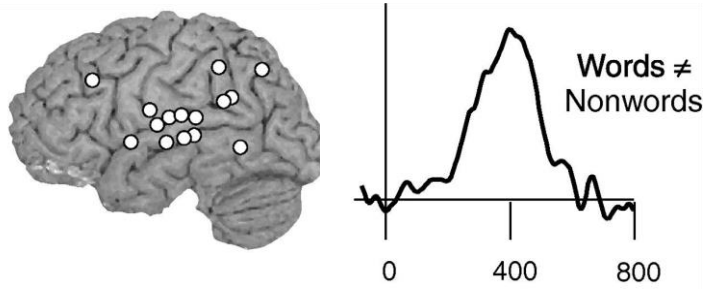
- Introduction to neural basis of language comprehesion
- Deep neural networks and brain alignment: brain encoding and decoding
- Research questions
- Implications to Neuro-AI

Mechanistic understanding of language processing in the brain: four big questions

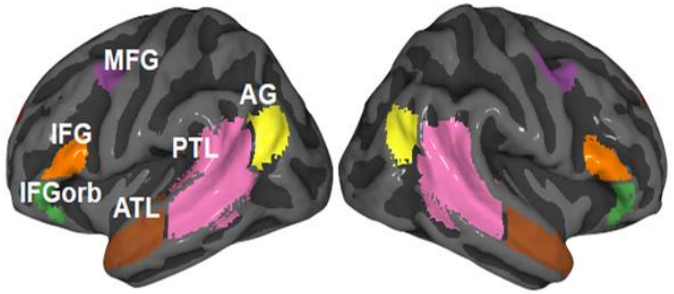
What



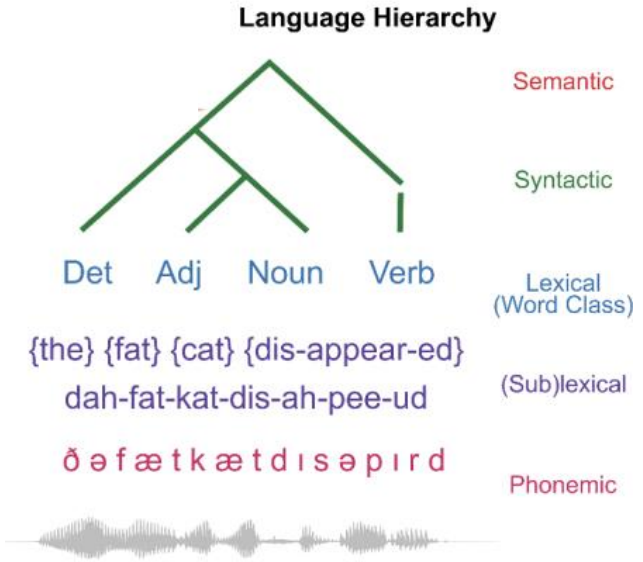
When



Where

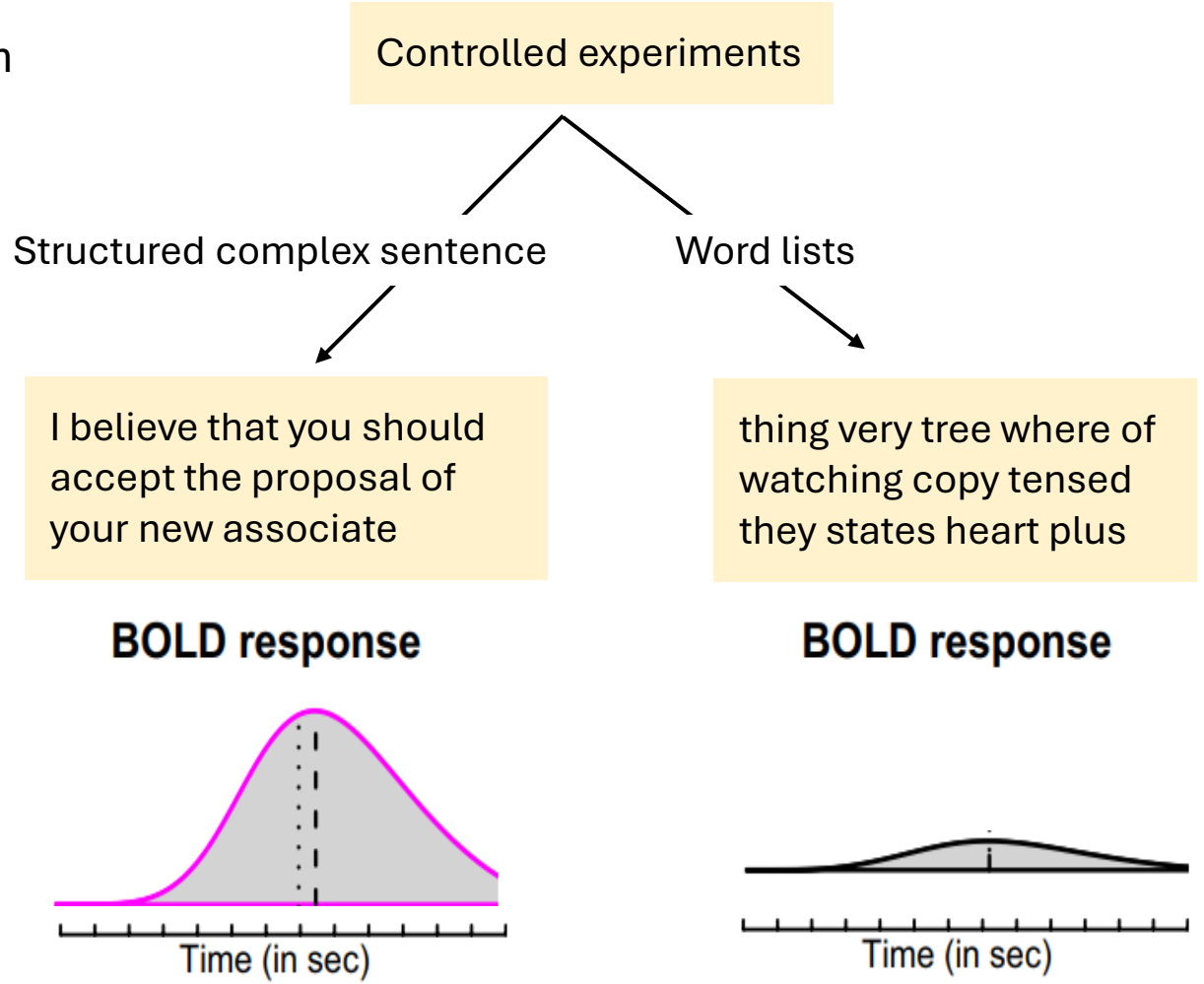


How



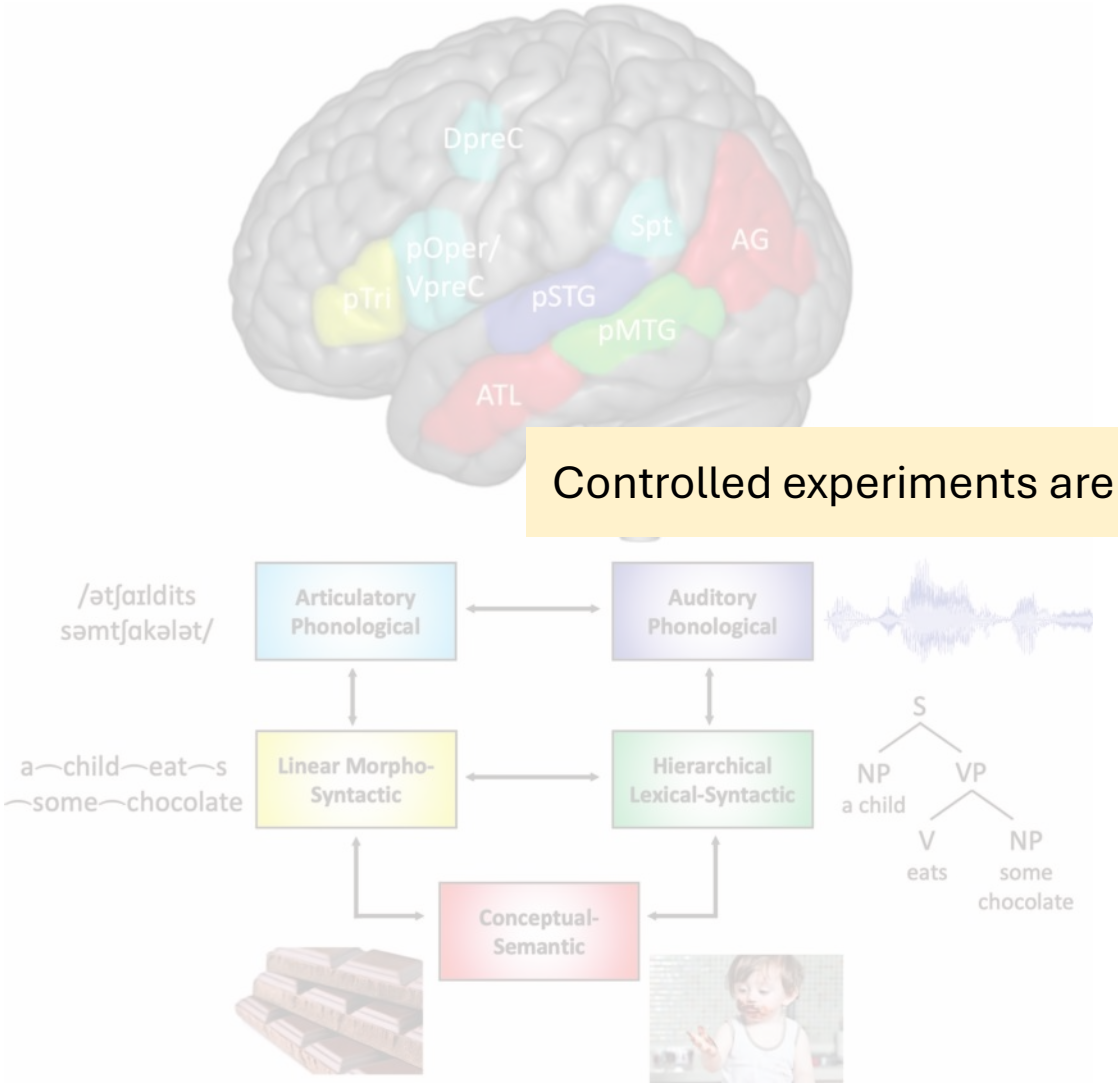
Typical studies of language processing with controlled experiments

- How the human brain computes and encodes syntactic structures?
  - **Syntax:** how do words structurally combine to form sentences and meaning?



Pallier et al. 2011

Language organization in the brain



Controlled experiments are task-based and not ecological

uage at different features
Hierarchical syntactic information occurs in the cortical zone situated between auditory-phonological and semantic zones.

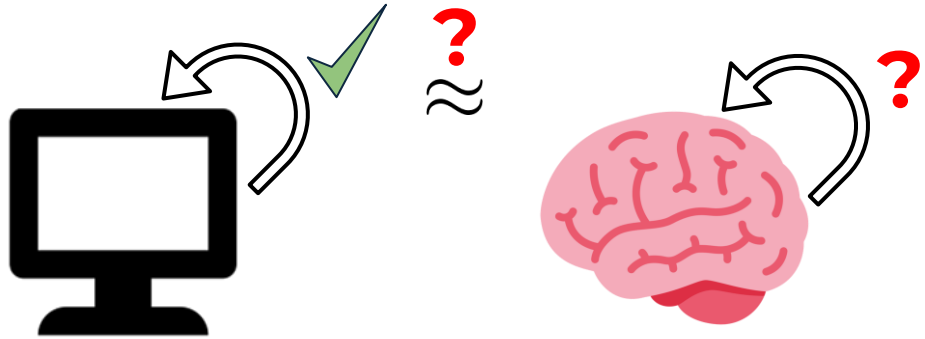
Matchin & Hickok et al. 2022

Increasingly available open source ecological stimuli datasets

Dataset	Modality	Subj	1-TR	# TRs
Full-Moth-Radio-Hour	Listening	8	2.0045s	9932
Subset-Moth-Radio-Hour	Reading	6	2.0045s	4028
Subset-Moth-Radio-Hour	Listening	6	2.0045s	4028
Narratives (21^{st}-Year)	Listening	18	1.5s	2250
Harry-Potter	Reading	8	2s	1211

With advancement of **ecological stimuli datasets** and **open source language models**, recent studies looked at interesting open questions?

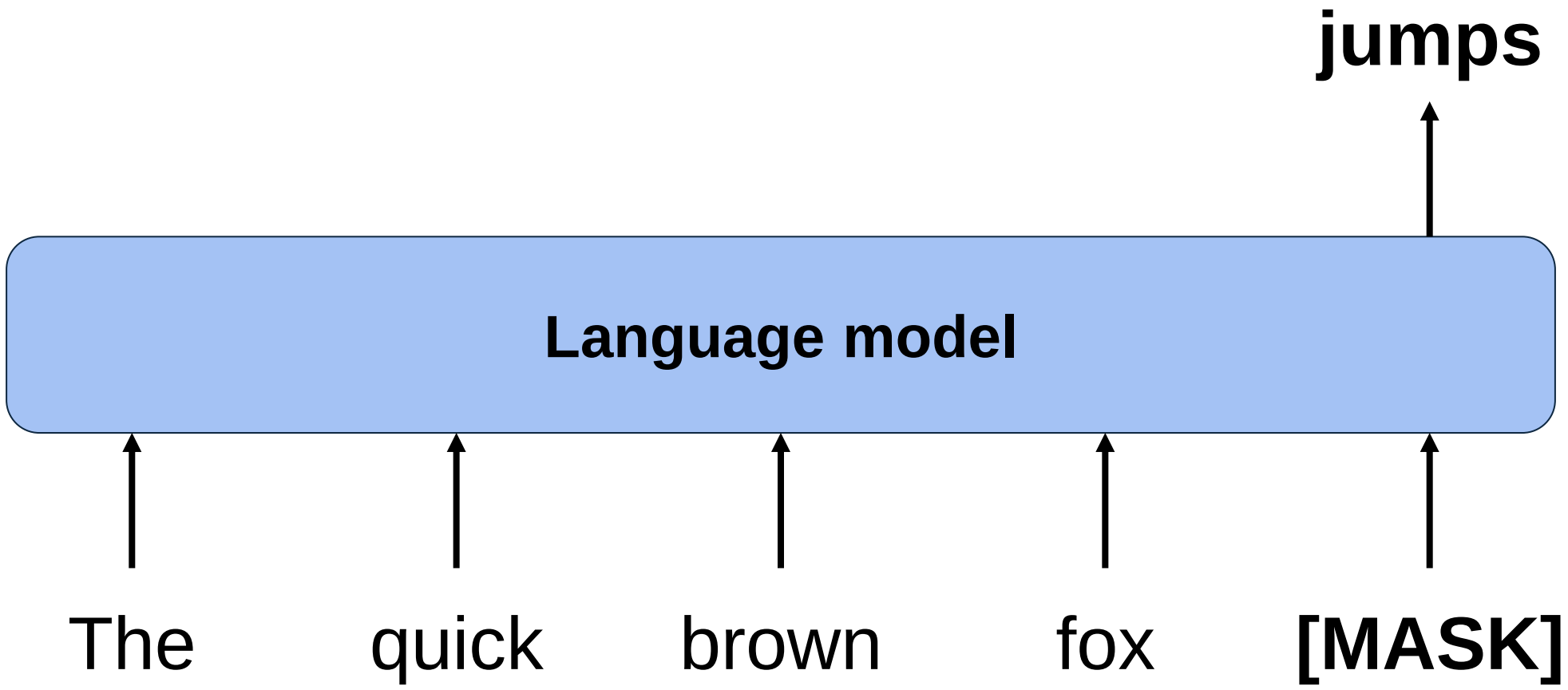
Is the “how” of the NLP system the same as “how” of the brain?



How is information aggregated by the brain during language comprehension?

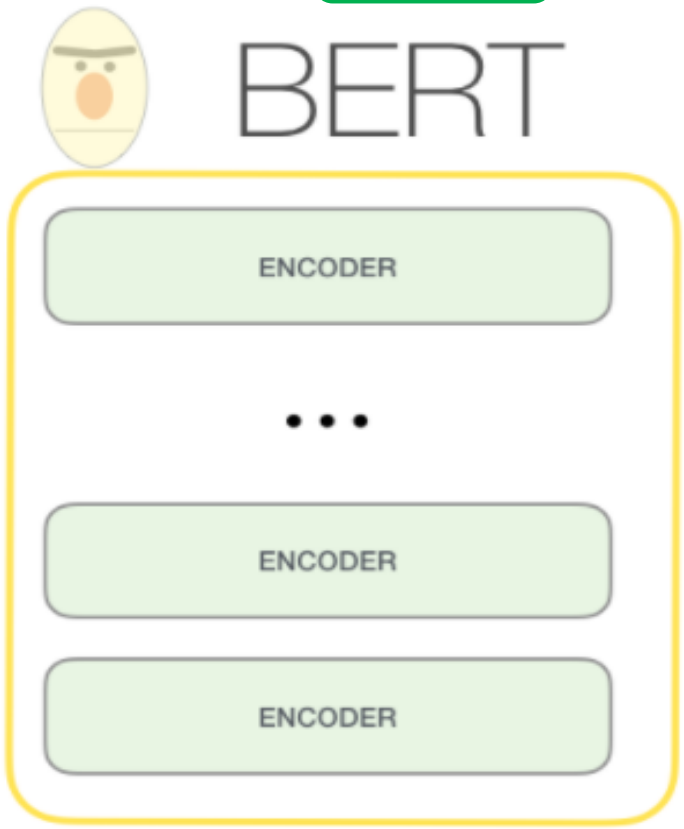
Deniz et al. 2019 Lebel et al. 2022 Nastase et al. 2021 Li et al. 2022 Zhang et al. 2021

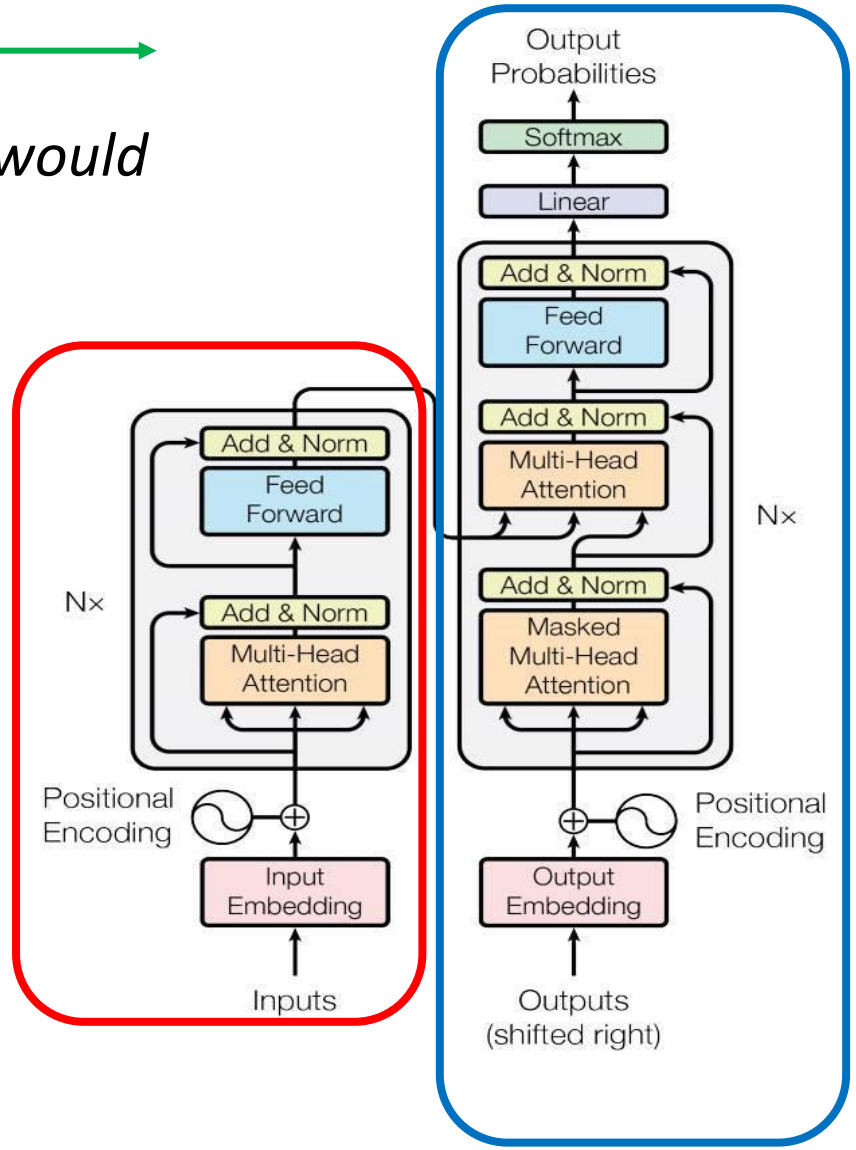
Language models are trained to predict missing words



Transformer: two popular language models (BERT & GPT-2)

Harry never thought he would





Harry never thought he ***???***

Devlin et al. 2018

Vaswani et al. 2017

Radford et al. 2019

Emerging abilities of language models

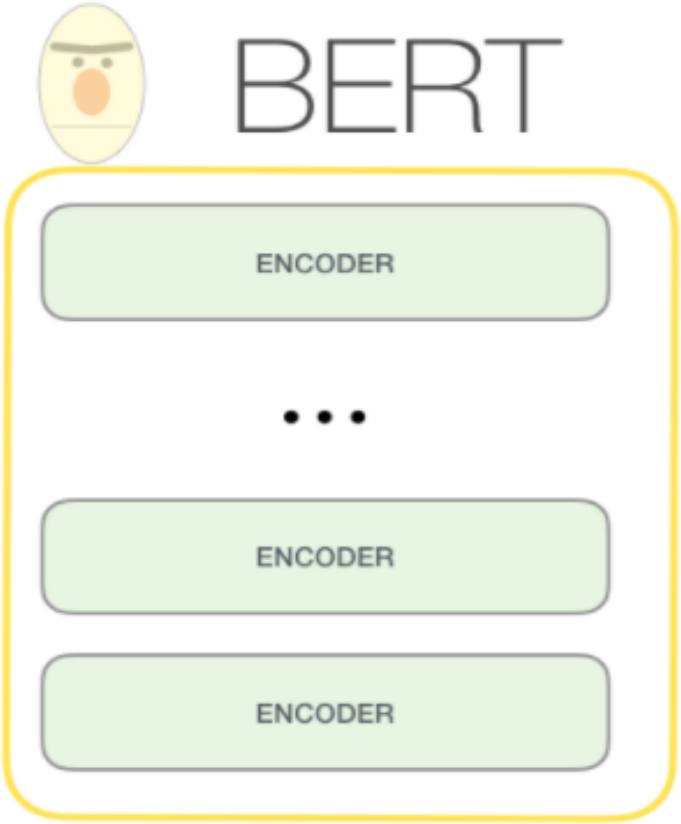
Advantages are:

- Deals with **longer context lengths** (e.g. 4096 sequence length in longformer model)
- Models are pretrained on **different domain specific datasets**, and have **reasoning capabilities**

https://huggingface.co/HuggingFaceH4/zephyr-7b-beta

Extracting representation of a word from LMs

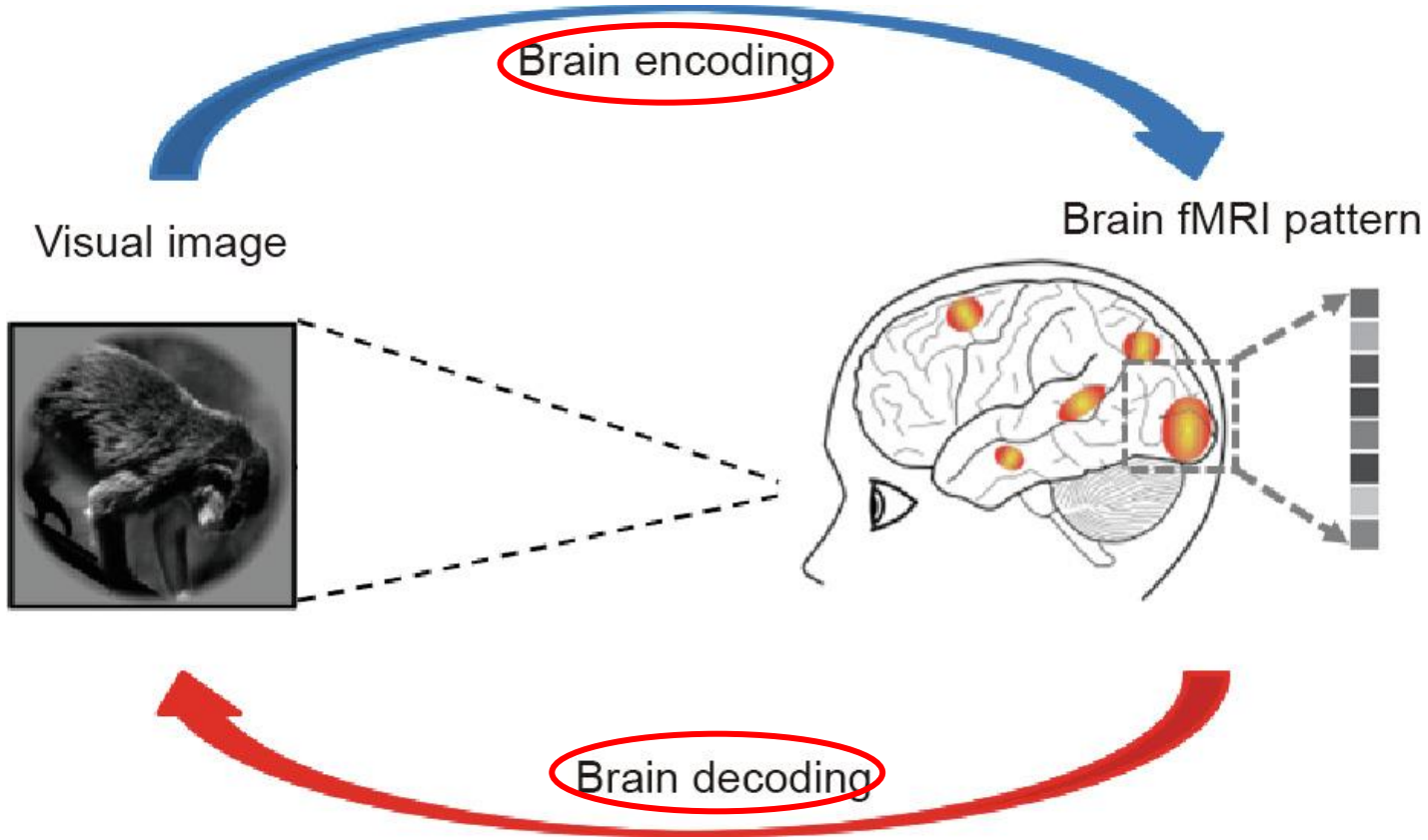
- What layer should be considered?
- How much context / what context ?
- Attention heads, feed forward layers, Weight activations?



Devlin et al. 2018

Brain Encoding and Decoding

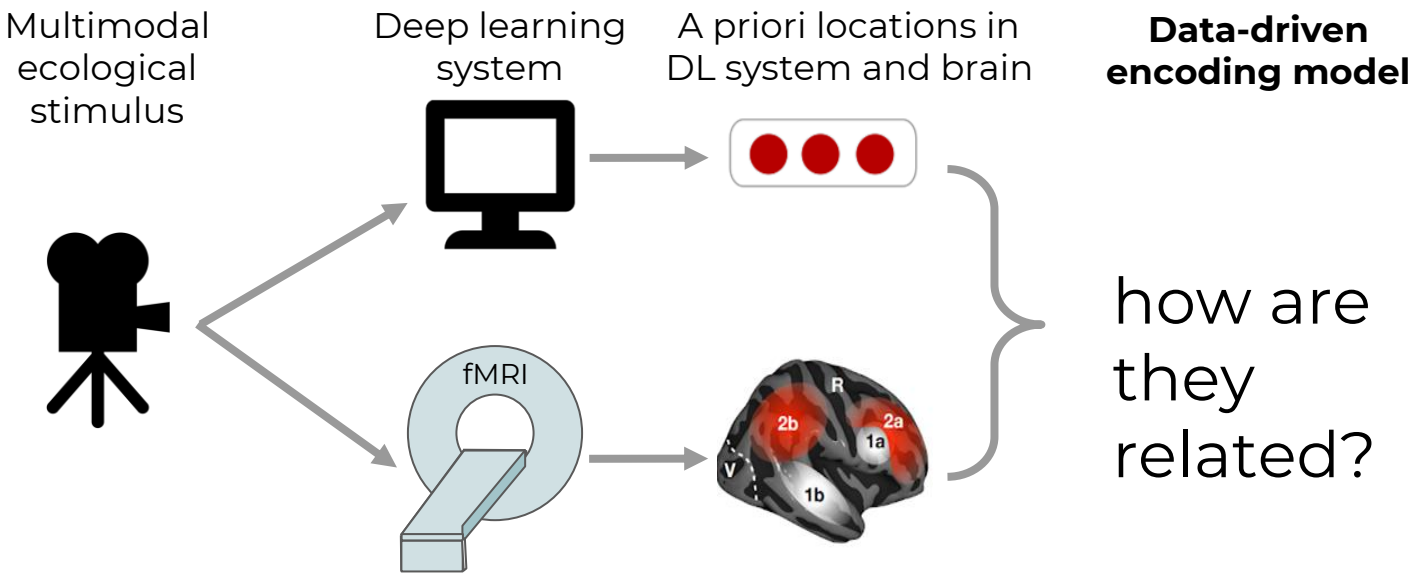
How is the stimulus represented in the brain?



Reconstruct the stimulus, given the brain response?

Ivanova et al. 2022

Deep neural networks and brain alignment: brain encoding and decoding

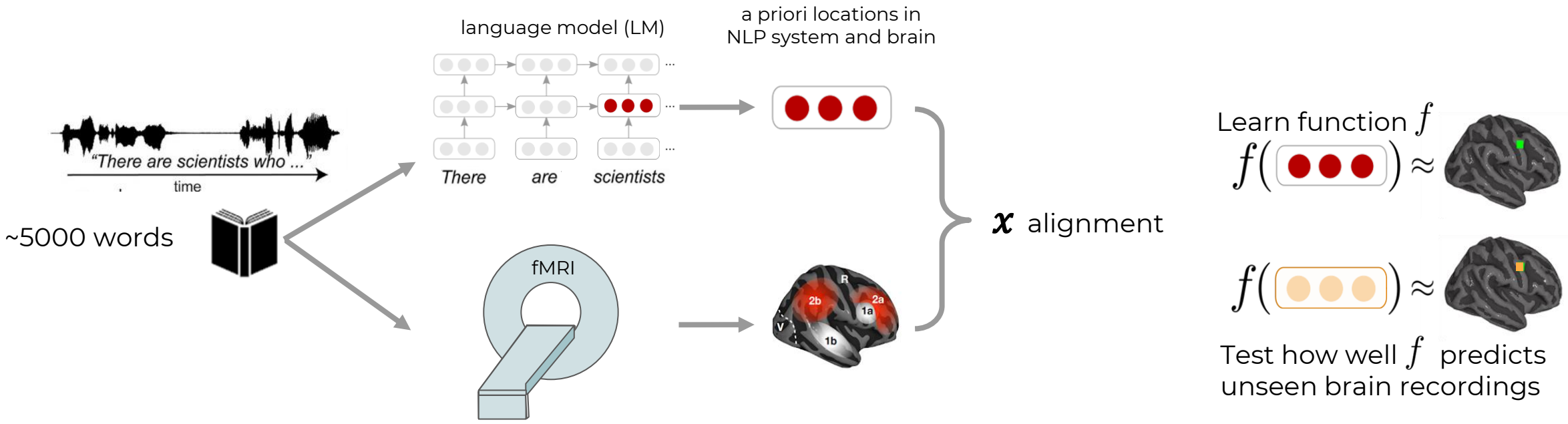


Wehbe et al. 2014,
Jain and Huth 2018,
Gauthier and Levy 2019

Toneva and Wehbe 2019,
Caucheteux et al. 2020,
Toneva et al. 2020

Jain et al. 2020,
Schrimpf et al. 2021,
Goldstein et al. 2022
...

General encoding pipeline to evaluate brain-LM alignment



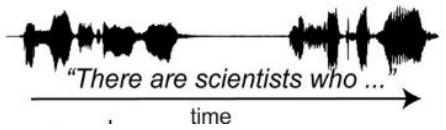
Brain alignment of a LM ⇒ how similar its representations are to a human brain's

Wehbe et al. 2014,
Jain and Huth 2018,
Gauthier and Levy 2019

Toneva and Wehbe 2019,
Caucheteux et al. 2020,
Toneva et al. 2020

Jain et al. 2020,
Schrimpf et al. 2021,
Goldstein et al. 2022
...

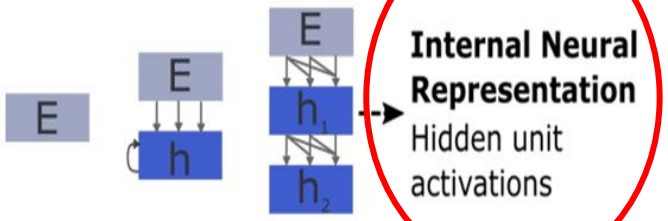
Encoding schema



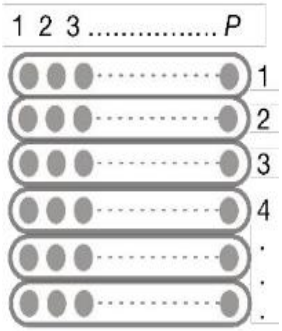
Stimulus

Input

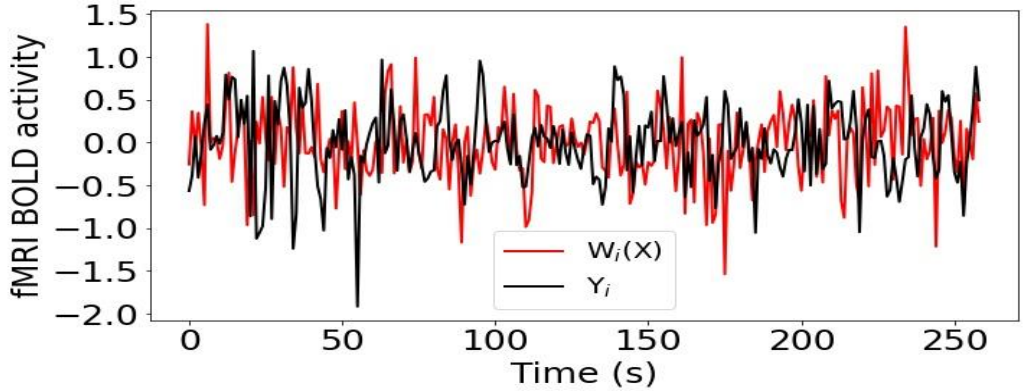
Models



Internal Neural Representation
Hidden unit activations



Pearson Correlation (R) = Corr(Y, W(X))



X

Input

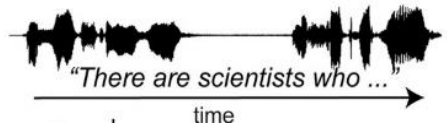
W

Ridge Regression

Y

Output

Internal Neural Representations
fMRI



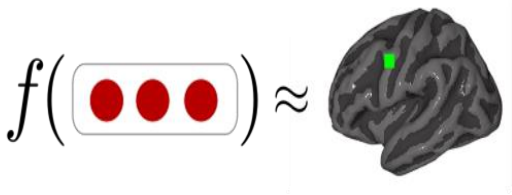
Stimulus

Present

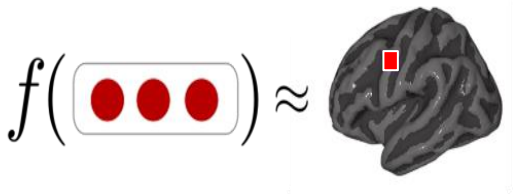
Encoding: training independent models

- Independent model per participant

P1

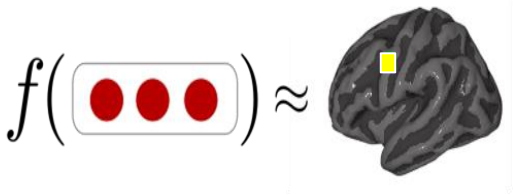


P2



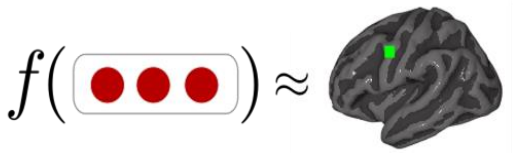
...

PN

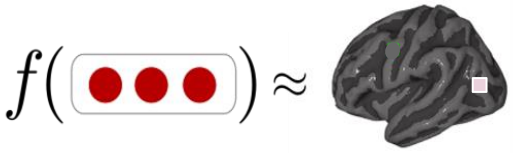


- Independent model per voxel / sensor-timepoint

P1, v1

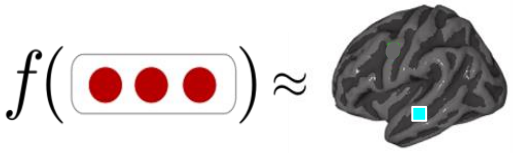


P1, v2

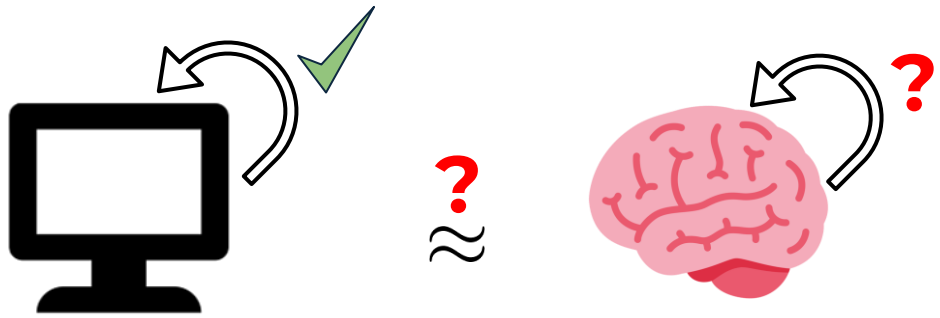


...

P1, vm



Research questions



Language models predict the next word from nearby words, but semantics are crucial for language comprehension.

Is the **how brain utilizes context** through time to process words **during narrative story listening**?

Do **language models** that process **longer sequences** align better with human participants during **long narratives story listening**?

fMRI measures BOLD suffers from delay due to Hemodynamic Response Function

How: ???

Do diverse language regions within the brain is **impacted by delays?**

Grounded language acquisition: how infants can learn language by observing their environments.

How does the **model's ability** to learn **semantic concepts** through cross-situational learning in **noisy supervision?**

Objectives

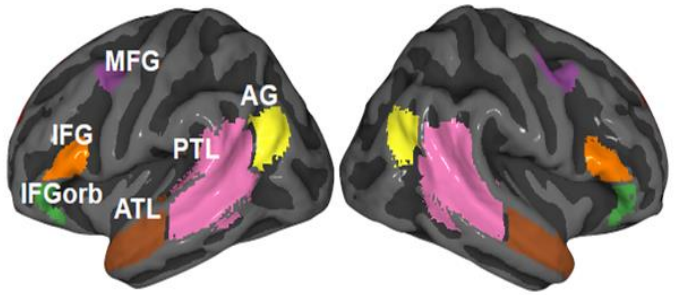
- What happens during narrative story listening?
  - Whether a new word representation is combined with previous context?

- We use MEG activity to trace
  - How is context represented in the brain?
  - How does previous context help for new word meaning?

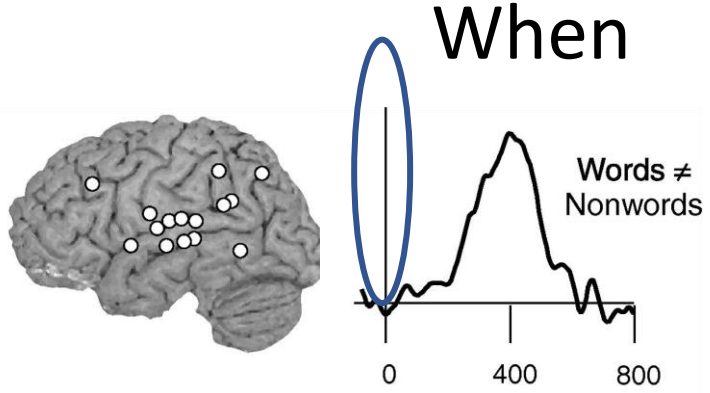
With MEG we can analyze sub-word time course

- MEG recording data at very fast temporal resolution
- So, we can look at sub-word process
- fMRI recording data at very high-spatial resolution

Where



When



Text: Word Contexts

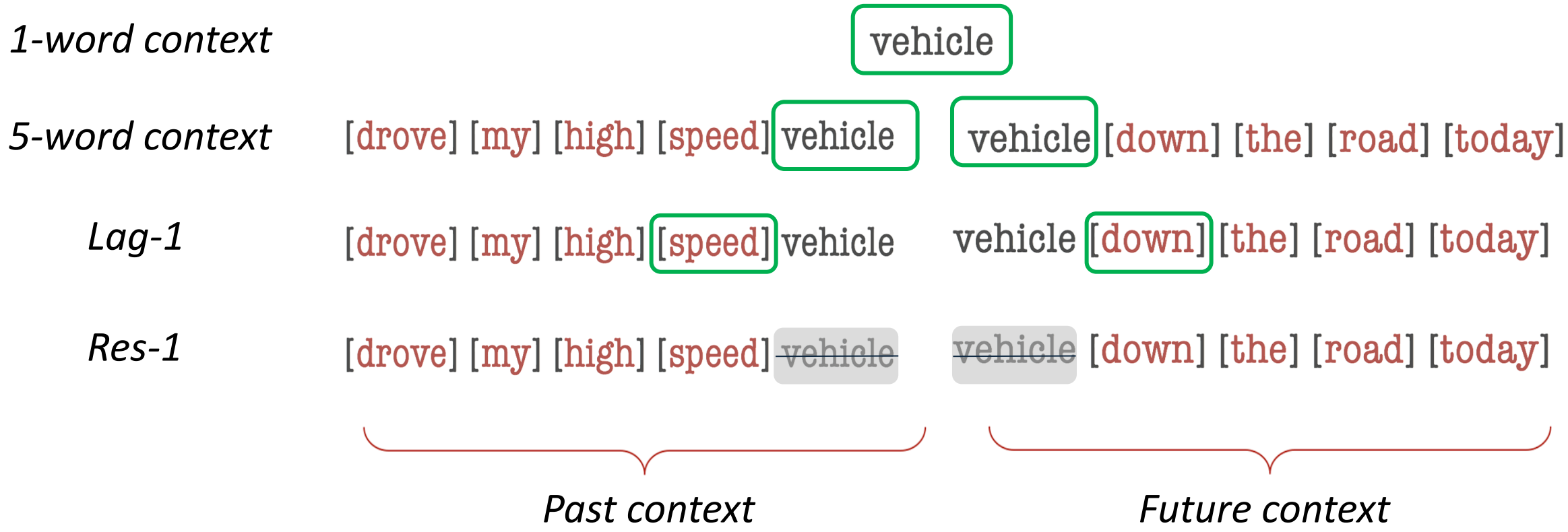
[Context word] [Context word] *Current word*

Selected embedding

How to extract contextual word represention?

How does previous context help for new word meaning?

Isolating current word meaning is a type of intervention

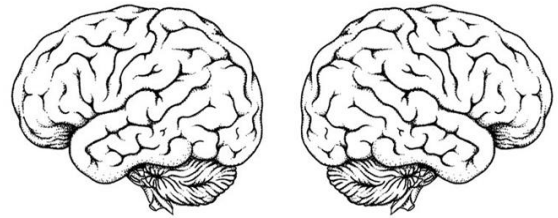


Listening data target: human brain MEG recordings

- We use MEG-MASC story listening dataset:
  - 27 subjects (8-subjects used),
  - 4-stories (11,002 words)
- **Alignment performed between MEG signal and word representations around every word onset**
  - 800ms signal window around word onset: 200ms before (baseline correction), 600ms after

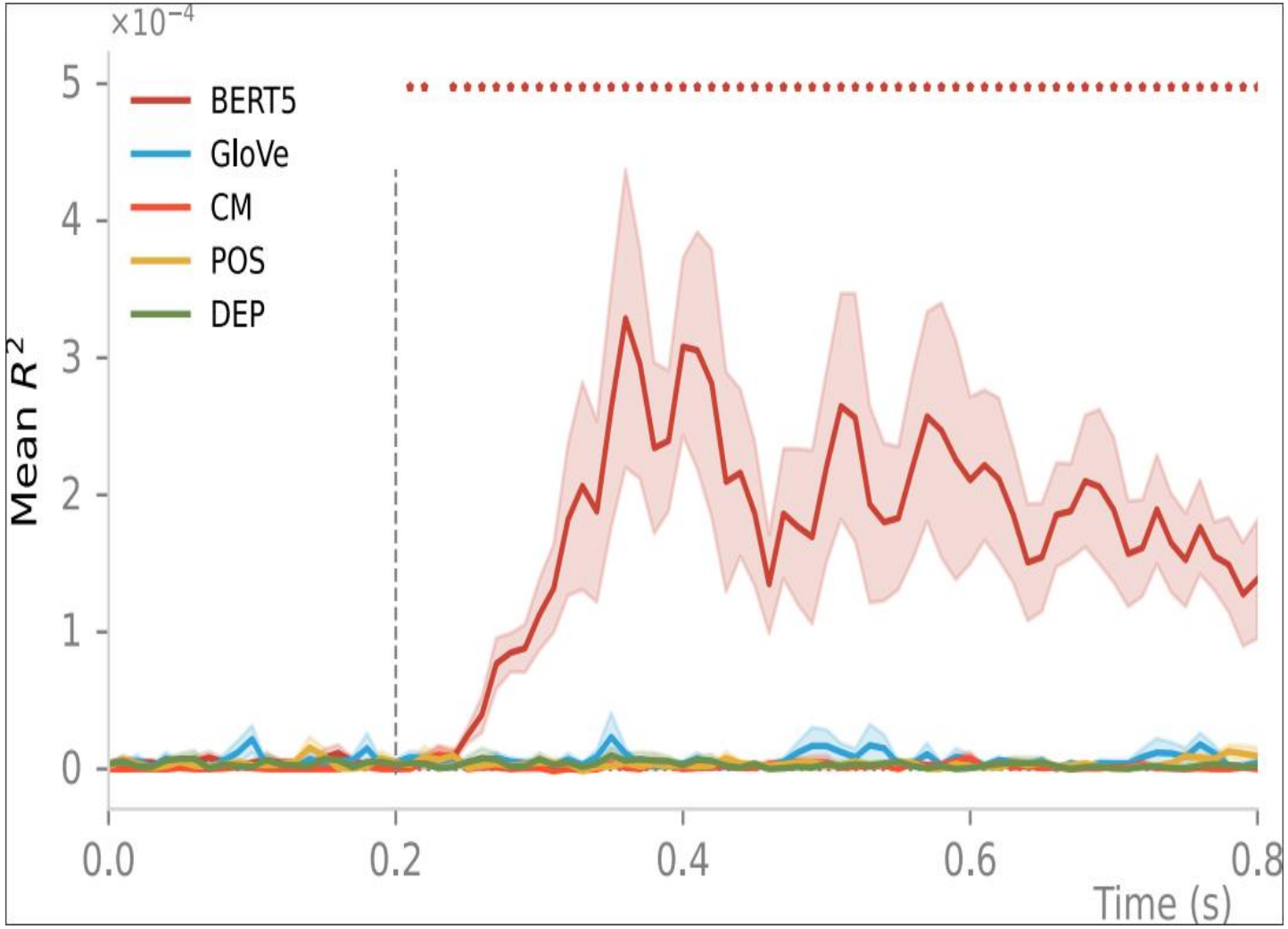
MEG-MASC Dataset

- MEG recordings
- 27 participants
- 2 hours of story listening
- 2 repeated sessions
- structural MRIs
- audio, phonetic and word annotations
- standardised BIDS structure



Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkanen, David Poeppel, Jean-Remi King. "MEG-MASC: a high-quality magneto-encephalography dataset for evaluating natural speech processing" Arxiv (2022).

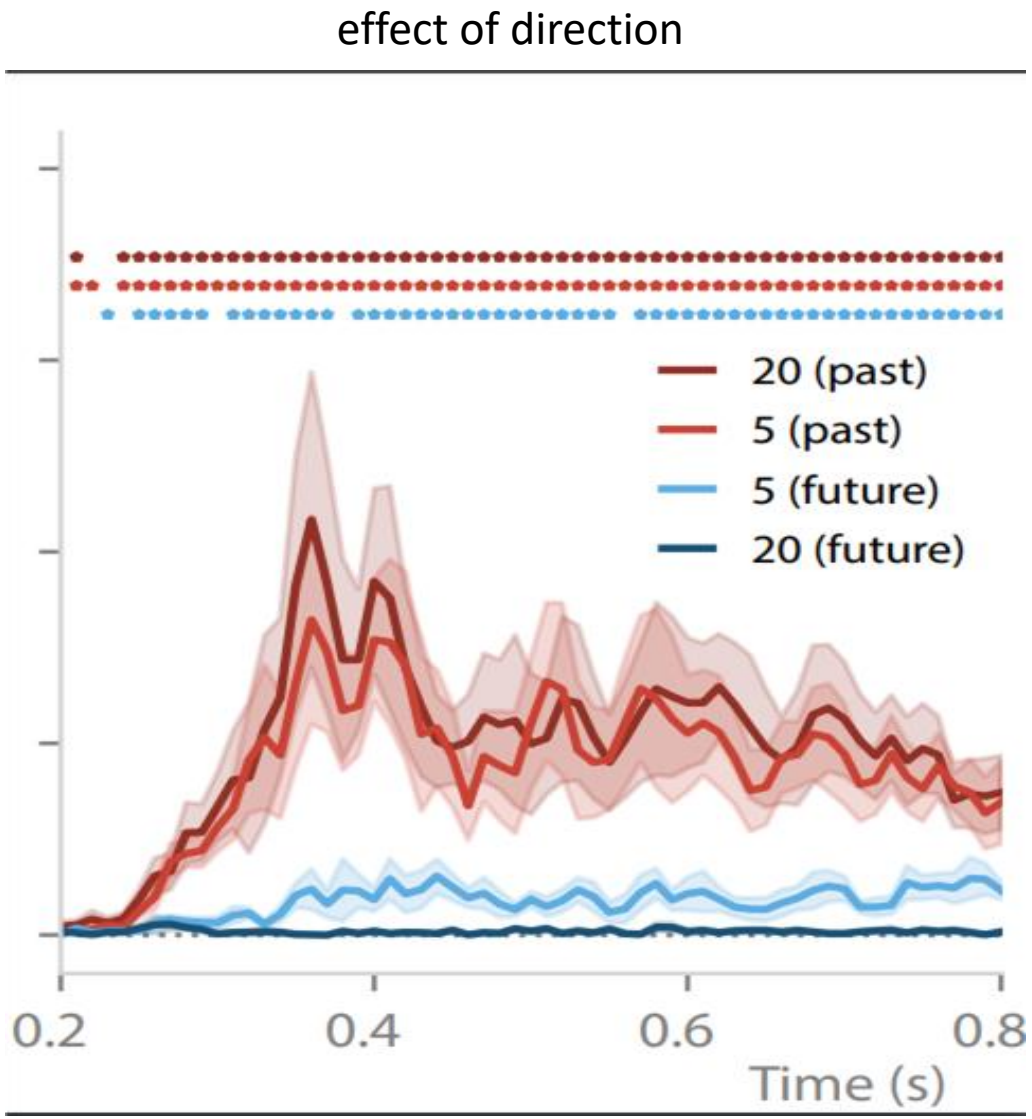
Encoding Performance of Syntactic and Semantic Methods



- Simple syntactic features (Complexity Metric, Parts of Speech, Dependency tags)
- Non-contextual word representations (GloVe)
- BERT contextual representations

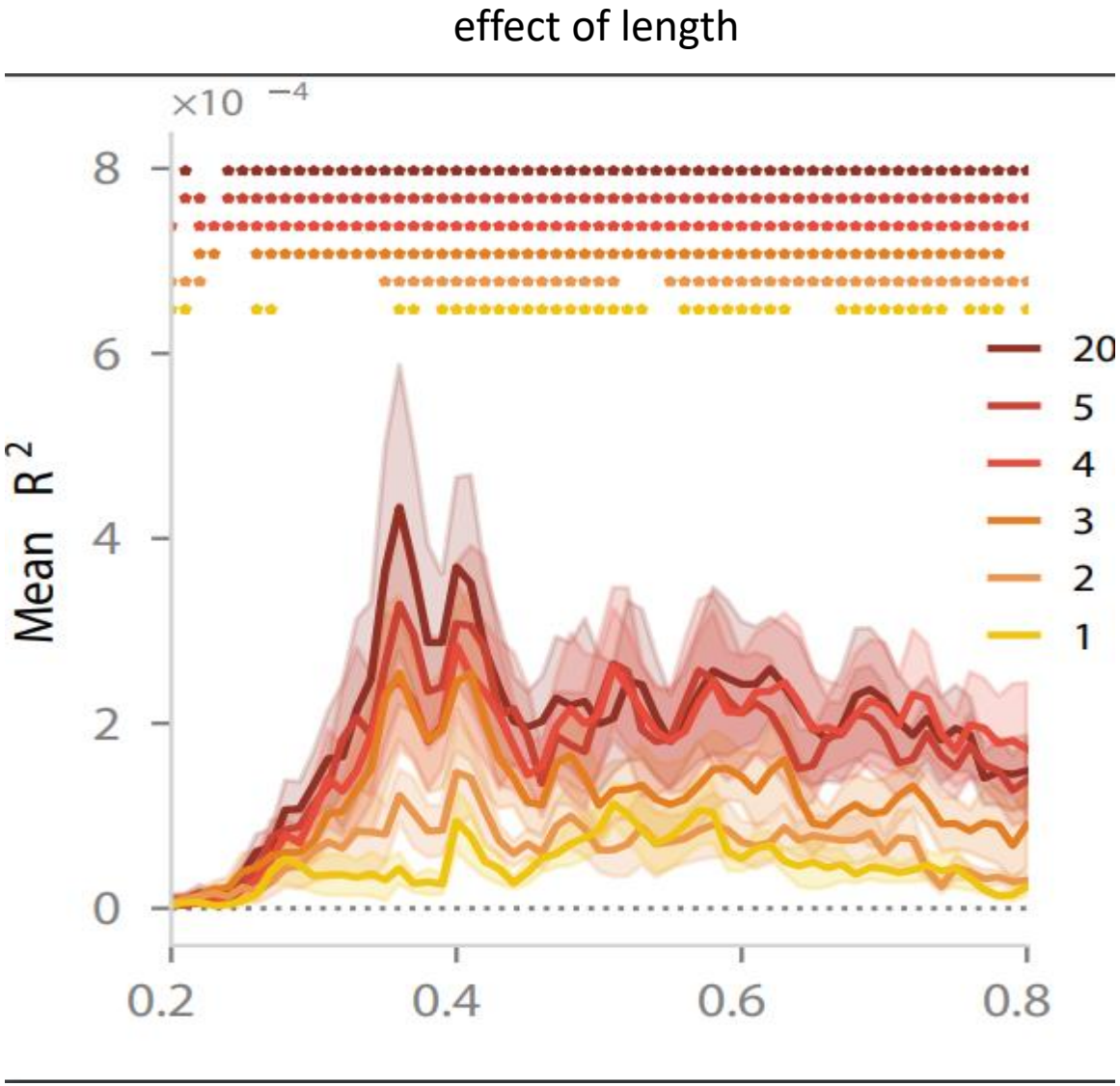
Due to limited context information, basic syntactic (CM, POS and DEP), and non-contextual semantic features (GloVe) are, on average, not correlated with the considered window of MEG activity

How is context represented in the brain



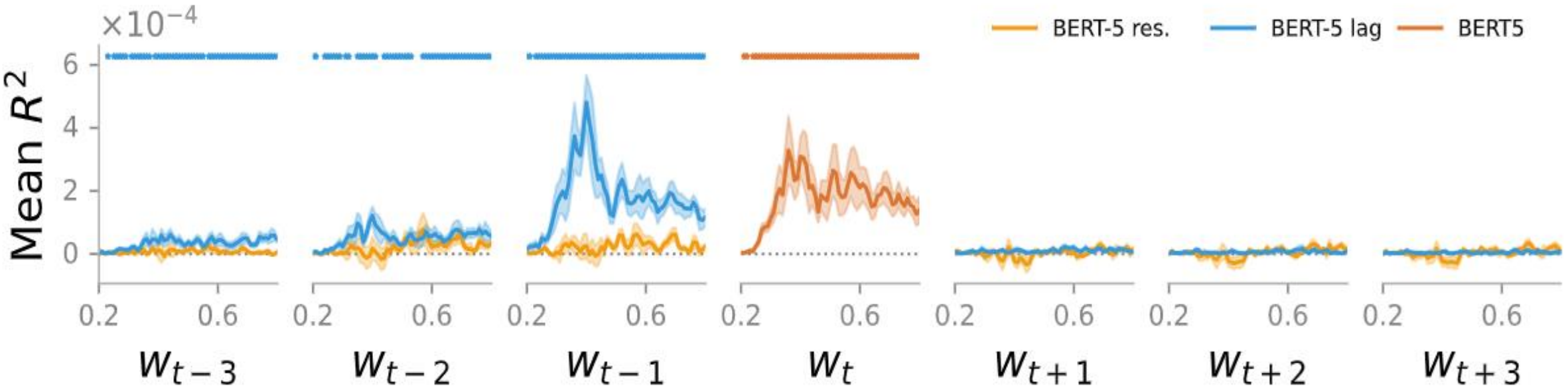
Long past contexts enable better encoding than future or short-scale present contexts

Contextual BERT Embeddings:



Context length plays a crucial role in predicting MEG (300 to 425 ms).

Contextual BERT Embeddings (Residuals vs. Lag)



MEG is sensitive to mostly the current and previous words

Past word context is crucial in obtaining significant results.

Oota, Subba Reddy et al. 2023. "MEG Encoding using Word Context Semantics in Listening Stories."

Partial Conclusions

Similar to language models, human brain process words in time through close past words

By varying context lengths, we showed that semantics are crucial for brain language comprehension

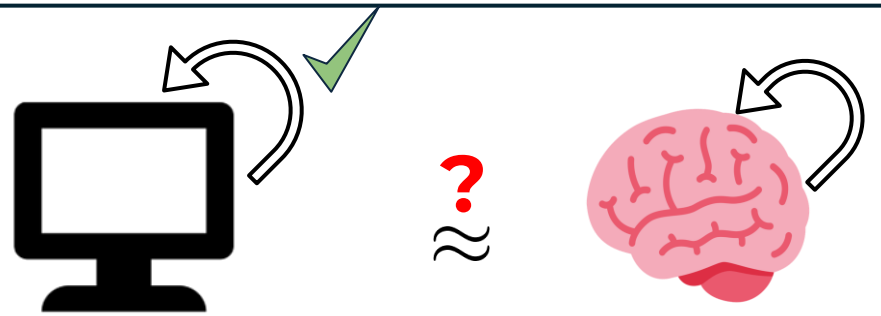
Coherent with previous studies:

- Gwilliams et al. 2022 showed that the several past phonemes information (with position and order in sequence) are kept in memory

Research questions

Language models predict the next word from nearby words, but semantics are crucial for language comprehension.

Is the **how brain utilizes context** through time to process words **during narrative story listening**?



Do **language models** that process **longer sequences** align better with human participants during **long narratives story listening**?

fMRI measures BOLD suffers from delay due to Hemodynamic Response Function

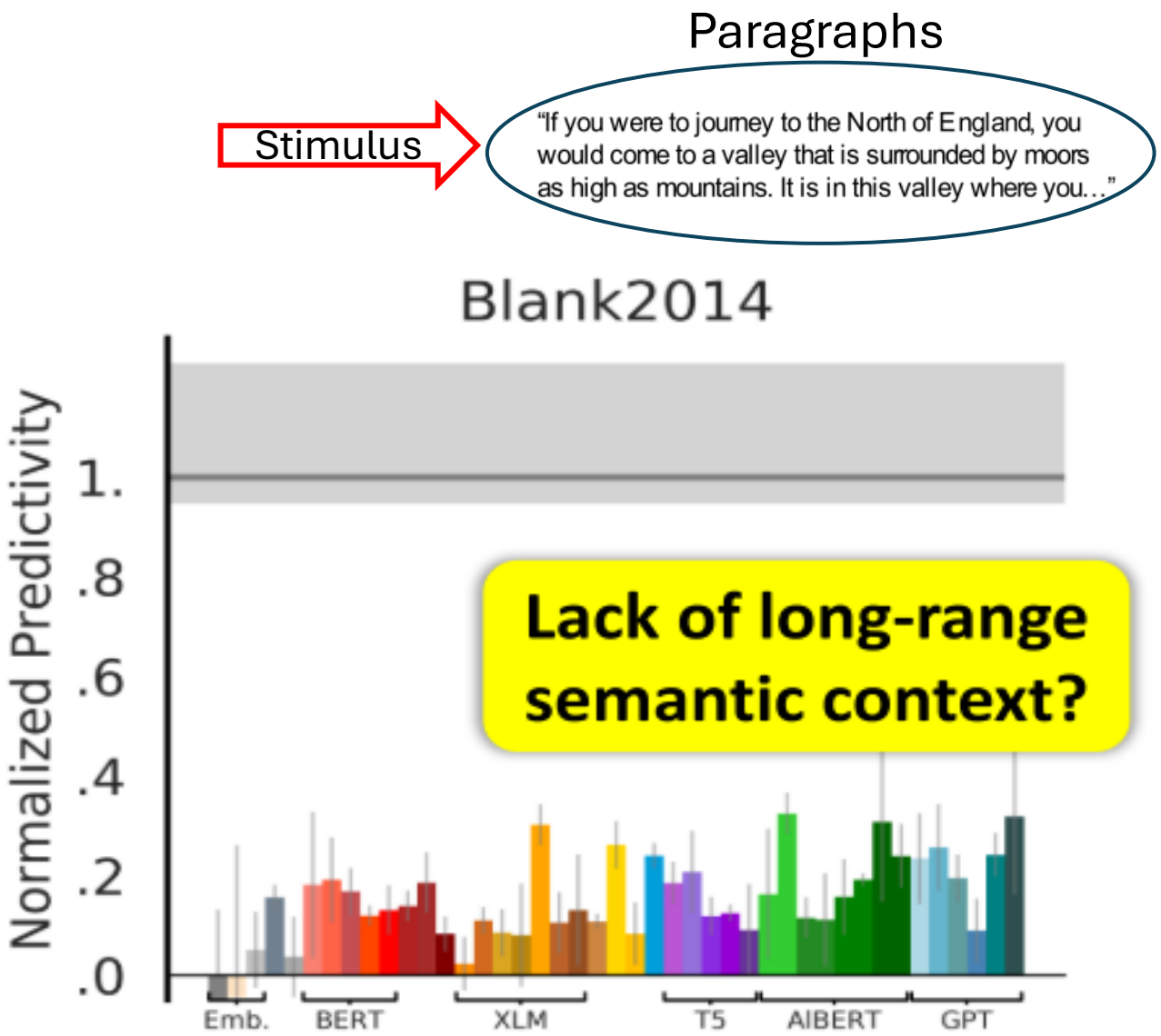
How: ???

Do diverse language regions within the brain is **impacted by delays?**

Grounded language acquisition: how infants can learn language by observing their environments.

How does the **model's ability** to learn **semantic concepts** through cross-situational learning in **noisy supervision?**

Can current language models deal with long-term dependencies?



- Transformer language models (BERT & GPT-2) are unable to handle the long-term dependencies
  - **sequence** length is fixed to 512 words
- LSTMs still lacks investigation of the long-term memory cognitive plausibility and its link to fMRI data

Vaswani et al. 2017, Schrimpf et al. 2021

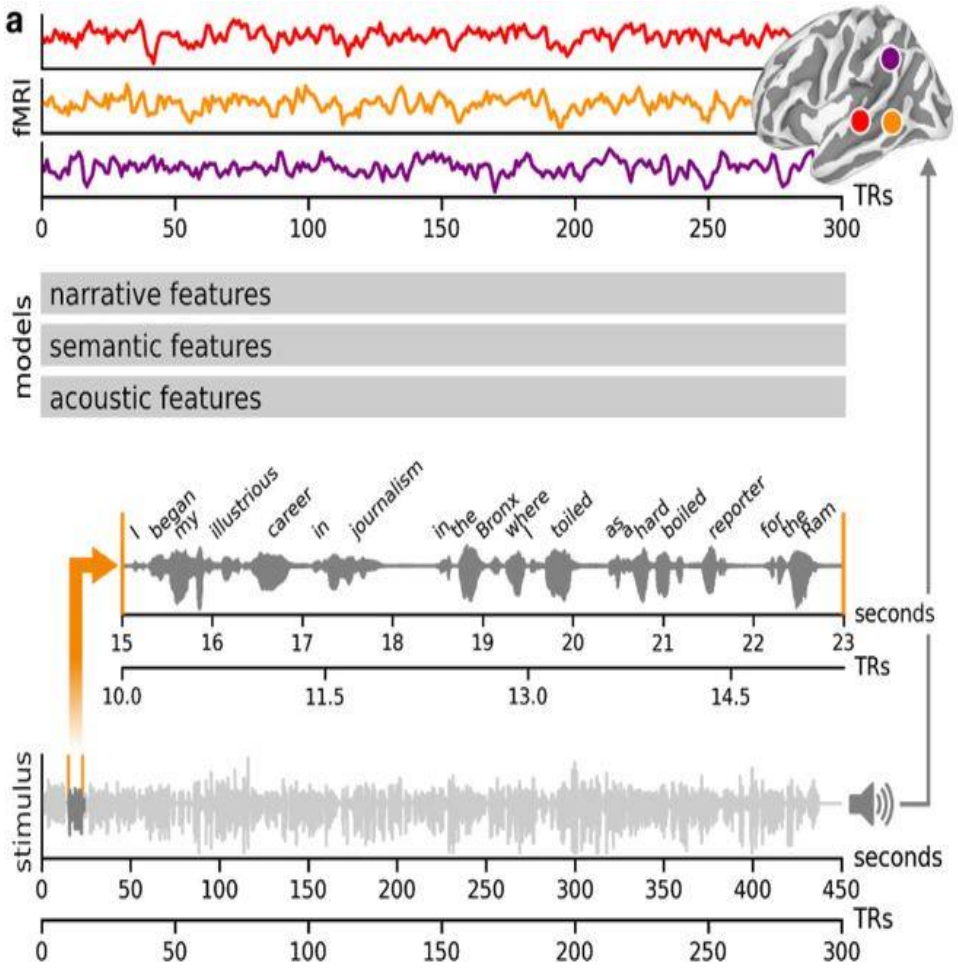
What kind of language models can represent long-term dependencies?

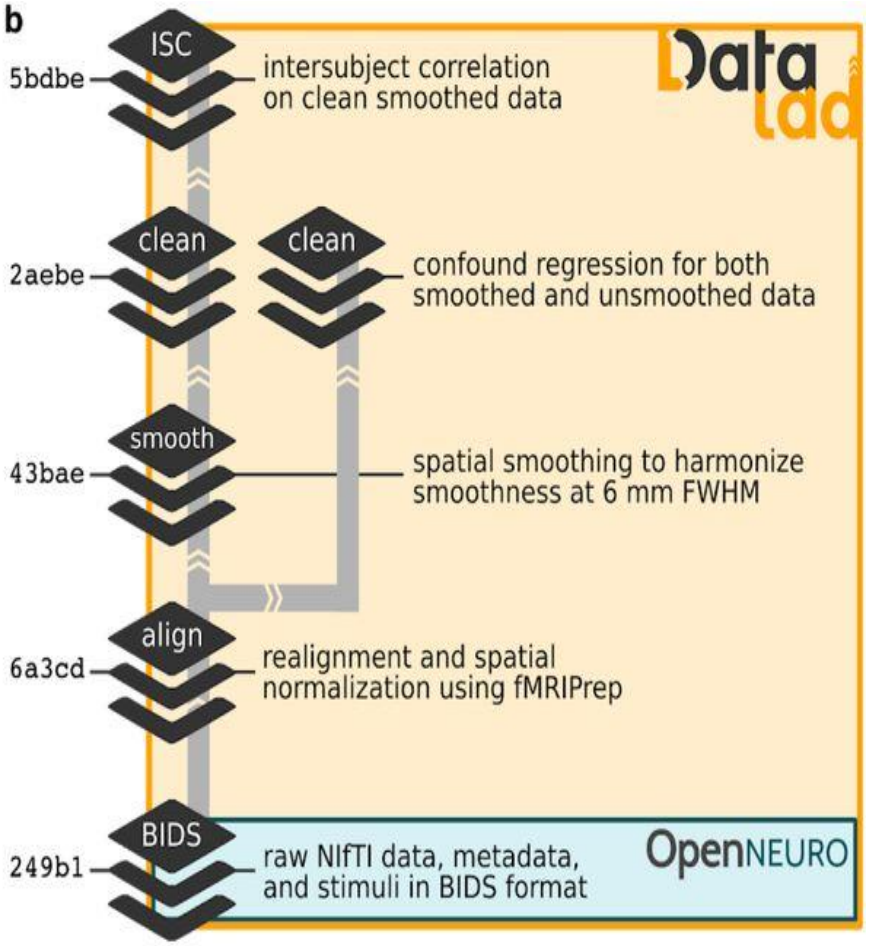
could they also predict higher cognition while subjects are engaged in longer stories ?

The data target: human brain recordings

- We use Pieman story listening:
  - 82 subjects,
  - 282 TRs (repetition time)
  - here it is 1.5 sec.

Example: "I began my illustrious carrier in journalism..."





Nastase et al. 2021 fMRI

Text: Feature Representaions

Static word embeddings

GloVe

Recurrent models

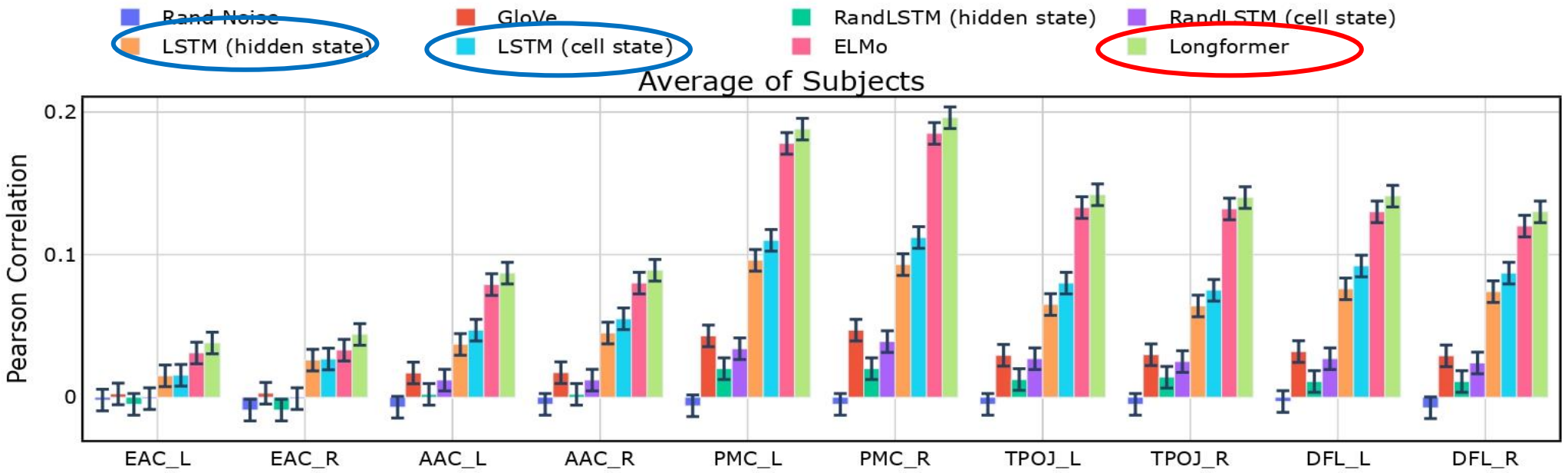
Long Short-Term Memory Networks (LSTM)

- Cell state (long memory)
- Hidden state (short memory)

Semantic embeddings

ELMo
Longformer

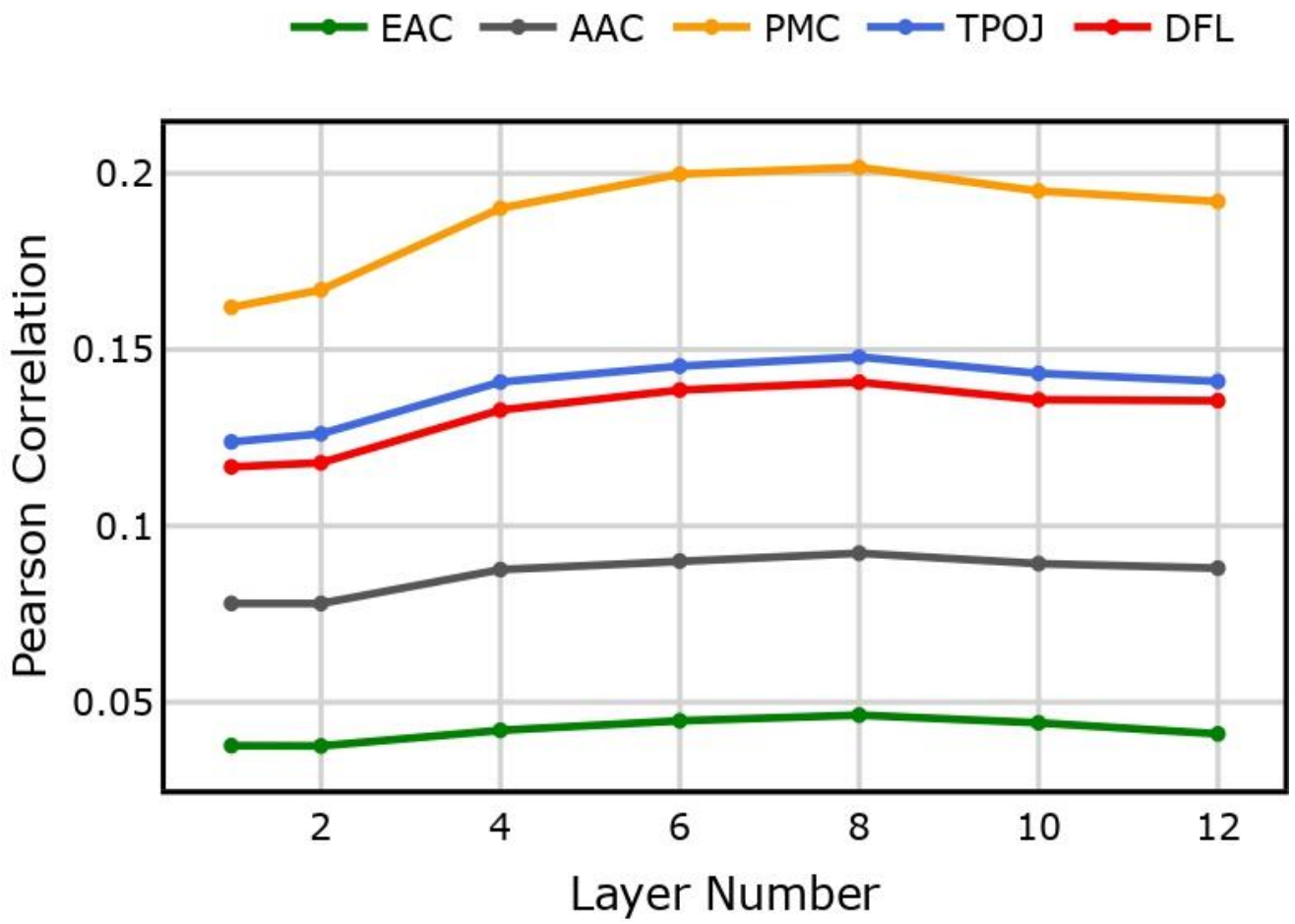
Encoding Performance of language models



- Longformer model representations have high Pearson correlation across language rois and sensory regions (early auditory cortex)

- LSTM cell state representations display better brain alignment than hidden state representations

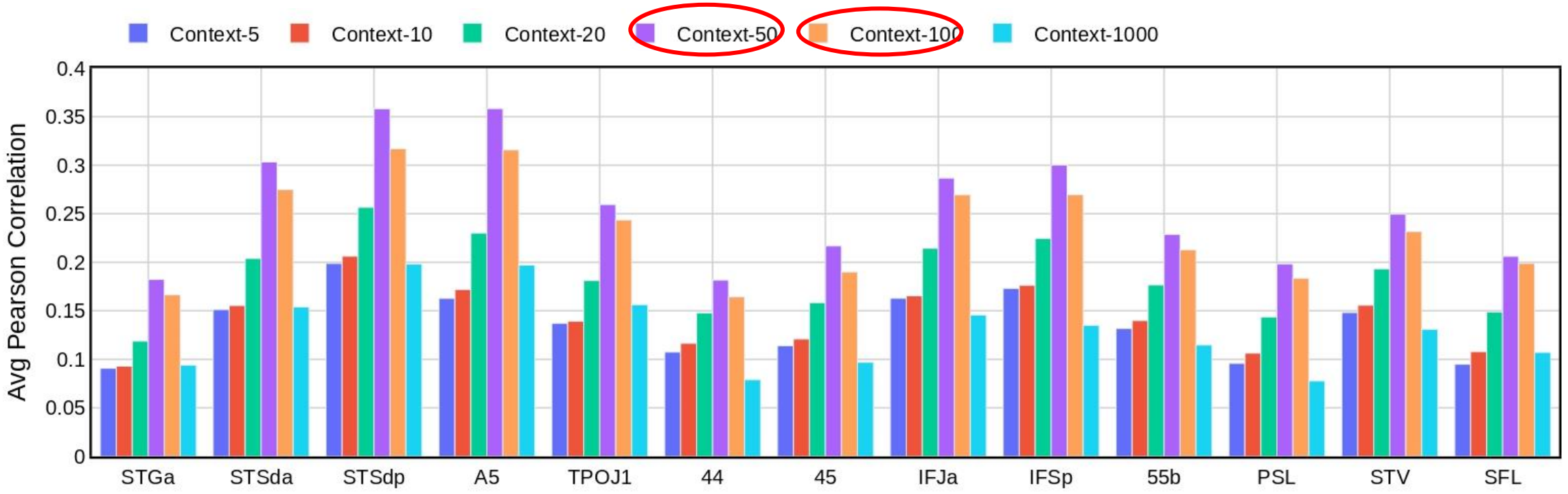
Layer-wise encoding performance: Longformer



best alignment with fMRI in middle layers

Effect of context length on longer-context language models

Longformer



brain alignment improves only when we provide longer input contexts (5-100)

Partial Conclusions

Use human brain recordings to evaluate how well representations from language models (static vs. recurrent vs. pretrained) can predict representations of the human brain during language comprehension

Richer representaions learned from language models, designed to integrate longer contexts, have improved alignment with human brain activity

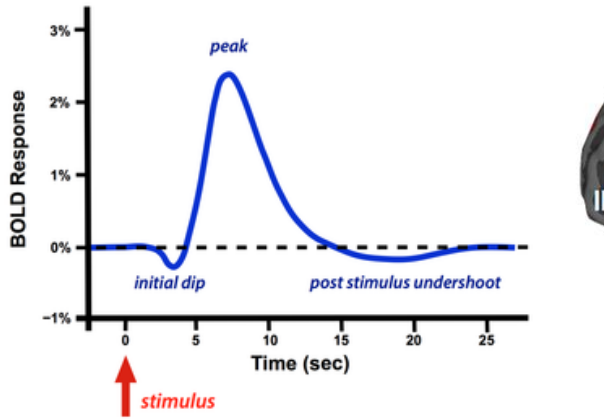
Pretrained language models significantly predict brain language regions that are thought to underlie language comprehension

Research questions

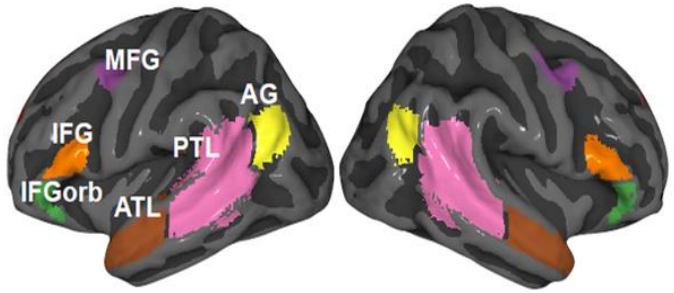
Language models predict the next word from nearby words, but semantics are crucial for language comprehension.

Is the **how brain utilizes context** through time to process words **during narrative story listening**?

Do **language models** that process **longer sequences** align better with human participants during **long narratives story listening**?



fMRI measures BOLD suffers from delay due to Hemodynamic Response Function



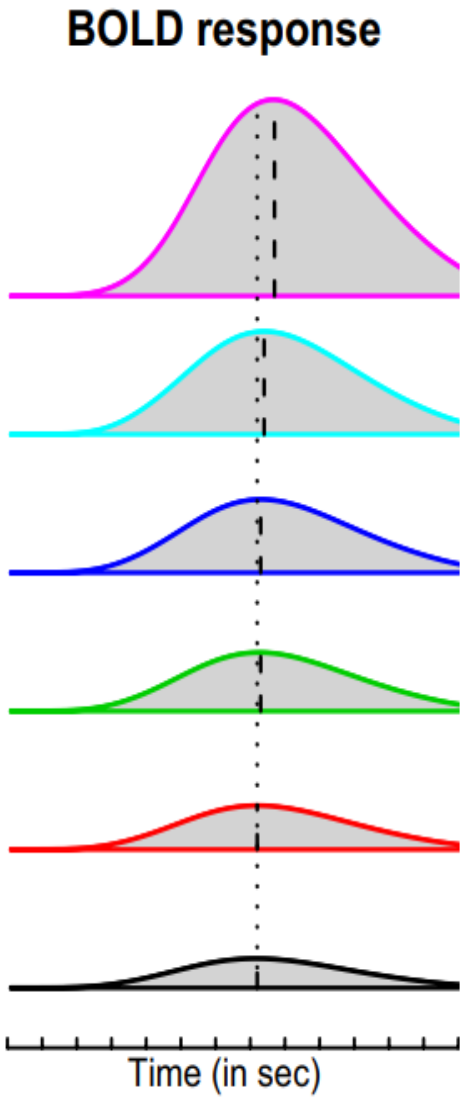
How: ???

Do diverse language regions within the brain is **impacted by delays?**

Grounded language acquisition: how infants can learn language by observing their environments.

How does the **model's ability** to learn **semantic concepts** through cross-situational learning in **noisy supervision?**

Hemodynamic Response Function (HRF) delay.

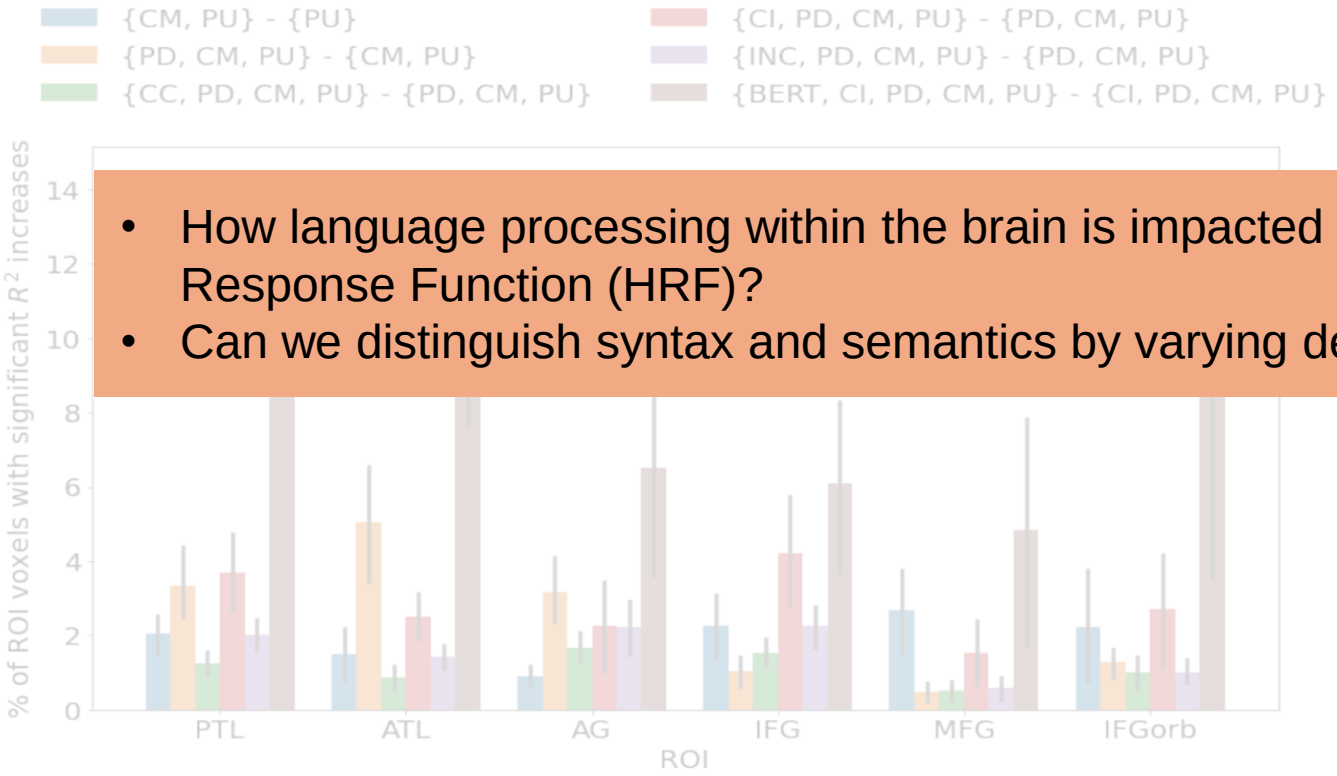


Current brain encoding studies focused on language processing at fixed HRF delay

Stimuli	Authors	Type	Lang.	Delays
Text	(Jain et al., 2020)	fMRI	English	8secs (4 TRs)
	(Jain and Huth, 2018)	fMRI	English	8secs (4 TRs)
	(Caucheteux et al., 2021)	fMRI	English	7.5secs (5 TRs)
	(Reddy and Wehbe, 2021)	fMRI	English	8secs (4 TRs)
	(Merlin and Toneva, 2022)	fMRI	English	8secs (4 TRs)
	(Aw and Toneva, 2022)	fMRI	English	8secs (4TRs)
	(Antonello et al., 2021)	fMRI	English	8secs (4 TRs)
	(Oota et al., 2022b)	fMRI	English	9secs (6 TRs)
	(Oota et al., 2023a)	fMRI	English	12secs (6 TRs)

What information is processed across language regions at fixed HRF delay?

Reddy et al. 2021



- How language processing within the brain is impacted by delays in the Hemodynamic Response Function (HRF)?
- Can we distinguish syntax and semantics by varying delays?

semantic are difficult to distinguish

Text: Feature Representations

Basic word-level syntax

Part-of-Speech Tags (POS) &
Dependency Tags (DEP)

Basic speech features

Phonological

Syntactic tree parsers

Constituent Complete (CC)

Constituent Incomplete (CI)

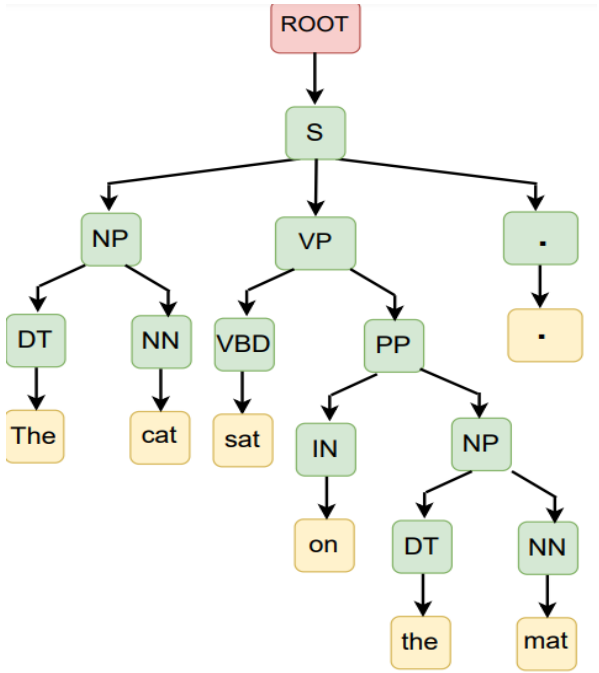
Semantic embeddings

Text models:
BERT, GPT-2
LLaMa-2

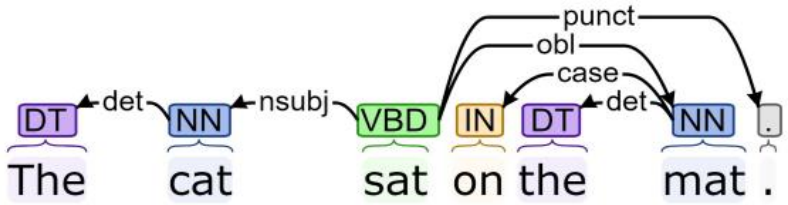
Speech models:
Wav2Vec2.0

Syntactic Parising

- Syntax: how do words structurally combine to form sentences and meaning?
- In natural language processing, there are two popular syntactic parsing methods



Constituency parsing

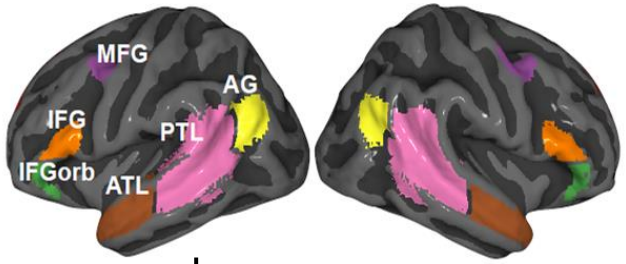


Dependency parsing

Listening data target: human brain recordings

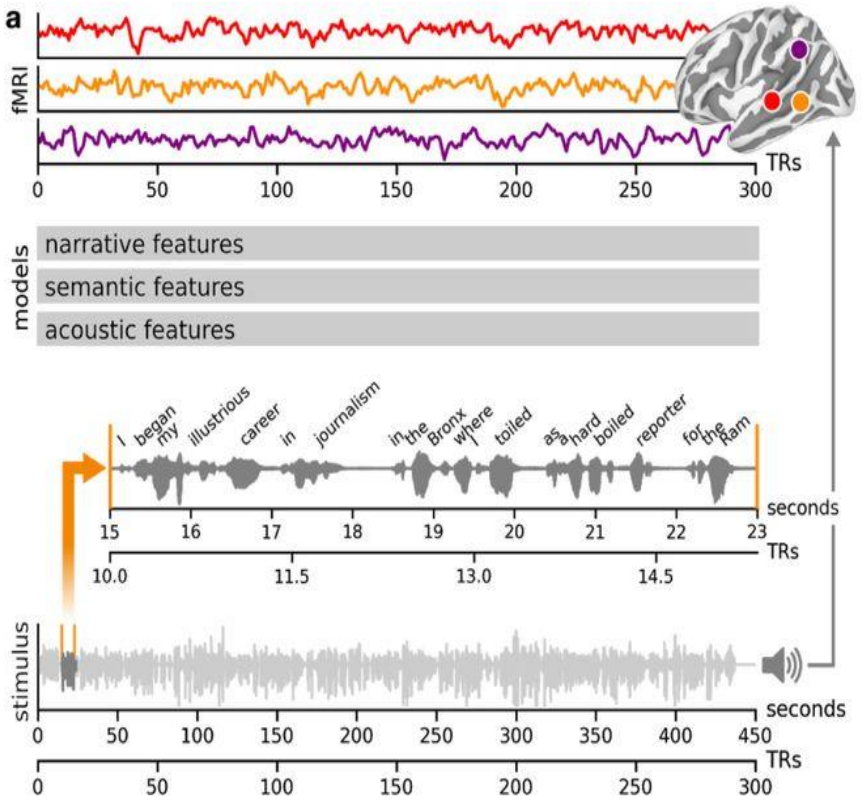
- We use Tunneling story:
  - 22 subjects,
  - 1023 TRs (repetition time)
  - here it is 1.5 sec.

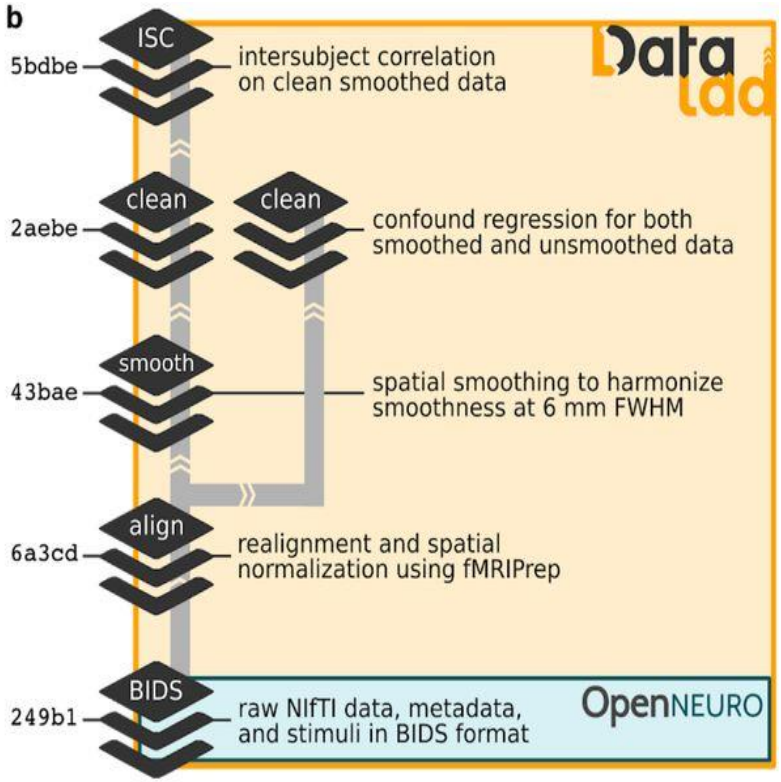
Example: "I began my illustrious carrier in journalism..."



Language Network

[Fedorenko et al. 2010, 2014]





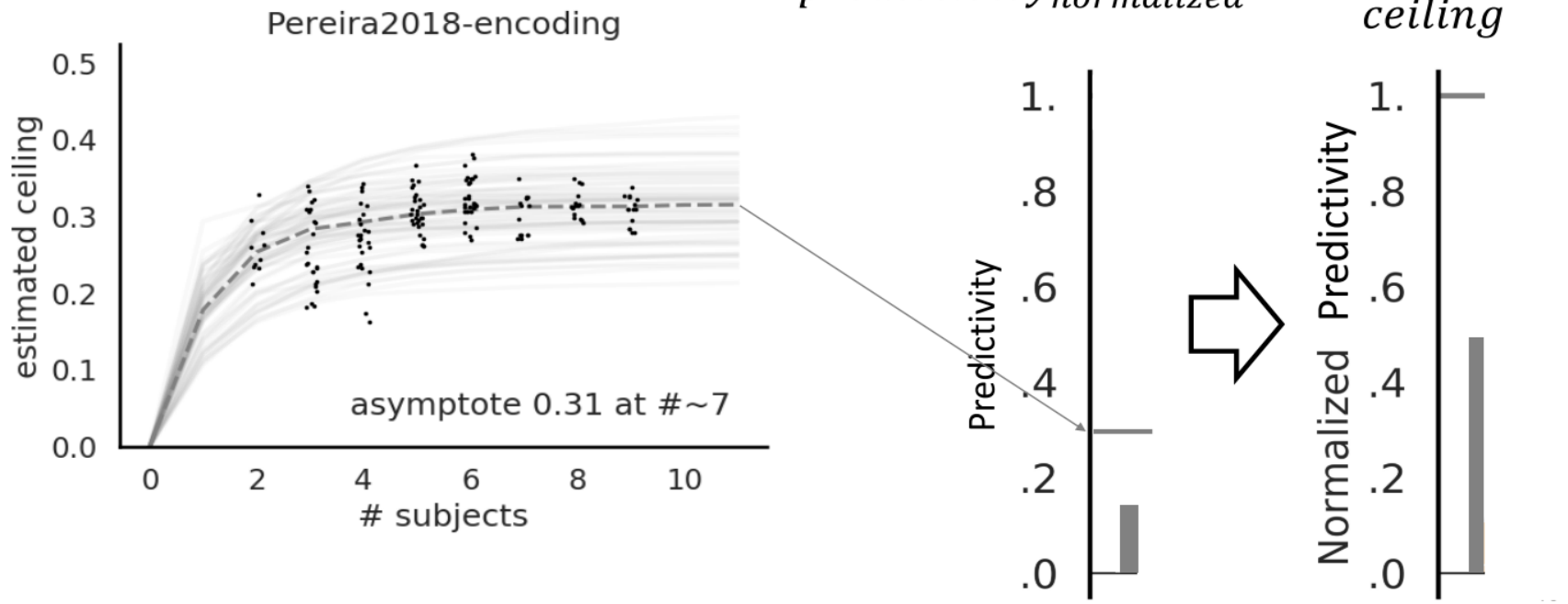
Nastase et al. 2021

Normalized Predictivity

"how close are we" – **ceiling**

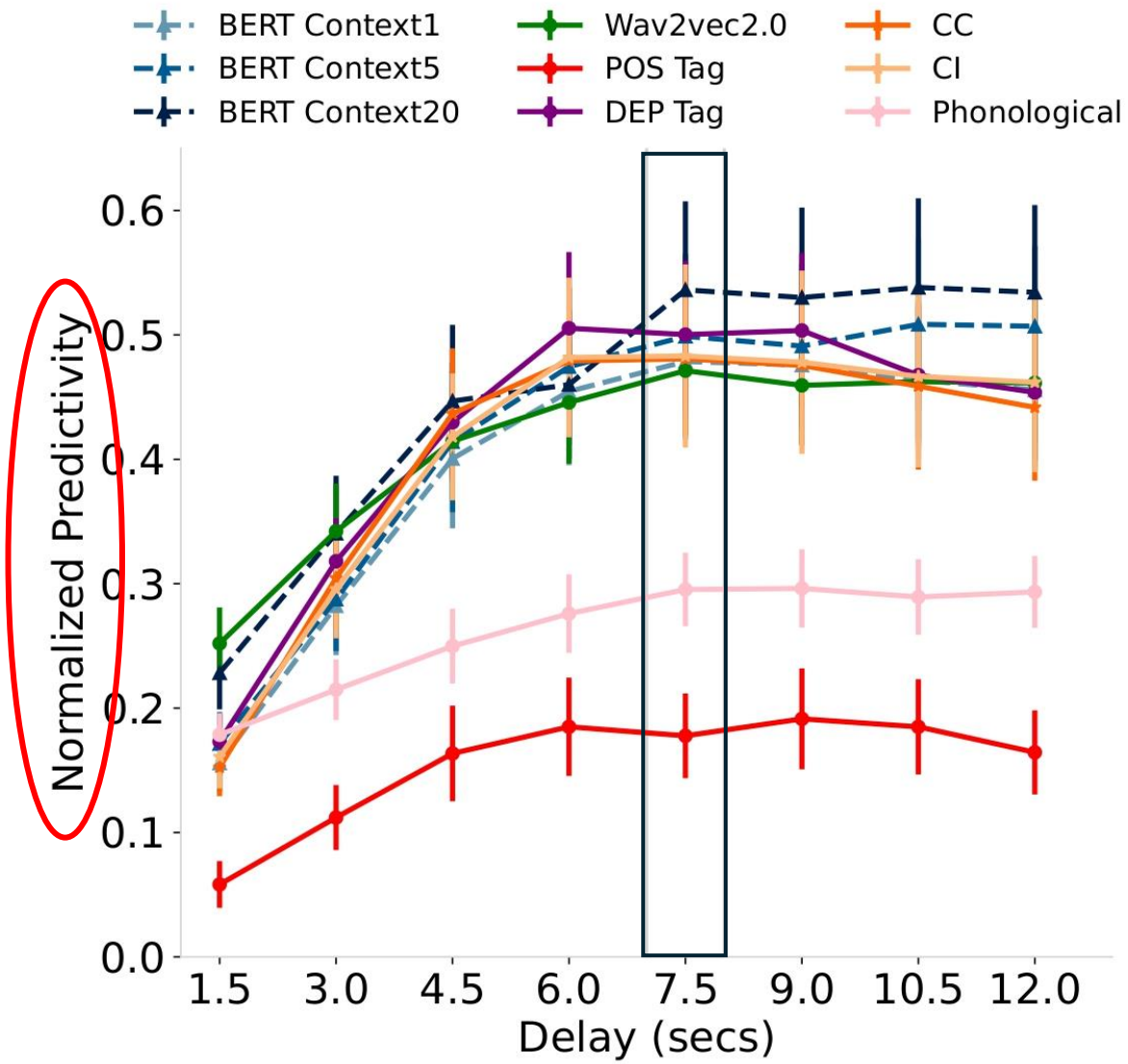
compute how well a pool of subjects predicts a held-out subject

$$predictivity_{normalized} = \frac{predictivity}{ceiling}$$



Schrimprf et al. 2021

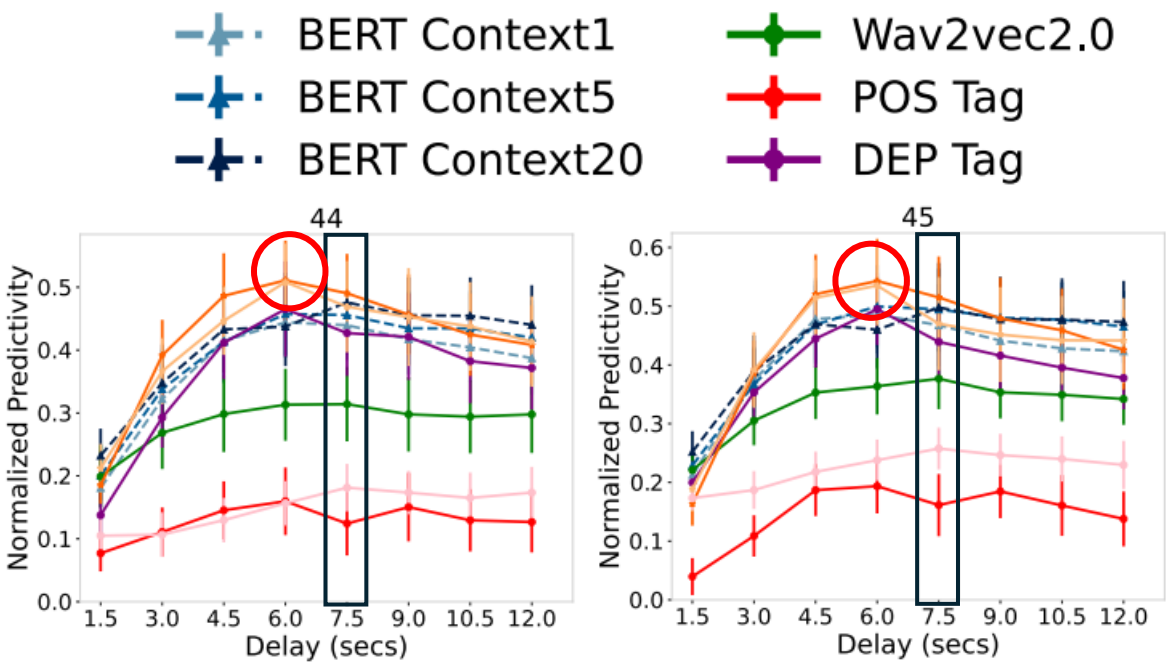
How language processing within the brain is impacted by delays in the Hemodynamic Response Function (HRF)?



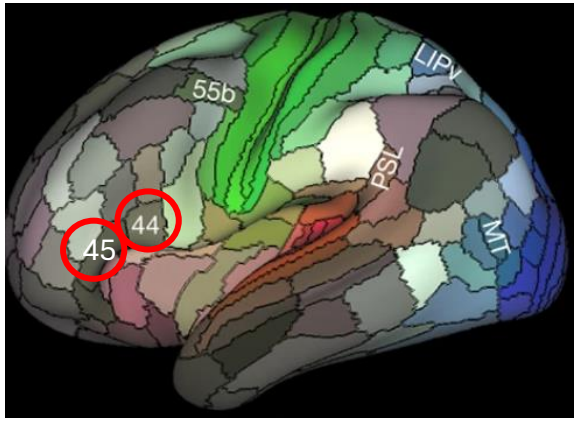
↓Models / Delays→	D1	D2	D3	D4	D5	D6	D7	D8
BERT Context1	15.58*	28.23*	40.06*	45.46*	**47.86**	47.57	46.36	45.58
BERT Context5	17.14*	28.75*	41.41*	47.44	49.88	49.1	**50.85**	50.69
BERT Context20	22.83*	34.05*	44.67*	46.0*	53.62	53.0	**53.81**	53.42
Wav2vec2.0	25.2*	34.22*	41.45*	44.57*	**47.12**	45.94	46.24	46.14
POS Tag	5.82*	11.2*	16.35	18.55	17.77	**19.14**	18.5	16.44
DEP Tag	17.31*	31.8*	42.98*	**50.53**	50.02	50.35	46.76	45.37*
CC	15.24*	30.45*	43.66*	47.91	**48.08**	47.55	45.88	44.18*
CI	16.08*	29.57*	41.81*	48.18	**48.30**	47.8	46.66	46.19
Basic Speech	17.86*	21.49*	24.98*	27.6*	29.53	**29.62**	28.93	29.34

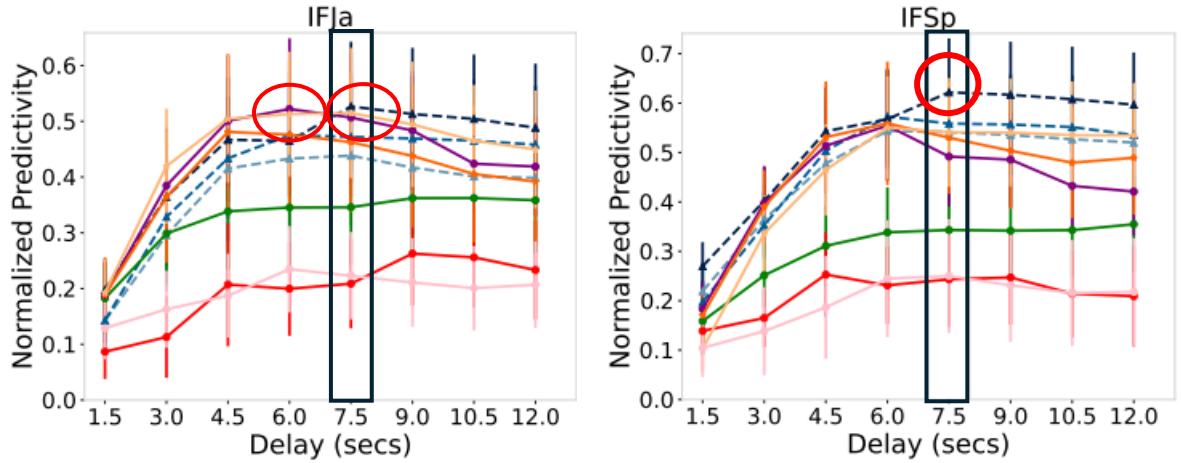
- Syntactic embeddings, including CC and CI, show higher brain activity in the early delays, particularly D4, with a decrease in activity at later delays (D6-D8).
- BERT wth Context 20 performs the best, implying that brain predictivity improves with increasing context length.

Can we distinguish syntax and semantics by varying delays?

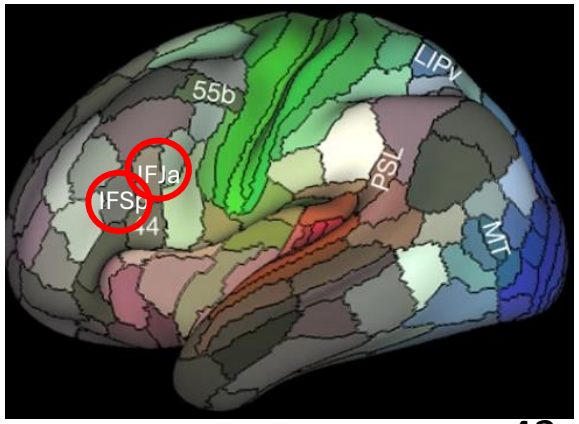


- Higher normalized predictivity is observed for syntactic embeddings at D4 for 44 and 45 regions





- IFJa region process syntax in early delays and semantics in later delays
- IFSp region process semantics in later delays (D5-D8)



Partial Conclusions

In ecological setting, our findings are consistent with Hierarchy of language processing (Matchin & Hickok)

Different optimal HRF delays for processing of syntax (6 secs) and semantics (> 7.5 secs) at early and later delays, respectively

Detailed region and sub-region analysis reveal that longer context may play a significant role in higher HRF delays

Research questions

Language models predict the next word from nearby words, but semantics are crucial for language comprehension.

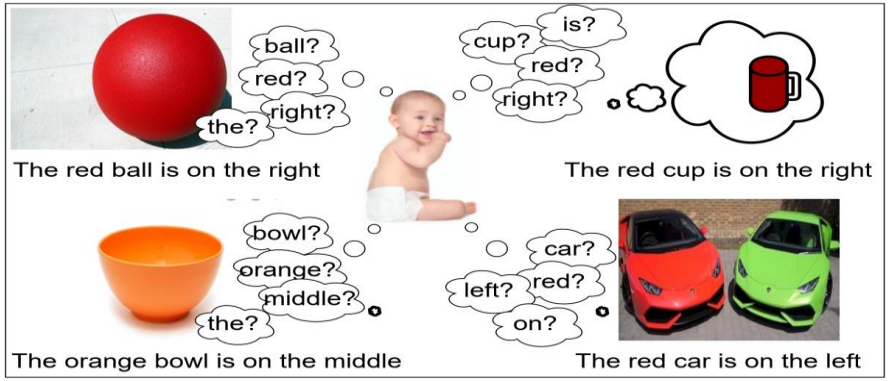
Is the **how brain utilizes context** through time to process words **during narrative story listening**?

Do **language models** that process **longer sequences** align better with human participants during **long narratives story listening**?

fMRI measures BOLD suffers from delay due to Hemodynamic Response Function

How: ???

Do diverse language regions within the brain is **impacted by delays?**

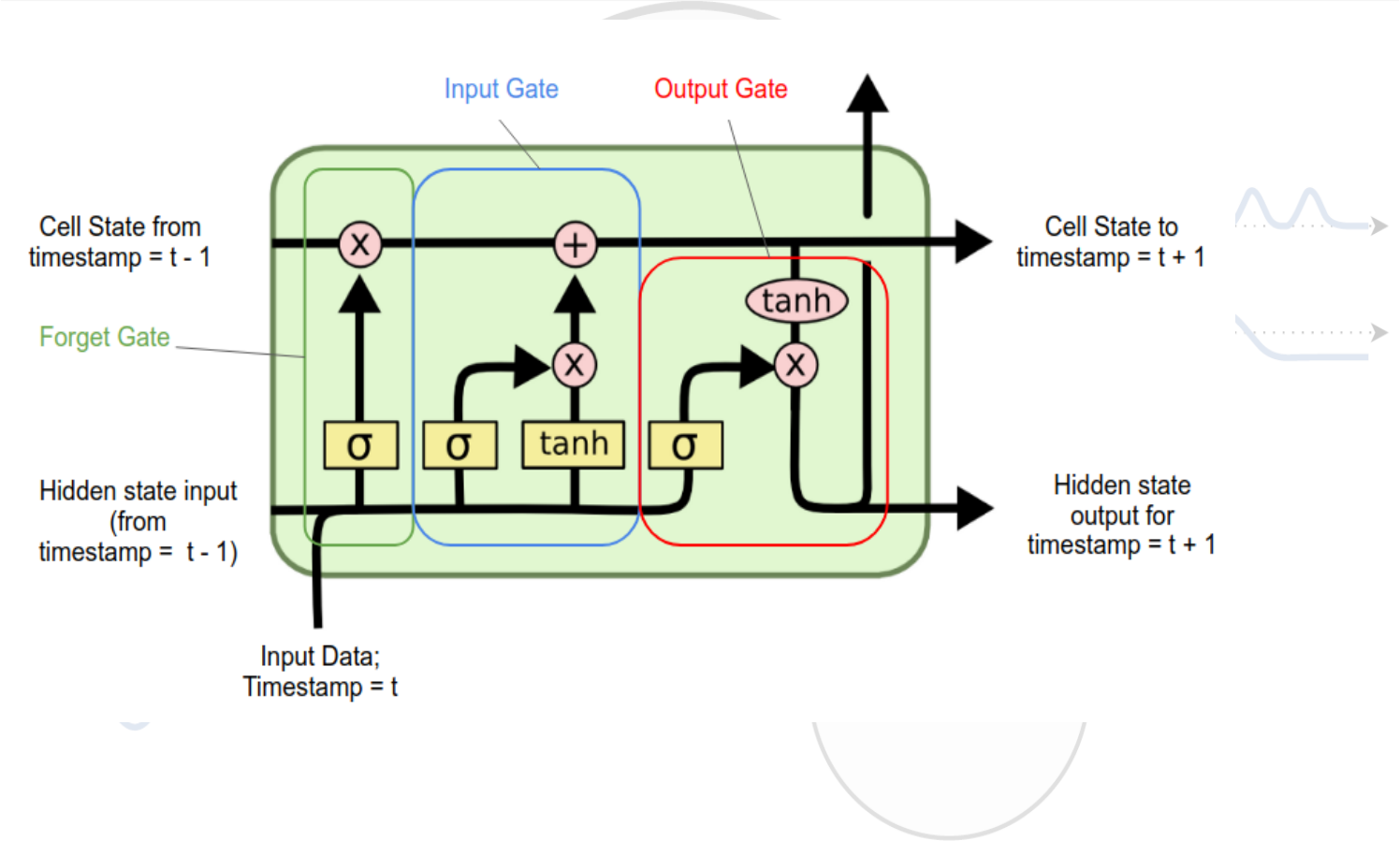


Grounded language acquisition: how infants can learn language by observing their environments.

How does the **model's ability** to learn **semantic concepts** through cross-situational learning in **noisy supervision?**

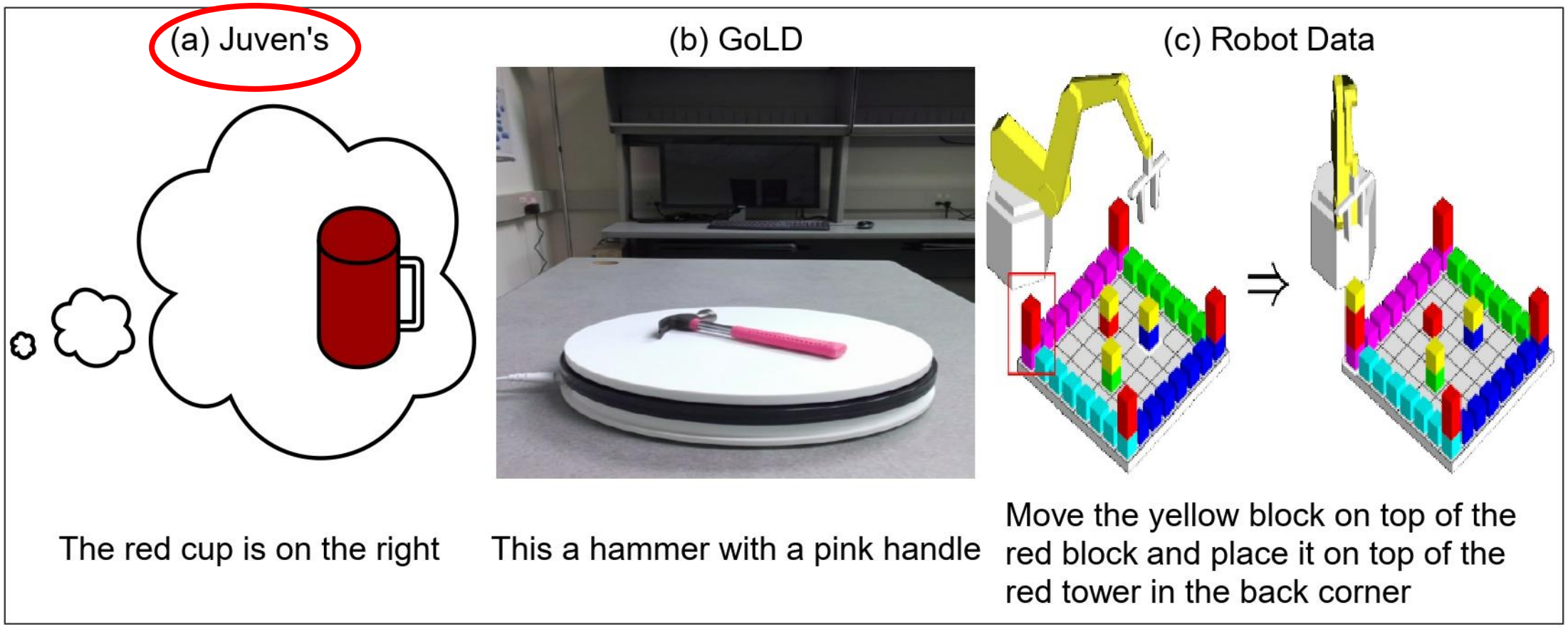
How language models can perform grounded language acquisition?

- We employ CSL task using two sequence-based language models:
  - Echo State Networks (i.e. Reservoir Computing)
  - Long Short-Term Memory Networks (LSTM)



Grounded language datasets

Dataset with simple sentences



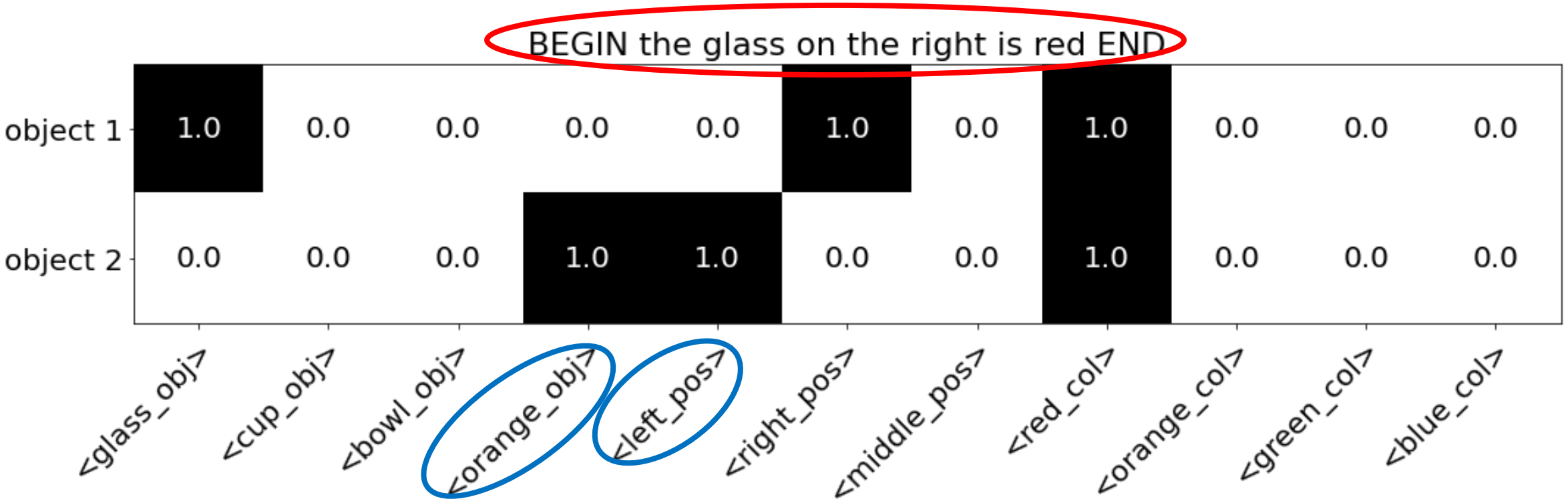
Evaluation metric

"the cup is on the right"

Imagined vision	is valid ?	is exact ?
(a) 	✗	✗
(b) 	✓	✗
(c) 	✓	✓

CSL Task: Noisy supervision output

- The target is a noisy supervision vector that contains additional concepts that are not present in the input sentence



Results

	Juven's CSL Data		GoLD Data		Robot Data
Model	**Valid**	**Exact**	**Valid**	**Exact**	**Exact**
ESN-offline + One-Hot	46.60±0.27	63.30±0.35	29.49±0.25	30.38±0.45	42.30±0.14
ESN-offline + GloVe	44.40±0.31	61.00±0.37	48.93±0.28	53.90±0.24	57.42±0.23
ESN-offline + fine-tuned BERT	20.70±0.16	40.20±0.18	44.57±0.26	47.48±0.41	43.00±0.11
ESN-offline + BERT	24.50±0.20	43.60±0.24	52.20±0.24	54.78±0.35	45.50±0.14
ESN-online FL + One-Hot	02.90±0.01	29.40±0.24	19.23±0.22	**26.92±0.29**	37.12±0.06
ESN-online FL + GloVe	06.00±0.07	40.20±0.31	20.27±0.26	32.56±0.24	38.09±0.14
ESN-online FL + fine-tuned BERT	**02.52±0.01**	**26.00±0.18**	**17.45±0.11**	28.89±0.19	**34.20±0.06**
ESN-online FL + BERT	02.72±0.01	28.50±0.20	27.24±0.12	54.40±0.21	35.34±0.10
ESN-online CL + One-Hot	18.64±0.13	39.52±0.31	21.69±0.46	32.48±0.48	57.10±0.55
ESN-online CL + GloVe	42.60±0.56	72.90±1.01	22.14±0.64	36.42±0.76	59.96±0.64
ESN-online CL + fine-tuned BERT	27.28±0.19	54.00±0.34	18.37±0.40	34.04±0.28	58.86±0.20
ESN-online CL + BERT	32.86±0.20	60.88±0.41	22.30±0.46	52.49±0.44	60.17±0.33
RandLSTM + One-Hot	100.0±0.0	100.0±0.0	71.11±1.61	75.34±1.82	79.53±1.51
RandLSTM + GloVe	100.0±0.0	100.0±0.0	84.48±2.32	84.83±2.10	88.88±1.04
RandLSTM + fine-tuned BERT	100.0±0.0	100.0±0.0	72.02±1.64	72.02±2.03	87.34±0.89
RandLSTM + BERT	100.0±0.0	100.0±0.0	76.31±1.45	80.17±1.67	87.91±1.21
LSTM + One-Hot	99.64±0.01	99.82±0.01	42.89±0.56	48.14±0.65	75.67±0.54
LSTM + GloVe	99.20±0.01	99.99±0.00	65.18±0.84	70.89±0.91	86.57±0.87
LSTM + fine-tuned BERT	97.84±0.01	98.90±0.01	44.18±0.46	47.26±0.68	72.47±0.41
LSTM + BERT	98.10±0.01	99.99±0.01	48.28±0.44	52.40±0.66	78.60±0.45

(b) Complex corpora

- LSTM outperforms the ESN on both Valid and Exact errors on Juven's

ESN-online FL outperformed all the models

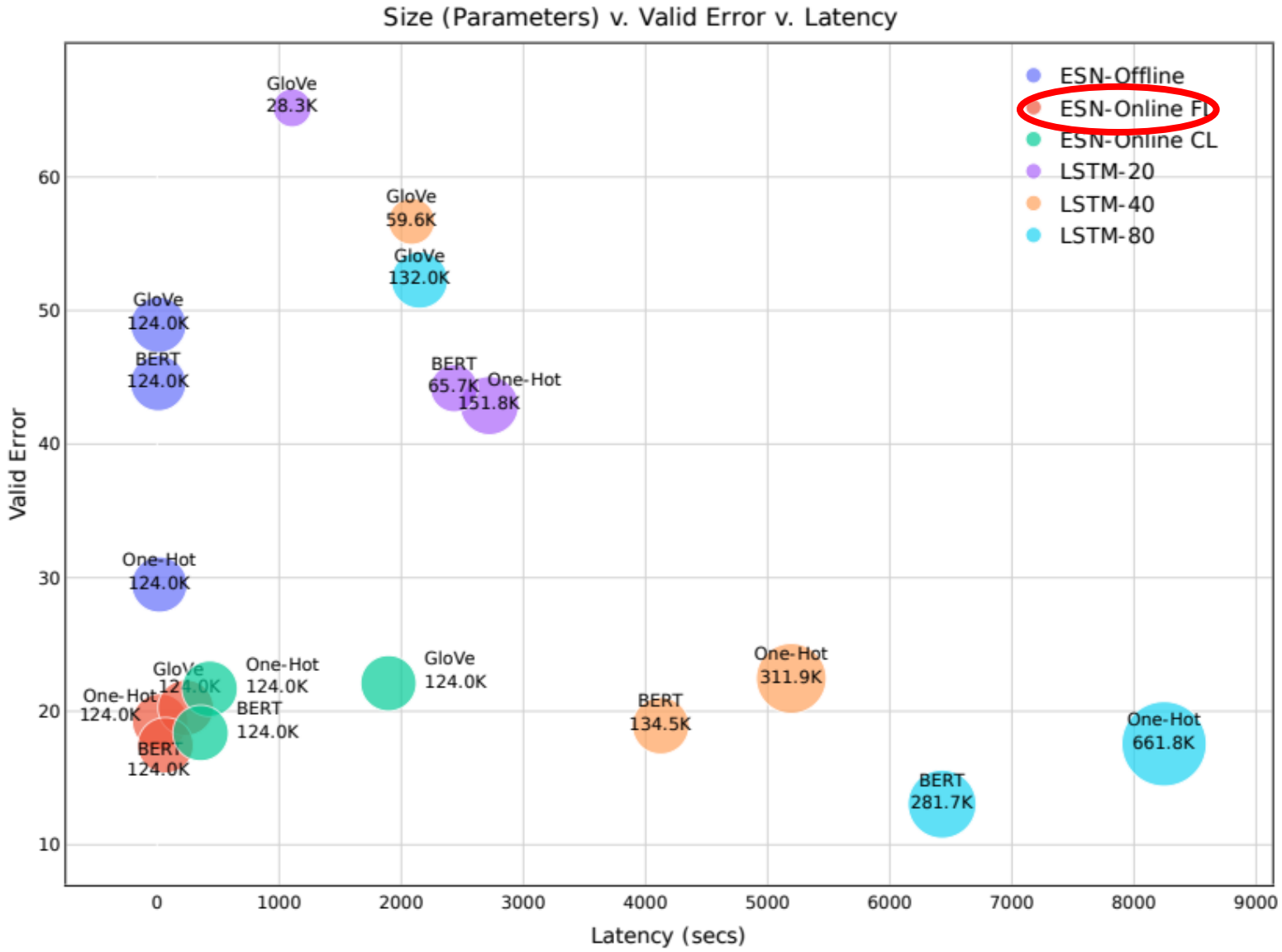
significantly better than ESN-online CL and offline methods.

(a) Reduced-size corpora

maller vocabulary of objects: (4 or Juven's; 10 for GoLD

Larger vocabulary of objects (50 for Juven's; 47 for GoLD, 11 for Robot Data)

Parameters vs. Valid Error vs. Training Latency



GoLD (Large Data)

ESN-online FL model showcases lower valid error using 124K parameters with a model training latency of 64 seconds

ESN model has better computational complexity in terms of latency and model size.

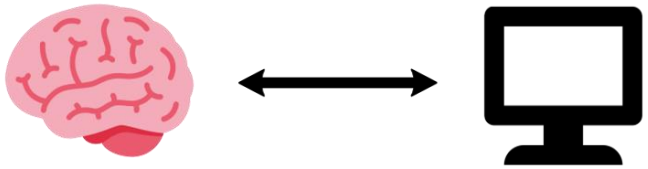
Partial Conclusions

Biologically plausible ESNs have a better trade-off on all three grounded language datasets with better prediction error and low latency.

Fine-tuned BERT representations are more efficient at capturing complex relationship between words.

ESNs with online learning models are making better predictions during the processing of a sentence.

Implications to Neuro-AI



NLP -> Neurolingustics

- Disentangling representations allowed us to distinguish syntactic and semantic peaks across language regions at different HRF delays
- Contextual representations allowed us to indicate that alignment with MEG depends on past context.

Neurolingustics -> NLP

- Need LM with deeper understanding of how humans relate characters, discourse during long narratives
- Need LM models to better evaluate meaningful explainable variance for individual participants
- Need better speech models for end-to-end language comprehension.

Thank You

Questions?