COMPARING MULTILINGUAL NLP MODELS TO THE HUMAN BRAIN

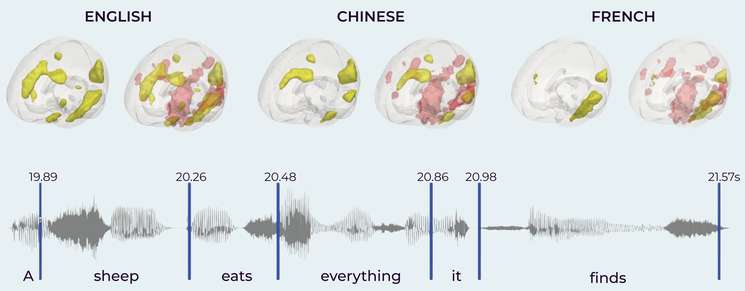
University of Amsterdam

BY DAVID FRÜHBUS ET AL.
Supervisor: Rochelle Choenni

FMRI

LE PETIT PRINCE MULTILINGUAL FMRI CORPUS

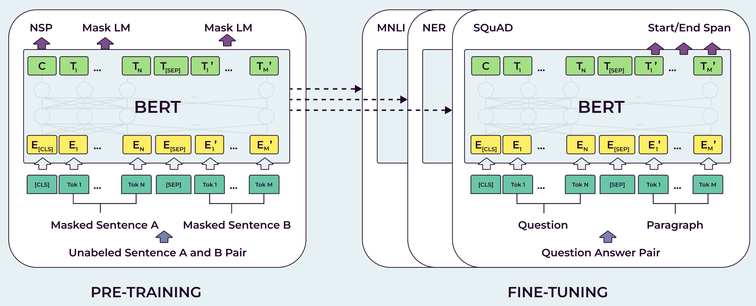
A multilingual dataset consisting of fMRI data from 112 participants listening to the audiobook "The Little Prince" in English, Chinese or French.



LANGUAGE MODEL

GETTING XLM-ROBERTA SENTENCE REPRESENTATIONS

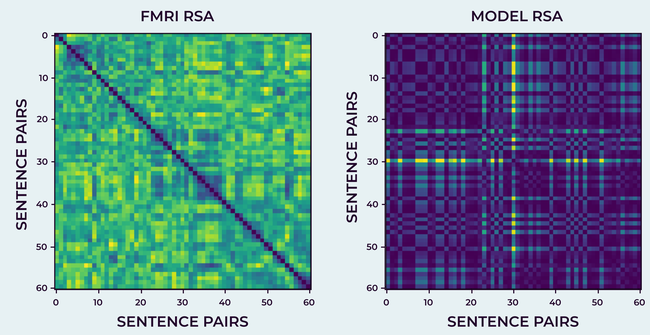
To get the model representations, we got the hidden state from all layers of XLM-RoBERTa after doing a forward pass on pairs of sentences from the Le Petit Prince text.



RSA

COMPUTING RSA MATRICES

We compute the Representational Similarity Matrices with the pairwise cosine similarity between representations of sentence chunks for FMRI and Model data.



ANALYSIS

Examine the brain representations in response to linguistic stimuli in French, Chinese, and English within the fMRI data.

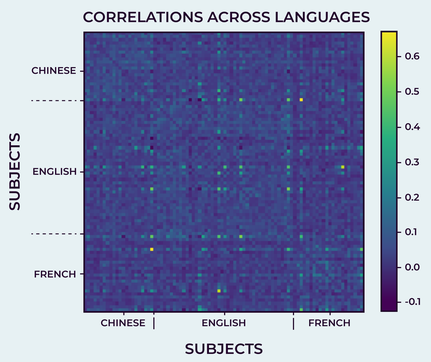
Retrieve and analyse the hidden representations from a multilingual NLP model processing linguistic stimuli in multiple languages.

Align the neural and brain representations across time and compute RSAs for subjects and model layers.

Compare the Pearson correlation between the RSAs of subjects and model layers across languages.

HUMAN MEETS HUMAN

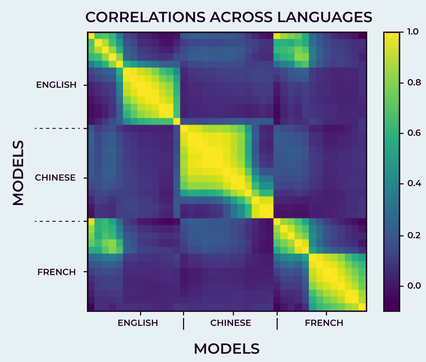
CORRELATION OF SUBJECT RSA'S ACROSS LANGUAGES



	CHINESE	ENGLISH	FRENCH
CHINESE	0.054912	0.038489	0.027926
ENGLISH	0.038489	0.047469	0.025302
FRENCH	0.027926	0.025302	0.041955

AI MEETS AI

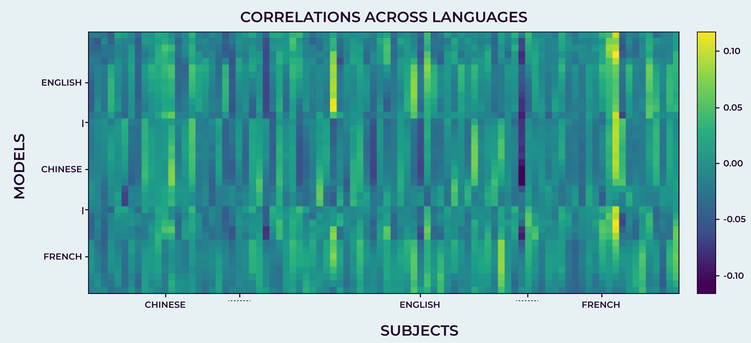
CORRELATION OF MODEL RSA'S ACROSS LAYERS AND LANGUAGES



	CHINESE	ENGLISH	FRENCH
CHINESE	0.495141	0.099466	0.177084
ENGLISH	0.099466	0.714532	0.075476
FRENCH	0.177084	0.075476	0.527678

AI MEETS HUMAN

CORRELATION OF MODEL AND SUBJECT RSA'S ACROSS LANGUAGES



	SUBJ ENGLISH	SUBJ CHINESE	SUBJ FRENCH
MODEL ENGLISH	0.037965	0.054821	0.022329
MODEL CHINESE	0.042433	0.048675	0.032421
MODEL FRENCH	0.045134	0.040563	0.026835

FINDINGS

For both model-to-model and subject-to-subject cases, **inter-correlation is stronger than any cross-correlation**. For model-to-model, Chinese is the most inter-correlated, while French and English are more cross-correlated.

Early and late-stage model layers have strong inter-correlation, linked to syntactic and semantic representations. **English and French show strong early-layer correlation, implying syntactic similarity**. Chinese syntax requires more layers, suggesting higher syntactic complexity.

The similarity between subjects and model layers follows the similarity seen in brain representations. Intra-language correlation is not higher than cross-language correlation. The strongest correlation is between the English model and Chinese brain representations.

FUTURE

Combine RSA with sentiment classification to analyze the influence of emotion on brain representations.

Compare brain representations with neural representations from multimodal (audio-language) models.

Extend the cross-linguistic analysis by incorporating more languages, especially those from different language families.

We used a BERT-based model, while future research could compare brain activations with GPT-based models.

REFERENCES

[1] Abnar, S., Beinborn, L., Choenni, R., & Zuidema, J. (2019). Blackbox Meets Blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains.

[2] Li, Jixing, et al. "Le Petit Prince multilingual naturalistic fMRI corpus." Scientific data 9.1 (2022): 530.

SCAN ME!