Multilingual Computational Models Reveal Shared Brain Responses to 21 Languages

Andrea Gregor de Varda[1,2], Saima Malik-Moraleda[2,3], Greta Tuckute[2], Evelina Fedorenko[2,3]

[1]University of Milano-Bicocca [2]Massachusetts Institute of Technology [3]Harvard University

Abstract

At the heart of language neuroscience lies a fundamental question: How does the human brain process the rich variety of languages? Recent developments in Natural Language Processing, particularly in multilingual neural network language models, offer a promising avenue to answer this question by providing a theory-agnostic way of representing linguistic content across languages. Our study leverages these advances to ask how the brains of native speakers of 21 languages respond to linguistic stimuli, and to what extent linguistic representations are similar across languages. We combined existing (12 languages across 4 language families; n=24 participants) and newly collected fMRI data (9 languages across 4 language families; n=27 participants) to evaluate a series of encoding models predicting brain activity in the language network based on representations from diverse multilingual language models (20 models across 8 model classes). We found evidence of cross-lingual robustness in the alignment between language representations in artificial and biological neural networks. Critically, we showed that the encoding models can be transferred zero-shot across languages, so that a model trained to predict brain activity in a set of languages can account for brain responses in a held-out language, even across language families. These results imply a shared component in the processing of different languages, plausibly related to a shared meaning space.

Acknowledgements: We are grateful to Hee So Kim and Agata Wolna for assisting with participant recruitment; to Megha Vemuri, Alexander Fung, and Sara Swords for their assistance with data collection; and to Ben Lipkin for help with preprocessing and curating the Natural Stories dataset. GT was supported by a Friends of McGovern Graduate Fellowship and the K. Lisa Yang ICoN Center Graduate Fellowship; EF was supported by NIH award U01-NS121471 and funds from the McGovern Institute for Brain Research, the Brain and Cognitive Sciences Department, the Simons Center for the Social Brain, and the MIT Quest for Intelligence.

Contributions: **AGdV**: Conceptualization, Methodology, Investigation, Data curation, Visualization, Formal analysis, Validation, Software, Writing-Original draft; **SMM**: Investigation, Data curation, Writing-Review and editing; **GT**: Investigation, Data curation, Writing-Review and editing; **EF**: Writing-Review and editing, Supervision, Project administration.

Introduction

Natural languages vary remarkably in how they convey meaning; this variety spans all levels of linguistic structure, from phonology to syntax (Evans & Levinson, 2009). Consider, for instance, the intricate verb conjugations in Italian, which convey tense, aspect, gender, and number, contrasting sharply with the more context-reliant structures in Mandarin, where verb forms are not inflected. In Finnish, the extensive case system can express complex structural relationships between nouns and other sentence elements, a feature largely absent in English. And the tonal modulations of Vietnamese create meanings distinguished by pitch, a concept foreign to non-tonal languages. Despite this variability, all human languages allow their speakers to express almost any idea in a sequence of words, thus constituting the most sophisticated communication system in the animal kingdom. Here, we test whether the human language network—a set of brain areas that support human linguistic ability (Fedorenko et al., 2024a)—encodes linguistic content in a consistent way across diverse languages.

Recent evidence has shown that the language network exhibits similar topography, including a left-hemispheric bias, strong inter-regional functional integration, and functional selectivity for language relative to non-linguistic inputs and tasks across languages and language families (Malik-Moraleda, Ayyash, et al., 2022). However, the nature of linguistic representations within and across languages remains an open question. A core challenge in addressing this question is to quantitatively capture linguistic information consistently across languages. Recent progress in multilingual Natural Language Processing (NLP) research has provided a promising approach to represent linguistic content in a language-general format, making it possible to project words and sentences onto comparable representation spaces across languages. Such advances have been driven by the introduction of Multilingual Neural Network Language Models (MNNLMs). MNNLMs are transformer-based language models trained simultaneously on text in several languages (up to 200 languages in the case of NLLB-200; NLLB Team, 2024). Some of these models, such as machine translation models (NLLB Team, 2024), are provided with explicit cross-lingual supervision during training, so that their training data comprises translation-equivalent sentences that facilitate cross-lingual alignment. For other models (e.g., mBERT; Devlin et al., 2019), cross-lingual correspondences are not explicitly provided in the training data, but are instead implicitly learned from the statistical regularities across languages. These different models have shown remarkable ability in learning language-independent representation spaces, suggesting potential universality in linguistic feature encoding, including semantic and syntactic information (Chi et al., 2020; de Varda & Marelli, 2023, 2024; Stanczak et al., 2022). An appeal of these language-general linguistic representations is that they are constructed in a data-driven way without any potential bias that may be introduced by a particular linguistic theory of framework. To quantitatively capture linguistic content across languages, our study employs contextualized word and sentence embeddings derived from MNNLMs as representational spaces to understand the brain's processing of linguistic inputs in diverse languages.

Several studies have now demonstrated that representations from neural network language models can successfully account for brain responses in the human language network (Antonello & Huth, 2024; Aw & Toneva, 2023; Caucheteux & King, 2022; Goldstein et al., 2022; Schrimpf et al., 2021; Tuckute et al.,

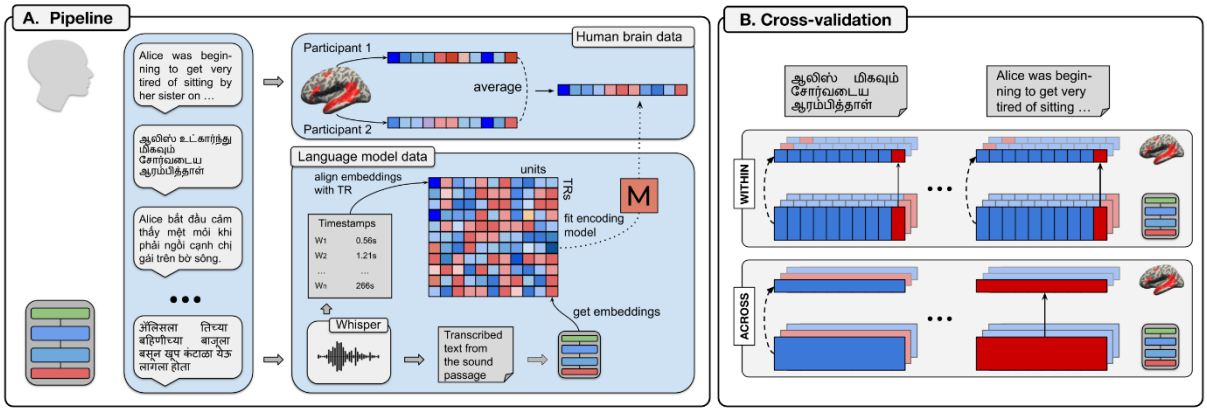
2024a, 2024b, *inter alia*). These findings have been taken as evidence of the sensitivity of the human language processor to the distributional statistics of language and its engagement in predictive processing. However—similar to other work in the cognitive neuroscience of language (see Malik-Moraleda, Ayyash et al., 2022 for discussion; also, Blasi et al., 2022; Forgey, 2024)—these studies have been mostly conducted in English, with a few exceptions involving Dutch, French, Japanese, and Chinese (Caucheteux & King, 2022; Millet et al., 2023; Nakagi et al., 2024; Sun et al., 2023; Yamashita et al., 2023). Thus, this research area is far from covering the vast diversity that characterizes natural languages. To fill this gap, we first assessed whether MNNLMs could predict brain responses in speakers of different languages, training several encoding models encompassing a relatively large sample of natural languages and neural model architectures. We also assessed whether some relevant model features similarly modulated their encoding performance in different languages. In particular, previous research in computational language neuroscience has established that a better fit to brain data is obtained (a) by language models with better performance on next-word prediction (Schrimpf et al., 2021; Caucheteux & King, 2022), underscoring the central role of predictive processing in language comprehension (cf. Antonello & Huth, 2024); (b) by auto-regressive (or causal) language models trained to predict upcoming words, as opposed to bidirectional encoders or encoder-decoder models (Schrimpf et al., 2021), stressing the importance of incrementality of language processing; and (c) by larger-capacity neural architectures (Schrimpf et al., 2021; Antonello et al., 2024), highlighting the representational richness of human linguistic representations. In our study, we tested whether these patterns are idiosyncratic to English or generalize to diverse languages, thus plausibly reflecting key computational principles of human language processing.

Beyond testing the cross-lingual replicability of previous findings in computational language neuroscience, we investigated our core question: whether linguistic representations are similar across languages. Leveraging existing fMRI data from a passage-listening task in 12 languages (Malik-Moraleda, Ayyash, et al., 2022), we trained encoding models to predict brain responses based on multiple languages and transferred them zero-shot to a new language on which they had not been trained (Study I). Next, we tested the generalizability of the encoding models under more diverse and stringent conditions (Study II): we trained the models on existing fMRI datasets varying in the number of languages, presentation modalities, and types of language stimuli (Blank et al., 2014; Malik-Moraleda, Ayyash, et al., 2022; Pereira et al., 2018; Tuckute et al., 2024a), and then applied them directly to newly collected fMRI data in 9 additional languages, none of which were included in the training data. To foreshadow our key finding, the two studies converged in demonstrating that neural encoding models can robustly predict brain activity across diverse languages, providing evidence for a shared component of linguistic representations.

Results

Study I

Malik-Moraleda, Ayyash, et al. (2022) used fMRI to record brain responses in speakers of diverse languages listening to a ~4.5-minute-long passage from the book *Alice in Wonderland* (Carrol, 1865) in their native language. We used data from 12 languages (Afrikaans, Dutch, Farsi, French, Lithuanian, Marathi, Norwe-



C. Datasets

		Study I	Study II				
			Train				Test
		Experiment I data	Experiment I data	NatStories	Pereira2018	Tuckute2024	Experiment II data
Participants		24 (2 per language)	24 (2 per language)	97 (28.44 per story)	10 (7.84 per sentence)	10 (10 per sentence)	27 (3 per language)
Materials	Modality	Auditory	Auditory	Auditory	Visual	Visual	Auditory
	Stimuli	1 passage per language 130 volumes/language 592.25 tokens/language 344.83 types/language	1 passage per language 130 volumes/language 592.25 tokens/language 344.83 types/language	9 passages 1,516 volumes 10,206 tokens 3,037 types	627 sentences 627 volumes 7,711 tokens 2,968 types	1,000 sentences 1,000 volumes 6,000 tokens 2,894 types	3 passages per language 390 volumes/language 1,751.11 tokens/lang. 851.78 types/language
	Reliability	r = 0.32	r = 0.32	r = 0.59	r = 0.35	r = 0.30	r = 0.30
	Languages	N = 12 Afrikaans, Dutch, Farsi, French, Lithuanian, Marathi, Norwegian, Romanian, Spanish, Tamil, Turkish, Vietnamese	N = 12 Afrikaans, Dutch, Farsi, French, Lithuanian, Marathi, Norwegian, Romanian, Spanish, Tamil, Turkish, Vietnamese	English	English	English	N = 9 Italian, Portuguese, German, Russian, Polish, Hindi, Arabic, Mandarin, Korean
	Similarity	Embed. similarity = 0.56	Embed. similarity = 0.55 Word overlap = 0.49	Embed. similarity = 0.53 Word overlap = 0.09	Embed. similarity = 0.48 Word overlap = 0.06	Embed. similarity = 0.49 Word overlap = 0.05	N.A.

Figure 1. Graphical summary of the study methods. **A. Experimental pipeline:** Passages from the book *Alice in Wonderland* are presented to human participants while in the fMRI scanner (top) and MNNLMs (bottom). *Human brain data (top):* Two participants listen to the passage in their native language. The fMRI time-series obtained from the two participants' language areas are averaged. *Language model data (bottom):* Written transcriptions of the passages and relative timestamps from the raw audio file are retrieved using the speech-to-text model Whisper-timestamped (Radford et al., 2023; Louradour, 2023). The MNNLM produces embedding representations for the words in the passages, which are then temporally aligned with the fMRI time-series. An encoding model is fitted to predict fMRI responses on the basis of the MNNLM embeddings. **B. Encoding model cross-validation:** The encoding models are evaluated either on a left-out data subset within each language, or a left-out language. *Within-language cross-validation (top):* The encoding models are trained within each language with 10-fold cross-validation. During training, MNNLM embeddings are mapped onto brain responses in nine data folds (blue blocks, dashed arrows). The encoding models then predict brain activity in the remaining fold (red blocks, solid arrow). *Across-languages cross-validation (bottom):* The encoding models are trained on the complete brain data in all but one language, and are transferred zero-shot to the held-out language. **C. Datasets:** Datasets employed in the two studies. For each data source, the table reports the number of participants, presentation modality, information on the stimuli (number of passages/sentences, brain measurements (volumes), tokens, and word types, i.e., unique tokens), split-half reliability, and language coverage. The last row indicates the similarity in the linguistic materials between train and test splits. In the case of Study I, this measure reports the similarity between the word embeddings of each passage across languages. In the case of Study II, it indicates the similarity between the source materials used as training data for the encoding models and the test data (in English), considering both embedding similarity and word overlap. The similarity scores based on the embeddings are averaged across words, sentences/passages, languages, layers, and models.

-gian, Romanian, Spanish, Tamil, Turkish, Vietnamese) belonging to 4 language families (Indo-European, Dravidian, Turkic, Austroasiatic). Our sample encompassed high-, mid-, and low-resource languages, differing in the extent to which they have been previously studied in language neuroscience (SI S17). In most languages, participants (n=2 per language) listened to the same passage, but in Afrikaans, Dutch, and French, a different passage from the same book was used. These data were selected from a larger dataset, restricting to languages for which inter-participant blood-oxygen-level-dependent (BOLD) signal correlations in the left-hemisphere language areas (functionally defined; Fedorenko et al., 2010) were reliable (see Methods, "Study I – Time-series extraction").

Our general procedure was similar to that in past studies (e.g., Antonello & Huth, 2024; Aw & Toneva, 2023; Schrimpf et al., 2021; Tuckute et al., 2024a) (**Figure 1A**) and consisted of extracting contextualized word embeddings from the language models for the stimuli in each language, and linearly mapping these representations onto fMRI activity in response to those linguistic stimuli via ridge regression. For each language, we extracted BOLD time-series during the passage in the five core areas of the left-hemisphere fronto-temporal language network of each participant (Methods, "fMRI data modeling and definition of functional regions of interest (fROIs)"; see Fedorenko et al., 2024a for a review on the language network). We averaged the time-series across the voxels within each of the five functional regions of interest (fROIs), then across fROIs, and lastly, across the two participants, to obtain a univariate time-series for each language (**Figure 1A**, top). To obtain MNNLM embeddings for the same stimuli (the passages in the 12 languages), we first derived textual transcriptions of the auditory recordings with Whisper-timestamped (Radford et al., 2023; Louradour, 2023), a multilingual speech recognition model that outputs word-by-word timestamps. We then extracted embedding representations of each word in the passage transcriptions from 20 candidate MNNLMs, including models trained on translation (NLLB), contrastive cross-lingual alignment (XLM-Align, InfoXLM), span corruption (i.e., predicting consecutive chunks of text based on bidirectional context; mT5), masked language modeling (mBERT, XLM-R), and causal language modeling (mGPT, XGLM). The models were based on different architectures, encompassing encoder-only, decoder-only, and encoder-decoder models (see Methods, "Language models" and Table 1 for a summary of all the models considered in the study). Next, we aligned the MNNLM contextualized word embeddings with the fMRI time-series using the time stamps from Whisper-timestamped (**Figure 1A**, bottom). And finally, we fitted several linear encoding models to predict the brain responses from the MNNLM embeddings. The encoding models were then evaluated either on left-out data subsets or left-out languages.

Language models predict brain responses in diverse languages

Most past work relating model representations to brain data has relied on data from English and a small number of other high-resource languages, such as Dutch, French, Japanese, and Chinese (Caucheteux & King, 2022; Millet et al., 2023; Nakagi et al., 2024; Sun et al., 2023; Yamashita et al., 2023). To test the ability of language models to capture brain responses for a wider array of languages, we fitted an encoding model within each language and evaluated it with sequential 10-fold cross-validation (henceforth the WITHIN condition; see **Figure 1B**, top for a schematic representation of the cross-validation scheme). The encoding

models' performance was evaluated as the Pearson correlation between the predicted language network response and the actual fMRI time-series. We calculated statistical significance by comparing the performance of the encoding models with a shuffled baseline where we randomized the correspondence between MNNLM embeddings and brain data (Methods, "Significance testing").

Our results show that MNNLMs can successfully predict the language network's response to diverse languages, as recorded with fMRI (**Figure 2A**, top; see SI S2 for the results in each individual language and SI S4 for the results of a similar analysis on the 9 languages from Study II), significantly outperforming the randomized baselines for all models (all $p < 0.001$). Causal (or auto-regressive) MNNLMs – i.e., MNNLMs trained to predict the next word in a text sequence – achieve the best brain encoding performance, generally outperforming models trained on translation, cross-lingual contrastive learning, span corruption, and masked language modeling, in line with previous observations based on data from English (Schrimpf et al., 2021). Furthermore, within each model family, larger-capacity models outperform their smaller counterparts in predicting brain responses. Indeed, the best average performance is obtained by XGLM_{xl} ($r = 0.50$, $SE = 0.10$), the largest auto-regressive model considered in the study; its encoding score is immediately followed by the other large-scale, autoregressive MNNLMs ($\text{XGLM}_{\text{large}}$, $r = 0.49$, $SE = 0.11$; mGPT, $r = 0.46$, $SE = 0.07$). The only exception to this trend is $\text{XGLM}_{\text{small}}$ ($r = 0.39$, $SE = 0.12$), which performs similarly to XGLM_{med} ($r = 0.39$, $SE = 0.10$). Encoder-decoder models, both trained on translation (NLLB family) and span corruption (mT5 family), are associated with lower average performance, demonstrating reduced effectiveness in capturing language-specific brain responses compared to other model types.

The WITHIN encoding performance is critically influenced by the depth of the layer within the model. **Figure 2C** displays the encoding performance scores for each layer, normalized by their position within the model. For most model families, intermediate layers (as seen in XLM-Align, InfoXLM, XLM-R, DeBERTa) and deeper layers (as observed in mBERT, mT5, mGPT) demonstrate the highest predictivity in relation to the fMRI signal, also in line with prior work (Schrimpf et al., 2021; Antonello et al., 2023). The layer-wise performance patterns are less clear for the NLLB and XGLM models. Encoding scores averaged across models increase until a normalized depth of ~.70, where they reach a plateau (**Figure 2C**, top right).

Encoding models can be transferred across languages

To test for the existence of a shared component of linguistic representations *across* languages, we trained the encoding models on the data in all but one language, and tested them zero-shot on the left-out language (the ACROSS condition; see **Figure 1B**, bottom for a visualization of the cross-validation scheme).

Encoding models can be successfully transferred zero-shot to unseen languages, highlighting the strong generalizability of the link between distributed language representations learned from statistical word co-occurrence patterns and brain responses in the human language network (**Figure 2A**, bottom). Despite achieving lower average predictivity compared to the WITHIN condition (see comparison with Figure 2A, top), all encoding models significantly outperform the randomized baselines (all $p < 0.001$). The best performance is achieved by a masked language model from the XLM-R family ($\text{XLM-R}_{\text{large}}$, $r = 0.35$, $SE = 0.05$), but inter-model differences in encoding performance are less pronounced than in the WITHIN condition.

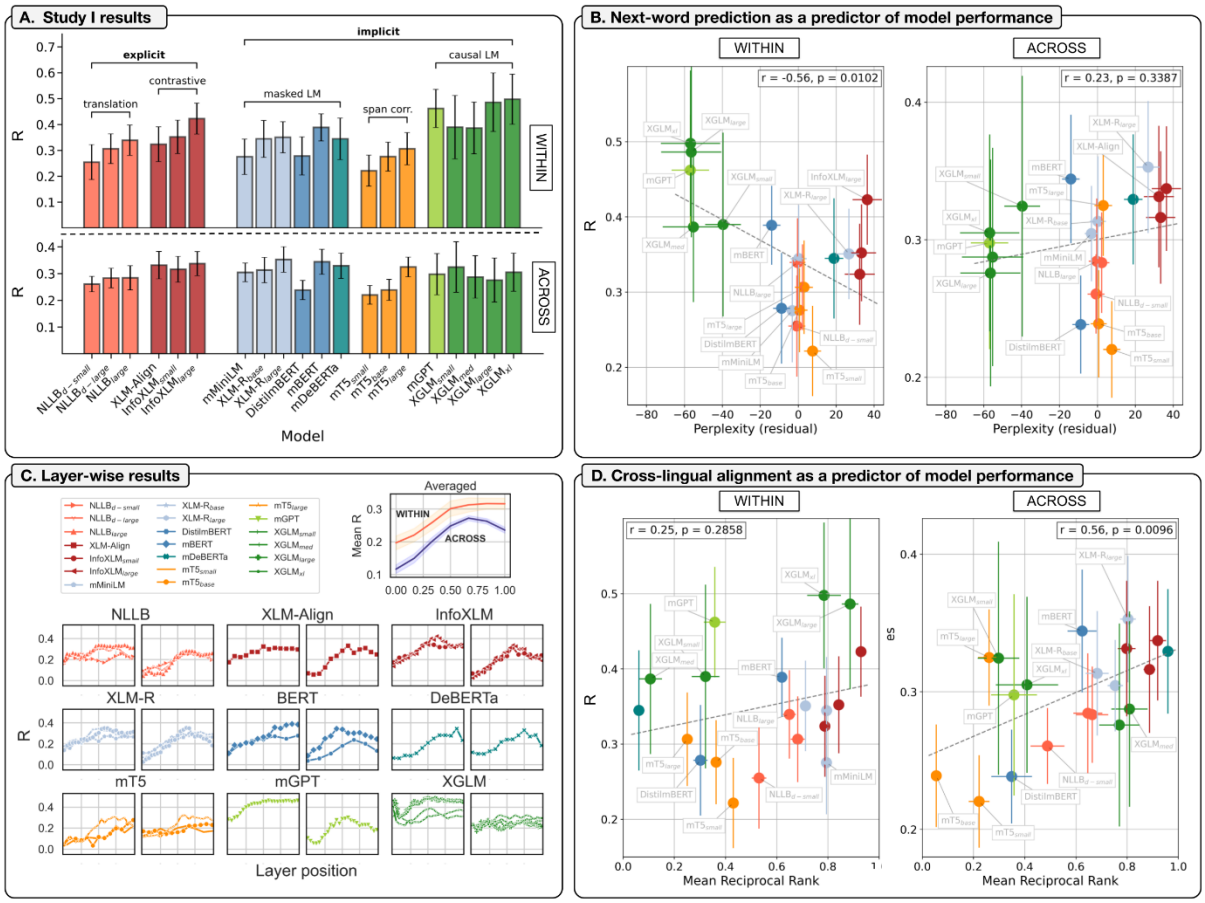


Figure 2. Summary of the results of Study I. **A. Study I results:** *Top:* Best-layer WITHIN encoding performance by model, averaged across languages. The error bars indicate the standard error of the encoding performance across languages. The models are first grouped based on whether they are exposed to implicit or explicit cross-lingual signal during training, and then by their training objective. Models in the same colors belong to the same model family. Causal (or auto-regressive) language models achieve the strongest brain encoding performance; within each model family, larger-capacity models are more effective in predicting brain activity. *Bottom:* Best-layer ACROSS encoding performance by model type. See SI S5 for the results obtained with the median-performing layer in the two conditions. **B. Next-word prediction as a predictor of model performance:** Relationship between the models' next-word prediction abilities and encoding performance, both in the WITHIN (left) and ACROSS (right) conditions. Next-word prediction (*x*-axis) is operationalized as perplexity on held-out data subsets, averaged across languages. Lower perplexity values indicate better next-word prediction accuracy. Encoding performance (*y*-axis) is measured as the best-layer encoding performance for each model considered. The error bars indicate the language-wise standard error of the mean on the two axes. The gray dashed line indicates the best linear fit. Note that the *y*-axes use different scales. **C. Layer-wise results:** Encoding performance by layer depth. The subplot on the top right reports the layer-wise encoding scores averaged across models; the shaded areas around the lines indicate the standard error of encoding performance across models. The layer positions (*x*-axis) are normalized (i.e., divided by the total number of layers in the models). The remaining subplots indicate the encoding performance by layer for all the models. For each model family (NLLB, XLM-Align, etc.) the first plot reports the layer-wise results obtained in the WITHIN condition, whereas the second plot depicts the scores obtained in the ACROSS condition. Encoding scores are averaged across languages. **D. Cross-lingual alignment as a predictor of model performance:** Relationship between the models' cross-lingual alignment and encoding performance in the WITHIN (left) and ACROSS (right) conditions. Cross-lingual alignment (*x*-axis) is measured as the mean reciprocal rank of the embedding distance of correct translations compared to unrelated sentences in an independent dataset. Higher values indicate stronger cross-lingual alignment. The error bars indicate the language-wise standard error of the mean on the two axes. The gray, dashed line indicates the best linear fit.

Similar to the WITHIN condition, larger models tend to outperform their smaller counterparts within each model class, with the exception of the XGLM model family, where numerically, the highest encoding performance is obtained with $XGLM_{small}$ (r = 0.32, SE = 0.09). Encoder-decoder models in the NLLB and mT5 families exhibit relatively lower performance, confirming their limited capacity in capturing brain responses compared to other model classes (as already shown in the WITHIN condition).

Layer depth within the model remains a significant factor in the ACROSS condition (**Figure 2C**). Intermediate-to-deep layers show higher levels of predictivity, paralleling the trends observed in the WITHIN encoding models and prior work. Encoding scores averaged across models display an inverse U-shaped relation with layer depth, increasing until a normalized depth of ~.70 followed by a decrease (**Figure 2C**, top right).

In a control analysis, we assessed whether brain responses to language could be predicted by a simple baseline of surface linguistic features, namely word length and word frequency, which explain substantial variance in linguistic processing difficulty as measured with reading times (Clifton et al., 2016; Kuperman, Schroeder, & Gnetov, 2024). Encoding models based on this baseline achieved low predictivity, demonstrating that MNNLM encoding models capture neurally relevant information beyond these shallow features (see SI S6 for details; also, de Varda et al., in prep.).

In another control analysis, we evaluated the spatial specificity of the encoding performance by testing the ability of the models to predict fMRI activity in two control sets of brain areas: the right-hemisphere language areas (homotopic to the core, left-hemisphere areas) and areas of the Multiple Demand network (Duncan et al., 2020), which plays a limited role in linguistic processing (Fedorenko & Shain, 2021). Encoding performance in the left-hemisphere language areas was significantly better than in either of these control sets of areas, especially in the ACROSS condition (SI S7 and S8). These findings establish that the information is not ubiquitously present across the brain and mirror past work on English (e.g., Tuckute et al., 2024a).

Next-word prediction ability explains inter-model differences in within-language encoding performance, and cross-lingual alignment of representation spaces supports transfer to new languages

Previous research in computational language neuroscience has shown that language models with better next-word prediction abilities provide a better basis for brain encoding models (Schrimpf et al., 2021; Caucheteux & King, 2022; Hosseini et al., 2024a). To test whether this observation extends to diverse languages, we evaluated the next-word prediction abilities of our 20 models by training a ridge classifier on top of the MNNLM embeddings to predict the next word in a multilingual corpus from Wikipedia. Next-word prediction scores for each model were quantified as the classifiers' cross-validated perplexity, averaged across languages (lower perplexity corresponds to better next word prediction ability). Extending prior findings from English (Schrimpf et al., 2021; Hosseini et al., 2024a) and Dutch (Caucheteux & King, 2022) to a wider array of languages, next-word prediction ability showed a significant relationship with encoding performance in the WITHIN condition (r = -0.56, p = 0.01; **Figure 2B**), indicating that more accurate language models provide a stronger basis for neural encoding. This pattern, however, did not hold for the ACROSS

condition (r = 0.23, n.s.; the difference between the two correlations was significant: z = 2.52, p = 0.01). The absence of a significant relationship between next-word prediction performance and cross-lingual encoding can be explained if we take into account the differences in the tasks that the encoding models need to perform in the two conditions. In the WITHIN condition, where the encoding models are trained and tested on the same language, the main determinant of encoding model performance is the quality of the underlying language representations. In contrast, in the ACROSS condition, the success of an encoding model hinges on the cross-lingual transferability of the representations. Even if a model produces extremely high-quality embeddings in all of its supported languages, the transfer of the encoding models in the ACROSS condition will fail unless the language representations in the various languages are projected onto the same space.

To assess whether the cross-lingual alignment of language representations is responsible for the successful cross-lingual transfer of the brain encoding models, we tested the relationship between the degree to which representations are aligned across languages within a language model and the ability of an encoding model to predict brain responses to a new language in the ACROSS condition. Here, we started with a set of semantically diverse English sentences (Pereira et al., 2018, Experiment 2 and 3 materials). We translated these sentences to the 12 languages in our sample using the Google Translator API, and obtained embeddings for these sentences in each language from each of our MNNLMs. Finally, we calculated to what extent the models projected translation-equivalent sentences onto close regions of the embedding space. For each model, for each language pair, we constructed a distance matrix between the embeddings of all the sentences, and calculated the distance-based mean reciprocal rank of the correct translation. Intuitively, this approach reflects the idea that semantically equivalent sentences should be closer in the embedding space than the other sentences, and the closer they are, the more aligned the two languages' representations are for a given model. In the end, for each model, we averaged the ranks across all language pairs to derive a single value per model.

Our results support the idea that higher representational inter-language alignment is associated with better cross-lingual transfer of the encoding models. **Figure 2D** shows a significant, positive correlation between the mean reciprocal rank of correct translations and brain predictivity in the ACROSS condition (r = 0.56, p = 0.01; see SI S15 for a replication with the data collected for Study II). As expected, this relation was specific to the ACROSS condition, and did not hold for the WITHIN condition (r = 0.25, n.s.), where representations need not be aligned for successful encoding, although the difference between the two correlations was not significant (z = 1.11, n.s.).

Study II

In our first study, we found that the relationship between language representations in artificial and biological systems is sufficiently general to be successfully transferable across languages. In our second study, we devised an even stronger test for the cross-lingual transferability of the encoding models. We fitted the encoding models employing fMRI data from Study I (passage listening in 12 languages) and from three additional fMRI datasets, all using English stimuli presented via auditory or visual modalities. Then, we collected new fMRI data on 9 additional languages (non-overlapping with those in Study I), using the same

paradigm of participants listening to passages from *Alice in Wonderland*, and then directly applied the trained models zero-shot to these newly collected brain responses. The three additional training datasets included two datasets where participants read English sentences (*Pereira2018*, Pereira et al., 2018, Experiment 2 and 3; *Tuckute2024*, Tuckute et al., 2024a) and an fMRI dataset where participants listened to naturalistic stories in English (*NatStories*, expanded from Blank et al., 2014; parts of this datasets were used in Shain, Blank et al., 2020, Shain et al., 2022, and Paunov et al., 2022; see Methods, "NatStories" for details). The datasets thus varied in the languages covered (one vs. multiple languages), type of stimuli employed (sentences vs. passages/stories), and presentation modality (listening vs. reading). **Figure 1C** summarizes the information about the datasets, including details on the participants, stimuli, response reliability, and similarity of the language materials to the test set. In contrast to Study I, where all participants listened to passages from the same book, in Study II, the source materials on which we trained the encoding models involve diverse language stimuli with little word overlap with the test materials and lower similarity in the embedding representations.

We collected new fMRI data from native speakers of 9 languages (Italian, Portuguese, German, Russian, Polish, Hindi, Arabic, Mandarin, Korean) belonging to 4 language families (Indo-European, Afroasiatic, Sino-Tibetan, Koreanic). For each language, three native speakers listened to three passages (~4.5 min each). As in Study I, the participants' language areas were identified with an independent language localizer task (Fedorenko et al., 2010; Malik-Moraleda, Ayyash et al., 2022), and we extracted the responses during the critical passage-listening task from these functionally defined areas and averaged across them, and then averaged the time-series across participants (two or three, depending on the reliability of the time-series; see Methods, "Study II – Time-series extraction"). Similarly, brain responses were averaged across voxels, fROIs, and participants in the training datasets, resulting in a single response per sentence (*Pereira2018, Tuckute2024*) and a single time-series per passage/story (Study I data, *NatStories*). Our general approach was identical to Study I with respect to pre-processing, extraction of time-series, temporal alignment of MNNLM embeddings and neural data, and model fitting.

Encoding models can be transferred to new languages across types of language materials and presentation modalities

The results of Study II show that encoding models trained on one language or a set of languages can be successfully applied to new languages even across variation in type and content of stimuli and presentation modalities (**Figure 3**). The strongest transfer performance is obtained by encoding models trained to predict brain responses in participants listening to passages or stories (Study I and *NatStories*; **Figure 3A**). Among the 20 encoding models based on the data from Study I, 15 significantly outperform the randomized baseline, reaching an encoding performance as high as $r = 0.16$ ($SE = 0.06$) in the case of XGLMxl. The encoding models trained on the *NatStories* data achieve an even stronger generalization performance, as all of them yield encoding scores significantly above chance, with correlations of up to $r = 0.21$ ($SE = 0.10$) with the actual brain responses for the mGPT model. We attribute the strong performance of the encoding models trained on *NatStories* to the high reliability of the fMRI responses in this dataset, which is likely due to a

relatively large number of participants per story (**Figure 1C**). Our findings also reveal that encoding models trained on data from sentence-reading tasks (*Tuckute2024, Pereira2018*) can predict brain responses to passage-listening tasks (**Figure 3B**). This cross-modality/cross-stimulus type of cross-lingual transfer highlights the robustness of the encoding models and suggests a deep alignment between the responses evoked by reading and listening in the language network, in light with a large body of prior work (e.g., Deniz et al., 2019; Fedorenko et al., 2010; Regev et al., 2013; Vagharchakian et al., 2012). In particular, models trained on the *Tuckute2024* dataset yield strong cross-modality and cross-lingual performance when mapped onto listening-based fMRI data, with 16 out of the 20 tested models exhibiting statistically significant predictive performance. The highest encoding score of the models trained on this dataset is achieved by $XLMR_{large}$, with a correlation of r = 0.13 (SE = 0.04) between predicted and recorded brain responses[1]. Models trained on the *Pereira2018* dataset exhibit a more modest cross-modality transfer performance, with only 5 of the 20 models predicting brain responses above chance, and the best-performing model, $XGLM_{xl}$, achieving an encoding score of r = 0.11 (SE = 0.07). We speculate that the lower performance obtained in the transfer from this dataset might be due to relatively low variation in syntactic structures (the dataset was constructed to be semantically variable, but most sentences use simple structures; see SI S13 for an analysis of the variability in syntactic structures across datasets).

The zero-shot generalization ability of MNNLM-based encoding models—shown in Study I and replicated and extended in Study II—indicates that MNNLMs capture properties of language processing that remain consistent across a range of linguistic and experimental contexts. This transferability highlights the potential of brain encoding models to reveal general principles of brain function related to language.

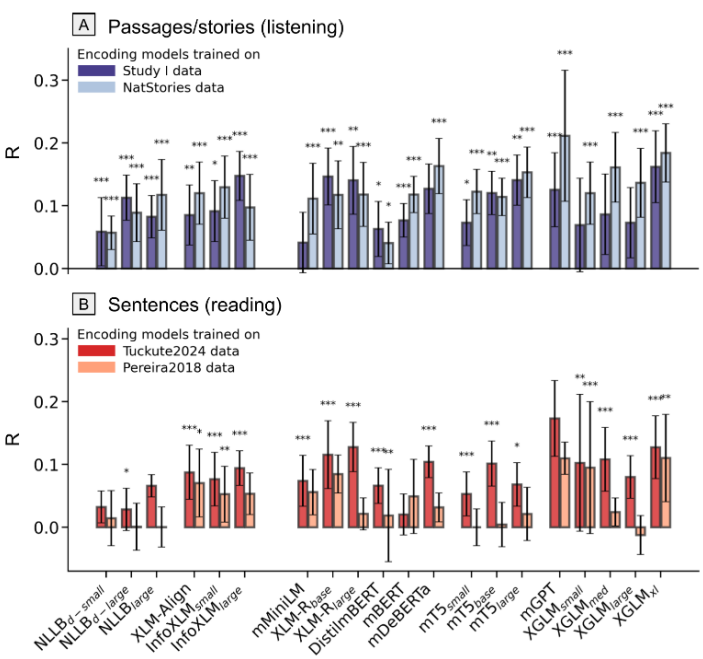


Figure 3. Summary of the results of Study II. The bars indicate how well the models trained on the independent data sources (Study I, *NatStories*, *Pereira2018*, *Tuckute2024*) transferred to the new data we collected. The encoding scores are averaged across passages and languages (see SI S3 for the results in the individual languages); the error bars indicate the standard error of the encoding performance across languages. The models are grouped according to the taxonomy defined for Study I. The asterisks indicate statistical significance (p < .05*; p < 0.01**; p < 0.001***). **A:** Transfer performance of the encoding models trained to predict brain responses to passages/stories presented auditorily (employing the data from Study I and *NatStories*). **B:** Transfer performance of the encoding models trained to predict brain responses to sentences presented visually (employing the data from the *Tuckute2024* and *Pereira2018* datasets).

[1] A numerically higher performance in the transfer from *Tuckute2024* was obtained by mGPT (r = 0.17, SE = 0.06), but the result was not significant (p = 0.06) due to the high performance of the randomized model in this transfer condition (r = 0.13).

Discussion

At the core of the language sciences lies a fundamental tension between the extensive variety of human languages and the uniformity of the biological tissue that supports them. Indeed, the human language network—an interconnected set of left-lateralized frontal and temporal brain regions—supports the processing of all human languages tested so far, irrespective of their profound surface differences that span most levels of linguistic analysis (Chee et al., 1999; Malik-Moraleda, Ayyash, et al., 2022), and even including constructed languages, such as Klingon (Malik-Moraleda et al., in press). This common neural substrate for language processing presumably emerged through biological and cultural evolution in response to similar functional demands and constraints. But beyond broadly cross-lingually similar anatomical and functional properties of the language network, are linguistic representations similar across languages? Recent technological advances in the form of multilingual neural network language models (MNNLMs) have made it possible to derive quantitatively specified linguistic representations in a language-general format. Building on these advances, we first replicated the findings from English and a handful of other high-resource languages (e.g., Schrimpf et al., 2021; Caucheteux & King, 2022; Millet et al., 2023; Nakagi et al., 2024; Sun et al., 2023; Yamashita et al., 2023) showing that encoding models based on these MNNLMs capture neural responses to language in 21 diverse languages (12 in Study I and 9 in Study II; see SI S4). Next and critically, we tested whether the link between language representations and brain responses (i.e., our encoding models) can be transferred across languages. Our results show that indeed the features encoded by MNNLMs can capture brain responses across languages, indicating the existence of a substantial component of the brain response to language that is governed by language-general principles.

What kind of linguistic information supports the cross-lingual transfer of the encoding models? Contextualized word embeddings are high-dimensional language representations that encode both surface (form-based) features of words and sentences, and meaning features, both lexical and compositional (Jawahar et al., 2019). Our results demonstrate that not all the information encoded in the MNNLMs' embeddings can be successfully transferred across languages: encoding models perform better in the WITHIN condition than in the ACROSS condition (Study I), despite being trained on substantially less data. However, identifying the exact subset of features that support cross-lingual transfer is not trivial and would likely require carefully controlled comparisons across many language pairs. Here, we consider two broad possibilities—whether the transfer effects are primarily driven by linguistic form versus by linguistic meaning.

Linguistic form is a compelling candidate for mediating cross-lingual transfer because human languages share a number of structural regularities driven by general cognitive and communicative pressures (see Fedorenko et al., 2024b; Gibson et al., 2019; Kemp, Xu, & Regier, 2018; Levshina & Moran, 2021 for reviews). For example, in all languages, frequent and predictable words tend to be shorter (Zipf, 1949; Piantadosi et al., 2011) and syntactic dependencies tend to be local (Futrell, Mahowald, & Gibson, 2015; Gildea & Temperley, 2010). Critically, brain activity in the language network is strongly sensitive to these lexical and syntactic features (e.g., Heilbron et al., 2022; Lopopolo et al., 2021; Sánchez, Carreiras, & Paz-Alonso, 2023; Shain, Blank et al., 2020; Shain et al., 2022; Tuckute et al., 2024a). By encoding these

statistical regularities, MNNLMs might therefore capture form-based features that modulate brain responses in a similar way across languages. In one of our control analyses, we ruled out the possibility that the simplest form-based features (word length and frequency; SI S6) explain cross-lingual transfer. Moreover, we performed a series of analyses aimed at testing the effects of form-based inter-language similarity (including phonetic, phonological, and syntactic similarity) on transfer. The results (reported in SI S14) show no evidence of form-based similarity at any level on transfer. Although these null results should be interpreted with caution due to potential measurement noise and limited statistical power in comparing encoding performance between languages (see also SI S16), they provide tentative evidence that form-level features may not be the primary determinant of cross-lingual encoding transfer. However, the role of form-based features should be investigated more thoroughly in future work.

Alternatively, cross-lingual transfer could be supported by higher-level information related to compositional meanings: after all, expressing complex meanings is the core objective of language use, and all human languages are capable of expressing diverse meanings about the outer and inner worlds. Although languages greatly vary in their surface features, cross-lingual variation in semantics, while attested (Everett, 2005; Frank et al., 2008; Majid et al., 2018; Berlin & Kay, 1969; Gibson et al., 2017), is comparatively limited relative to the massive overlap across languages in the concepts they can express. Two aspects of our results point to semantic information playing a critical role in the transfer. First, in the ACROSS condition, intermediate-to-deep model layers yield the best encoding performance. In contrast to early model layers, which encode low-level linguistic features such as morphological structure and parts of speech, deeper layers represent higher-level information such as compositional semantics (Belinkov et al., 2020; Niu et al., 2022; Tenney et al., 2019). And second, models that projected semantically equivalent sentences onto closer regions of the embedding space better predicted brain activity in the ACROSS condition. This result suggests that the models' ability to predict brain responses is tightly linked to how they encode semantic content across languages. The interpretation that the successful cross-lingual deployment of encoding models depends on semantic content further aligns with recent findings in computational language neuroscience. In particular, sentence meanings, as opposed to their syntactic structure, have been shown to be the main contributor to the alignment between representations from language models and fMRI recordings (Caucheteux et al., 2021; Gauthier & Levy, 2019; Kauf, Tuckute et al., 2023).

If the successful transfer of the encoding models relies on somewhat form-independent semantic representations in the brain, a critical question concerns the degree of abstractness of such representations. In other words, are the shared responses that we documented specific to language? Or would they be similarly well captured by semantic models grounded in other modalities, such as vision (Kriegeskorte, 2015)? Because language is often used to talk about the physical world, relations between words tend to reflect relations between the words' referents (Günther et al., 2020; Johns & Jones, 2012; Louwerse, 2011). Indeed, recent work has shown that semantic alignment emerges naturally across modalities, as both language and visual representations encode statistical regularities in the underlying reality (Roads & Love, 2020; Huh et al, 2024; cf. Hosseini et al., 2024b). Furthermore, representations from neural network language models can successfully predict brain responses in visual cortical areas (Doerig et al., 2022), even in macaques (Conwell

et al., 2024), and brain encoding models based on multimodal transformers (Xu et al., 2023) can be transferred across language and vision in putative semantic brain areas (Tang et al., 2024). It is important to note, however, that semantic information encoded in the human language network is not fully abstract: semantic information conveyed by meaningful non-linguistic stimuli elicits a much lower response than linguistically encoded meanings (Ivanova et al., 2021; Sueoka, Paunov et al., 2024). As a result, more work is needed to elucidate the extent to which shared responses across languages are preferentially linked to linguistically encoded meanings, as opposed to being fully abstract and shared with semantic information from other modalities.

Beyond showing cross-lingual transfer effects, our results corroborate previous observations from the computational neuroscience literature using a wider array of languages. Against the backdrop of English-centricity in the current literature, we show that language models can successfully account for brain responses even in low-resource languages, such as Marathi and Lithuanian. These findings resonate with the results of Hosseini et al. (2024a), who showed that language models align with brain responses even if trained on developmentally plausible amounts of data. More generally, they reinforce the idea that brain responses to language are modulated by its distributional statistics, which, in turn, are well captured by computational language models. Furthermore, not only can language models account for brain responses in diverse languages, but the model properties that have been shown to affect predictive accuracy for English extend to the languages in our sample. In particular, larger models outperform smaller models in predicting brain activity, underscoring the representational richness of human linguistic representations; and better next-word prediction performance of MNNLMs is associated with stronger ability to predict neural responses, corroborating the idea that language comprehension involves continuously generating and updating predictions about upcoming linguistic input (Heilbron et al., 2022; Levy, 2008; Shain et al., 2020; but see Antonello & Huth, 2024 for an alternative account).

One limitation of the current study is that in spite of being able to identify cross-lingual similarities in brain responses, its design is not equipped to evaluate fine-grained differences in brain encoding across languages (see SI S16). First, the MNNLMs used in this study were not developed to conduct rigorous cross-lingual comparisons, as their pre-training data distribution was highly unbalanced across languages. MNNLMs typically mitigate resource imbalances through strategies such as upsampling data for low-resource languages (Conneau et al., 2020; Devlin et al., 2019; Xue et al., 2021), but high-resource languages remain over-represented in the pre-training corpora, leading to inherent differences in representation quality across languages. Although there have been recent efforts to develop comparable monolingual models in hundreds of languages (Chang et al., 2024), extending a similar standardization to multilingual models has not yet been accomplished, yet is a necessary step for robust comparisons. Additionally, future studies evaluating inter-language differences will need to adopt a “deep sampling” approach to fMRI data collection, testing large groups of speakers for each language and carefully matching them on relevant factors. This is crucial because cross-lingual differences are likely to be subtle and must exceed the considerable variability among the speakers of the same language to be reliably detected (Malik-Moraleda, Ayyash, et al., 2022).

What could these cross-lingual differences in encoding look like? There are several factors that, in principle, could affect the fit between language representations in humans and language models. For example, languages with templatic (non-concatenative) morphology, such as Hebrew, may make it challenging for standard sub-word tokenization approaches to capture their full morphological complexity (de Varda & Marelli, 2023; Klein and Tsarfaty, 2020). In turn, this mismatch between sub-word segmentation in models and speakers' representations could affect the predictivity of the brain encoding models. Relatedly, languages where the written form underspecifies the content expressed in spoken form could pose additional challenges to text-based encoding models. Hausa distinguishes grammatical classes through tone but lacks tone marking in its orthography (Adebara & Abdul-Mageed, 2022); thus, the language models' input will inevitably lack some informative features which may contribute to brain alignment.

The English-centric approach to language hinders both cognitive science (Blasi et al., 2022) and NLP research (Joshi et al., 2020). Research at the intersection of these fields has struggled to widen its focus beyond English and provide a globally representative picture of the computational principles underlying language processing in humans. As a result, the generalizability of some of the claims in the computational cognitive science of language has been questioned, suggesting that the objective functions that neural network language models are trained to optimize might not be a suitable basis for language-general *in silico* models of human language processing (Blasi et al., 2022). Although these concerns are certainly warranted, our findings bring optimism to the field, demonstrating that many claims previously reported for English do, in fact, generalize across languages. This result suggests that the principles underpinning language processing in neural network language models are broadly applicable to the human language system, supporting the idea that these computational models capture fundamental aspects of human language processing.

Methods

Study I

fMRI experiments

Our first study utilized pre-existing functional magnetic resonance imaging (fMRI) data from Malik-Moraleda, Ayyash, et al. (2022) focusing specifically on brain responses elicited by the "passage listening" paradigm. In this setup, participants were exposed to passages approximately 4.5 minutes in duration, presented in their native language.

Participants

The original dataset included fMRI recordings from 86 right-handed participants, with a balanced gender distribution (50% females), ranging in age from 19 to 43 years (mean = 27.52, SD = 5.49). The dataset encompassed 45 languages across 12 language families. For 41 of these languages, there were two participants each, whereas the remaining four languages were represented by just one participant each. The number of participants considered in our study was subsequently reduced based on the reliability assessment

of the fMRI responses (see Methods, “Study II – Time-series extraction”). The languages with only one participant were not considered in our study due to the inability to establish response reliability.

Critical fMRI task

In Malik-Moraleda, Ayyash, et al.’s (2022) study, participants engaged in a variety of tasks while undergoing fMRI scanning. These tasks included a standard language localizer in English (all participants were proficient English speakers), a critical localizer in the participants' native languages, two non-linguistic tasks, a resting-state scan, and a naturalistic passage listening task. Our analyses focused on the passage listening task, where participants listened to ~4.5 min long passages from the book *Alice in Wonderland* (Carrol, 1865), recorded by a female native speaker of each represented language. Because the passage recordings varied slightly in length across languages, they were all trimmed to 260 s, and the final two seconds of the recording included a gradual volume fade-out. In addition, each run included 12 seconds of silence at the beginning and end, for a total duration of 284 s (4 min 44 s).

fMRI data acquisition

Structural and functional data were collected on the whole-body 3 Tesla Siemens Trio scanner with a 32-channel head coil at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. T1-weighted structural images were collected in 179 sagittal slices with 1 mm isotropic voxels (TR = 2,530 ms, TE = 3.48 ms). Functional, blood oxygenation level dependent (BOLD) data were acquired using an EPI sequence (with a 90° flip angle and using GRAPPA with an acceleration factor of 2), with the following acquisition parameters: thirty-one 4mm thick near-axial slices, acquired in an interleaved order with a 10% distance factor; 2.1 mm × 2.1 mm in-plane resolution; field of view of 200mm in the phase encoding anterior to posterior (A >> P) direction; matrix size of 96 × 96; TR of 2000 ms; and TE of 30 ms. Prospective acquisition correction was used to adjust the positions of the gradients based on the participant’s motion one TR back. The first 10 s of each run were excluded to allow for steady-state magnetization.

fMRI data preprocessing

Data preprocessing was performed with SPM12 (using default parameters, unless specified otherwise) and custom scripts in MATLAB. Preprocessing of anatomical data included normalization into a common space (Montreal Neurological Institute [MNI] template) and segmentation into probabilistic maps of the gray matter (GM), white matter (WM), and cerebro-spinal fluid (CSF). A GM mask was generated from the GM probability map and resampled to 2 mm isotropic voxels to mask the functional data. The WM and CSF maps were used as described below. Preprocessing of functional data included motion correction (realignment to the mean image using second-degree b-spline interpolation), normalization (estimated for the mean image using trilinear interpolation), resampling into 2 mm isotropic voxels, and smoothing with a 4 mm full-width, half-maximum (FWHM) Gaussian filter.

fMRI data modeling and definition of functional regions of interest (fROIs)

The language localizer data (here and in Study II) and the data from the additional training datasets in Study II were all modeled using the same setup. Effects were estimated using a General Linear Model (GLM) in which each experimental condition was modeled with a boxcar function convolved with the canonical hemodynamic response function (HRF) (fixation was modeled implicitly, such that all timepoints that did not correspond to one of the conditions were assumed to correspond to a fixation period). Temporal autocorrelations in the BOLD signal timeseries were accounted for by a combination of high-pass filtering with a 128 seconds cutoff, and whitening using an AR(0.2) model (first-order autoregressive model linearized around the coefficient a=0.2) to approximate the observed covariance of the functional data in the context of Restricted Maximum Likelihood estimation (ReML). In addition to experimental condition effects, the GLM design included first-order temporal derivatives for each condition (included to model variability in the HRF delays), as well as nuisance regressors to control for the effect of slow linear drifts, subject-motion parameters, and potential outlier scans on the BOLD signal.

Language fROIs were defined in individual participants (Saxe et al., 2006; Fedorenko et al., 2010) by combining two sources of information: (1) the participant's activation map for the language localizer contrast (Fedorenko et al., 2010) and (2) group-level constraints ('parcels') that delineated the expected gross locations of activations and were sufficiently large to encompass the variability in the locations of individual activations (all parcels are available for download from https://www.evlab.mit.edu/resources-all/download-parcels). The language localizer, described in detail in Fedorenko et al. (2010), consisted of reading English sentences and lists of unconnected nonwords in a standard blocked design, with stimuli presented one (non)word at a time. We used the English localizer to ensure comparability with previous studies and because it has been established that the localizer for a language works well as long as participants are fluent in that language (Chen et al., 2021; Malik-Moraleda, Ayyash et al., 2022). The parcels were derived from a probabilistic activation overlap map using watershed parcellation, as described by Fedorenko et al. (2010), for the *sentences* > *non-words* contrast in 220 independent participants and covered extensive portions of the lateral frontal, temporal, and parietal cortices. Five language fROIs were defined (as the 10% of voxels with the highest t-values for the *sentences* > *nonwords* contrast) in the dominant hemisphere (left for all participants but one Marathi speaker): three on the lateral surface of the frontal cortex and two on the lateral surface of the temporal and parietal cortex. (We also defined five homotopic language fROIs and 20 fROIs in the Multiple Demand network (Duncan et al., 2020), which were both used in a control analysis.)

Time-series extraction

The passages were 260 s (4 min 20 s) long, surrounded by 12-s periods of silence. fMRI responses were recorded at the temporal resolution of 2 s, resulting in 142 data points (volumes or TRs) (130 TRs after the exclusion of the silence periods). For each participant, BOLD time-series were extracted from each voxel within the language network; then, they were averaged across the voxels in each of the five language fROIs, and then across fROIs; finally, they were averaged across participants. To account for the lag in the hemodynamic response, we shifted the time-series by 6 seconds (this choice was corroborated by encoding

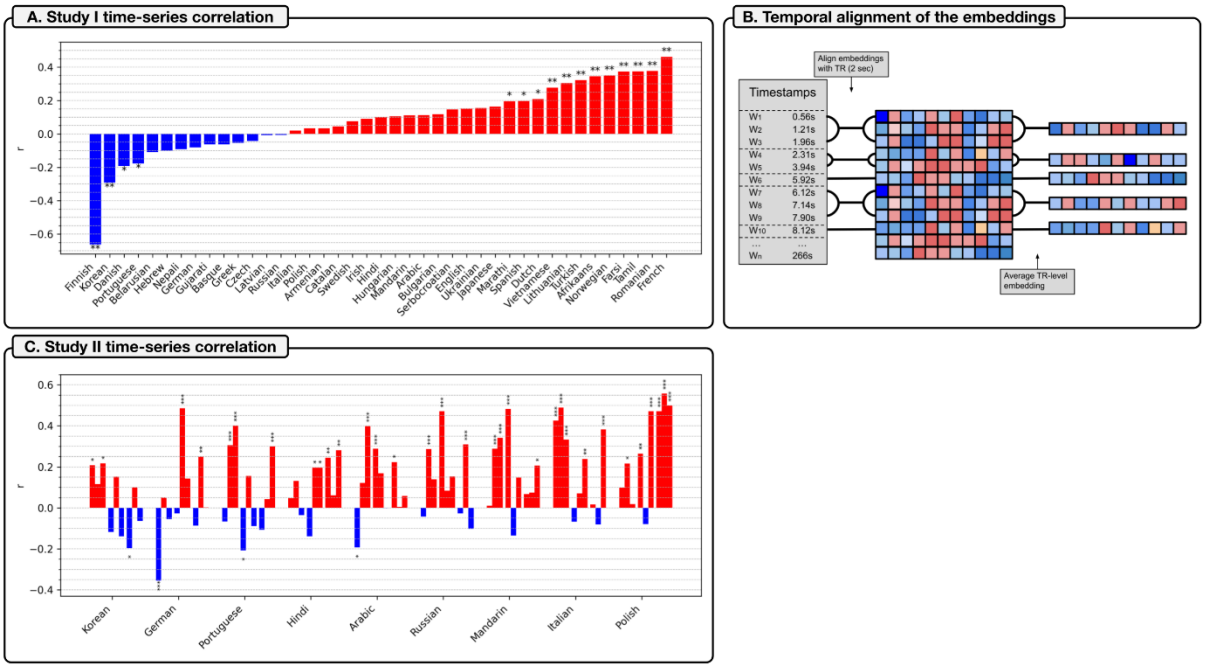


Figure 4. fMRI time-series data reliability and temporal alignment of the embeddings with the time-series. **A. Study I time-series correlation:** Correlation between the time-series of the two participants for each language (Study I data, from Malik-Moraleda, Ayyash, et al., 2022). The asterisks indicate statistical significance ($p < .05$*; $p < 0.01$**). **B. Temporal alignment of the embeddings:** Visual summary of the alignment of contextualized word embeddings and brain responses. Whisper-timestamped produces word-by-word transcriptions of the audio files and the relative timestamps. We used the timestamps to downsample the embedding matrix produced by the MNNLM to the temporal resolution of the fMRI time-series, averaging over the word representations corresponding to each TR interval. **C. Study II time-series correlation:** Correlation between the time-series of the three participants for each language × passage combination (Study II data). For each language, we have 9 bars: each triplet of bars corresponds to a passage (bars 1-3 correspond to passage 1, bars 4-6 – to passage 2, and bars 7-9 – to passage 3), and the individual bars in the triplet show the correlations for all possible participant pairs (participants 1 and 2, 1 and 3, and 2 and 3). The asterisks indicate statistical significance ($p < .05$*; $p < 0.01$**; $p < 0.001$***).

analyses in a separate dataset; SI S9). As noted above, we focused on the subset of languages where the time-series of the two participants showed a positive and statistically significant Pearson correlation ($p < .05$; see SI S10 for the results of a sanity-check analysis showing that the correlations are higher within a language than between languages). This selection criterion left 12 languages: Afrikaans, Dutch, Farsi, French, Lithuanian, Marathi, Norwegian, Romanian, Spanish, Tamil, Turkish, Vietnamese (see **Figure 4A**), excluding 29 languages.

Timestamped speech recognition

Participants were exposed to continuous auditory stimuli. To convert these audio stimuli into textual data suitable for language model analysis, we employed an automated transcription process using the Whisper-timestamped tool (Louradour, 2023). This tool is an extension of the Whisper model (Radford et al., 2023) and it outputs precise word-by-word timestamps. We utilized the largest available model variant (1,550M parameters); this model supported transcription in all the selected languages.

Language models

Following the acquisition of the textual input associated with each passage, we generated contextualized word embeddings using a variety of MNNLMs. The specific models employed are detailed in Table 1 and feature a range of architectures, including encoder-only, decoder-only, and encoder-decoder models, with parameter counts ranging from 118 million to 4.5 billion. Some of the models were explicitly optimized to align representations across languages; some examples include models trained on translation (NLLB), word alignment in translation pairs (XLM-Align), and contrastive cross-lingual training (InfoXLM). The majority of the models, however, were not trained with an explicit cross-lingual objective; instead, their pre-training tasks involved the reconstruction of the input text that had been altered or corrupted in some way. The models trained on masked language modeling (mBERT, XLM-R) were optimized to predict randomly masked tokens within the input by relying on bidirectional context, and the models based on span corruption (mT5) were similarly trained to reconstruct consecutive spans of text; mDeBERTa was trained to detect tokens that had been replaced in the input. Causal language models (mGPT, XGLM) were trained to predict each token sequentially based on the preceding context only. We also considered two models (DistilmBERT, mMiniLM) trained by distilling the language knowledge from larger, more complex masked language models (mBERT, XLM-R) into smaller, more efficient ones. All models were utilized as-is, without any fine-tuning or adaptation, and were accessed via the transformers Python package (version 4.45.2; Wolf et al., 2020).

Model	Architecture	N° lang.	Layers	Parameters	Pre-training task	Citation
NLLB$_{d\text{-}small}$	Encoder-decoder	200	12	615M	Distillation (from machine translation model)	NLLB Team, 2024
NLLB$_{d\text{-}large}$	Encoder-decoder	200	24	1,371M	Distillation (from machine translation model)	NLLB Team, 2024
NLLB$_{large}$	Encoder-decoder	200	24	1,371M	Machine translation	NLLB Team, 2024
XLM-Align	Encoder-only	94	12	278M	Denoising word alignment	Chi et al., 2021a
InfoXLM$_{small}$	Encoder-only	94	12	278M	Cross-lingual contrastive learning	Chi et al., 2021b
InfoXLM$_{large}$	Encoder-only	94	24	560M	Cross-lingual contrastive learning	Chi et al., 2021b
mMiniLM	Encoder-only	100	12	118M	Distillation (from masked language model)	Wang et al., 2020
XLM-R$_{base}$	Encoder-only	100	12	278M	Masked language modeling	Conneau et al., 2020
XLM-Rlarge	Encoder-only	100	24	560M	Masked language modeling	Conneau et al., 2020
DistilmBERT	Encoder-only	104	6	135M	Distillation (from masked language model)	Sanh et al., 2019
mBERT	Encoder-only	104	12	178M	Masked language modeling	Devlin et al., 2019
mDeBERTa	Encoder-only	100	12	278M	Replaced token detection	He et al. 2022

mT5$_{small}$	Encoder-decoder	101	8	172M	Span corruption	Xue et al., 2021
mT5$_{base}$	Encoder-decoder	101	12	390M	Span corruption	Xue et al., 2021
mT5$_{large}$	Encoder-decoder	101	24	973M	Span corruption	Xue et al., 2021
mGPT	Decoder-only	60	24	1,418M	Causal language modeling	Shliazhko et al., 2022
XGLM$_{small}$	Decoder-only	30	24	564M	Causal language modeling	Lin et al., 2021
XGLM$_{med}$	Decoder-only	30	24	1,733M	Causal language modeling	Lin et al., 2021
XGLM$_{large}$	Decoder-only	30	48	2,942M	Causal language modeling	Lin et al., 2021
XGLM$_{xl}$	Decoder-only	30	48	4,552M	Causal language modeling	Lin et al., 2021

Table 1. Summary of the MNNLMs employed in the study.

We extracted the hidden states from each layer of each MNNLM in response to the transcribed text and derived word-level embeddings by averaging the embeddings of sub-word tokens (see "Encoding models" section below for procedure in case words were split between TRs). In the case of encoder-decoder models, we did not consider the representations generated by the decoder. We added model-specific special tokens (e.g., beginning-of-sequence and end-of-sequence tokens) to the input text where appropriate, and then discarded their corresponding embedding representations.

All the models in our sample except the ones based on a decoder-only architecture were bidirectional, meaning that token representations could incorporate information about future tokens. To prevent the models from accessing future information, we iteratively slid a window of 100 words across the texts, ensuring that each token's embedding was influenced only by the preceding context within the window. This method allowed us to approximate a causal setup in bidirectional models, making the processing scheme more comparable with human language processing, where comprehension unfolds incrementally based on prior context without knowledge of future input. In SI S11, we considered an alternative approach of obtaining contextualized embeddings where we reset the context at each cross-validation fold boundary, ensuring that the embeddings could not carry over contextual information from one fold to the next. This analysis showed that preventing context bleed-over had a limited effect on performance.

To get a sense of the similarity in the embedding representations across languages, we calculated the average cosine similarity between the contextualized embeddings corresponding to all word pairs across languages. The resulting similarity values were then averaged across words, languages, layers, and models (**Figure 1C**).

Brain encoding models

We aligned the embeddings generated from the MNNLMs with the brain responses of participants exposed to the same linguistic stimuli. For each of the 130 TRs, TR-level embeddings were computed by averaging the embeddings of the words pronounced within that interval (**Figure 4B**). Words that were pronounced at the TR boundaries (e.g., a word whose pronunciation started in a given TR and ended in the subsequent one)

were included in the embedding calculation of the second TR. We normalized with *z*-scoring both the embedding representations (per unit) and the fMRI time-series (per univariate response). The normalization parameters were estimated in the training set and directly applied to the test set. The mapping of the embeddings to brain responses was accomplished using ridge (i.e., *L2*-penalized) regression. The α parameter, which controls the strength of the regularization, was set with nested leave-one-out cross-validation by considering 10 candidate values ranging from 10^{-5} to 10^{4}, logarithmically spaced. This was achieved with the RidgeCV function of the package scikit-learn (version 1.2.1; Pedregosa et al., 2011).

WITHIN encoding models

As a first step of our study, the encoding models were trained and tested on non-overlapping data subsets within each language. In these encoding models, the predictor variable was each TR-level embedding representation, whereas the dependent variable was the response within the language network, averaged across voxels, fROIs, and participants. We developed separate encoding models for every combination of language, model, and layer. Note that although most models were fitted for all the 12 languages considered in the study (see SI S4 for data from the 9 languages from Study II), the encoding models based on the XGLM embeddings were fitted in 5 languages (Spanish, Vietnamese, Tamil, Turkish, French) and the encoding models based on mGPT were fitted in 10 languages (Afrikaans, Farsi, French, Lithuanian, Marathi, Romanian, Spanish, Tamil, Turkish, Vietnamese), as the remaining languages were absent from the models' pre-training data. The encoding performance was assessed via 10-fold cross-validation with sequential splits[2]. We adopted a sequential train-test split since with shuffled train-test splits encoding models can achieve high predictivity by simply relying on the temporal autocorrelation of the signal (Feghhi et al., 2024; Kauf, Tuckute et al., 2023). The predicted responses for each left-out fold were then concatenated to reconstruct a full time-series, which was then compared to the recorded brain responses via Pearson correlation. A graphical summary of the encoding procedure is provided in **Figure 1B**, top.

ACROSS encoding models

In a subsequent step, we tested whether the encoding models could be transferred to new participants, materials, and languages. To this end, for each individual language $l_i \in \mathrm{L}$, we fitted an encoding model in all the remaining languages $l_j \in \mathrm{L} \setminus \{l_i\}$. Then, the obtained regression weights were employed to predict the fMRI responses in l_i, which was not included in the training data. The encoding performance was evaluated as the Pearson correlation between the predicted and the actual brain response in l_i. A schematic visualization of the cross-lingual encoding models is displayed in **Figure 1B**, bottom.

[2] In SI S11, we report the encoding performance obtained in a stricter condition where we reset the models' context at each fold boundary to control for possible context contamination. Our results showed that context contamination did not substantially impact the results.

Significance testing

To evaluate the statistical significance of the encoding performance scores, we compared the results of the encoding models with the scores obtained by equivalent encoding models where the fMRI responses were randomly shuffled (i.e., destroying the correspondence between stimulus and brain response). Random seeds were fixed for reproducibility purposes. Statistical significance was evaluated via a two-step analytical approach measuring whether the correlations obtained by the encoding model were systematically higher than the ones obtained after shuffling the response variable. First, we calculated the difference between corresponding pairs of Pearson correlation coefficients from the original and the shuffled encoding models, where each pair of coefficients corresponds to the scores obtained by a given encoding model in a specific language. Given the non-normal distribution of Pearson correlation scores, we applied Fisher's *r*-to-*z* transformation to convert these coefficients into *z*-scores (Fisher, 1921; see Eq. 1). The difference between the obtained *z*-scores was divided by the standard error of the difference, i.e., $\sqrt{\frac{2}{n-3}}$, where n = 130 is the sample size used to calculate the correlation coefficient. This procedure allowed us to generate a *z*-statistic for each correlation pair, reflecting the magnitude and direction of the difference between the correlation coefficients obtained by the original and the randomized encoding models.

$$(1)\quad z = \frac{1}{2} log\left(\frac{1+r}{1-r}\right)$$

To assess the overall significance across all pairs of correlations obtained in the various languages by a given MNNLM, we combined the individual *z*-statistics derived in the previous step using Stouffer's method (Stouffer et al., 1949), which is more conservative than Fisher's combined probability method in that it does not assume independence between observations. This method involves summing the *z*-statistics and dividing by the square root of *k*, the number of statistics combined (Eq. 2).

$$(2)\quad z_{tot} = \frac{\sum_{i=1} z_i}{\sqrt{k}}$$

The combined z_{tot} statistic was then converted into a two-tailed *p*-value, reflecting the probability of observing a combined difference in correlations as extreme as, or more extreme than, the one observed given a null hypothesis. This significance calculation procedure was carried out independently for each encoding model, both in the context of WITHIN and ACROSS encoding models.

Next-word prediction performance

To assess whether next-word prediction performance explained how well the model representations fit to the brain data, we evaluated all models on a word prediction task using a custom multilingual dataset. To create this dataset, we first scraped 50,000 words from random Wikipedia pages in each of the 12 languages considered in the study (6M words in total). We removed titles and headings and considered only text paragraphs. Since meaningful comparisons in next-word prediction ability can be drawn only in the context

of similarly sized vocabularies, we defined a vocabulary in each language as the 5,000 most common words in that language. Then, we divided the continuous text streams into chunks of 10 words each by sliding a 10-words window on the text data.

Next, we extracted word embeddings from all 20 language models considered in this study; for each 10-word chunk, we considered the contextualized embeddings of the last word, as derived from the last layer of the model. Lastly, we trained a ridge multi-class classifier to predict the following word (encoded as a one-hot 5,000-dimensional vector) from the contextualized embedding of the previous one. As for the neural encoding models, the α parameter was set with nested leave-one-out cross-validation. We trained the classifier independently in each language with stratified 10-fold cross-validation, to ensure that all classes were present in both the training and the test set. We evaluated the models' next-word prediction ability as the perplexity on held-out data, averaged over folds. Random seeds for all random processes in the pipeline were set to ensure reproducibility. In preliminary analyses, we found that the perplexity values obtained by the models in each language were highly correlated with the perplexity values obtained by a randomized baseline, possibly due to the fact that some languages are more predictable than others due to different vocabulary sizes and word repetition rates. To isolate the next-word prediction abilities of the MNNLMs, we used the perplexity estimates generated by the randomized models to linearly predict the models' actual perplexity estimates, and used as a next-word prediction metric the residual of the prediction, effectively decorrelating the target estimates with the ones produced by the randomized models. Similar results were obtained using the raw perplexity estimates. The relationship between next-word prediction scores and encoding performance was assessed as the Pearson correlation between each model's residual perplexity values and its best-layer encoding score obtained in the WITHIN and ACROSS conditions.

Alignment of representations across languages

To evaluate the cross-lingual alignment of the MNNLM embeddings, we used a semantically diverse dataset of English sentences (Pereira et al., 2018, Experiments 2 and 3). We translated these sentences from English into the 12 target languages using the Google Translator API. For each MNNLM in our sample, we extracted sentence embeddings in all the languages covered by the MNNLM (10 for mGPT; 5 for XGLM; 12 for all the other models) by averaging over word embeddings. Then, we measured the proximity of translation-equivalent sentences in the embedding space across language pairs. For each pair of languages, we constructed a distance matrix between the embeddings of all sentences, where each entry represented the cosine distance between sentence embeddings. We then calculated the mean reciprocal rank (MRR) of the correct translations within this matrix, where higher MRR values reflect stronger cross-lingual alignment. MRR scores for each model were obtained by averaging over all language pairs (45 for mGPT; 10 for XGLM; 66 for the remaining models). To assess the relationship between cross-lingual alignment and encoding performance, we calculated the Pearson correlations between each model's MRR value and its best-layer encoding score obtained in the WITHIN and ACROSS conditions. In SI S15, we repeated this analysis using the newly collected fMRI data from Study II.

Study II

fMRI experiments

Similar to Study I, our second study was based on fMRI data from participants listening to passages in their native language. In this setup, participants were exposed to three passages approximately 4.5 minutes in duration. All the methodological details (fMRI data acquisition, preprocessing, definition of fROIs) were similar to Study I except that the MRI scanner had undergone a Prisma upgrade (see Methods, "Study II – fMRI data acquisition").

Participants

We recruited 27 participants (three in each of the 9 target languages, i.e., Italian, Portuguese, German, Russian, Polish, Hindi, Arabic, Mandarin, Korean). All but two participants were right-handed; the remaining two were left-handed and ambidextrous; participants ranged in age from 20 to 42 years (mean = 27.37, SD = 5.18), and 18 of them were females (66.66%). None of the participants had taken part in Malik-Moraleda, Ayyash, et al.'s (2022) study, so there was no overlap in participants between Study I and Study II.

Critical fMRI task

The critical fMRI task was equivalent to the one in Study I, except that participants listened to three passages (as opposed to one) in their native language. The passages all came from the book *Alice in Wonderland* (Carrol, 1865) and were recorded by a female native speaker; the same criteria were applied with respect to the passages' durations.

fMRI data acquisition

Structural and functional data were collected on a whole-body 3 Tesla Siemens Prisma scanner with a 32-channel head coil at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. T1-weighted, Magnetization Prepared Rapid Gradient Echo (MP-RAGE) structural images were collected in 208 sagittal slices with 1 mm isotropic voxels (TR = 1,800 ms, TE = 2.37 ms, TI = 900 ms, flip = 8 degrees). Functional, blood oxygenation level-dependent (BOLD) data were acquired using an SMS EPI sequence with a 90° flip angle and using a slice acceleration factor of 3, with the following acquisition parameters: seventy-two 2 mm thick near-axial slices acquired in the interleaved order (with 10% distance factor), 2 mm × 2 mm in-plane resolution, FoV in the phase encoding (A >> P) direction 208 mm and matrix size 104 × 104, TR = 2,000 ms, TE = 30 ms, and partial Fourier of 7/8. The first 10 s of each run were excluded to allow for steady state magnetization.

Time-series extraction

As in Study I, the passages were 260 s long, surrounded by 12-s periods of silence. With a temporal resolution of 2 s, each passage contributed 142 data points (130 after the exclusion of the silence periods). For each participant, BOLD time-series were extracted from each voxel within the language network; then, they were

averaged across the voxels in each of the five language fROIs, and then across fROIs; lastly, they were averaged across participants (two or three, depending on the correlations in the time-series). For each language, correlations between the participants' time-series were assessed across the three passages. **Figure 4C** displays the correlations between participant pairs within each passage for each language. Although some individual language × passage combinations displayed near-zero or negative correlations, the correlations were generally positive and significant across languages. In line with our approach in Study I, we retained only participant pairs with time-series that were significantly correlated at the level of each individual passage. Importantly, this procedure allowed for variation in participant pairs across passages within the same language. For instances where two out of three participant pairs showed significant correlations in their time-series, all three participants were included, since all three time-series correlated with at least one of the others. Using these criteria, we obtained reliable time-series for all three passages in each language, with the exception of five specific language-passage combinations: Korean (passages 2 and 3), German (passage 1), Portuguese (passage 2), and Hindi (passage 1). SI S10 reports the results of a sanity-check analysis showing that the correlations in the time-series are higher within a language than between languages.

Training data

In our second experiment, we fitted several encoding models from different data sources and transferred them zero-shot to the newly collected data. The original data sources encompass different types of stimuli and modalities of presentation. In particular, in addition to the data employed in our first study, we considered two datasets where participants read sentences (*Pereira2018*, Pereira et al., 2018; *Tuckute2024*, Tuckute et al., 2024a), and one additional dataset where participants listened to stories in English (*NatStories*, expanded from Blank et al., 2014). Across all datasets in Study II, we extracted responses from the same five language fROIs (see "fMRI data modeling and definition of functional regions of interest (fROIs)") and averaged them across fROIs, and then across participants, thus following the same procedure as in Study I.

Pereira2018

The *Pereira2018* dataset (Pereira et al., 2018; Exp. 2 and 3) includes fMRI recordings from 10 unique participants (9 in Exp. 2 and 6 in Exp. 3) reading 627 sentences (384 in Exp. 2, 243 in Exp. 3), grouped into passages of three to four sentences. Each participant saw each sentence between 1 and 4 times, with the majority seeing each sentence 3 times. The two sets of materials were constructed independently, and each spanned a broad range of content areas. Sentences were 7–18 words long in Experiment 2, and 5–20 words long in Experiment 3. The sentences were presented visually on a screen one at a time for 4 seconds (followed by 4 seconds of fixation, with additional 4 seconds of fixation at the end of each passage).

Tuckute2024

The *Tuckute2024* dataset (Tuckute et al., 2024a) includes fMRI recordings from 10 unique participants reading 1,000 sentences. The materials were extracted from written and transcribed spoken corpora and were designed to maximize content and stylistic diversity. All sentences were 6 words long. The sentences were

presented visually on a screen one at a time for 2 seconds (followed by 4 seconds of fixation). Each participant read each sentence once, and data were acquired across two or three separate scanning sessions.

fMRI preprocessing (shared between Pereira2018 and Tuckute2024)

The same preprocessing pipeline was used as for the localizer data in Study I and Study II.

First-level Analysis (Pereira2018)

The same modeling setup was used as for the localizer data in Study I and Study II. Each sentence trial was modeled as a separate condition.

First-level Analysis (Tuckute2024)

Effects were estimated using a GLM in which each experimental condition (i.e., each sentence trial) was modeled using GLMsingle (Prince et al., 2022). Using this framework, a general linear model (GLM) was used to estimate the beta weights that represent the BOLD response amplitude evoked by each individual sentence trial (fixation was modeled implicitly, such that all time points that did not correspond to one of the conditions (sentences) were assumed to correspond to a fixation period). For each voxel, the HRF that provided the best fit to the data was identified (on the basis of the amount of variance explained). The data were modeled using a fixed number of noise regressors (five) and a fixed ridge regression fraction (0.05) (these parameters were determined empirically using a joint data modeling and data evaluation framework; see Tuckute et al., 2024a). By default, GLMsingle returns beta weights in units of percent signal change by dividing by the mean signal intensity observed at each voxel and multiplying by 100. To mitigate the effect of collecting data across multiple scanning sessions, the beta values were *z*-scored session-wise per voxel (Methods, "fMRI data modeling and definition of functional regions of interest (fROIs)").

NatStories

The *NatStories* dataset is a collection of fMRI recordings from participants listening to naturalistic stories (∼5 minutes each) while in the fMRI scanner. The dataset was collected in the Fedorenko lab between 2014 and 2020. The first findings were reported in Blank et al. (2014), and the data from the expanded dataset were used in several other studies (Paunov et al., 2022; Shain, Blank, et al., 2020; Shain et al., 2022; Sueoka, Paunov et al., 2024; Wehbe et al., 2021), but the dataset has never been used in its entirety, as we do here. The dataset features 9 stories: 8 were edited to include low-frequency syntactic constructions but maintaining fluency for native speakers (see Futrell et al., 2018 for details of these materials), and 1 was an expository text ("Tree") adapted from Wikipedia (see Paunov et al., 2022 for details). The number of participants per story varied from 7 to 70 (see SI S12). Because this is the first time the dataset is used in its current form, we conducted a reliability analysis which showed that the time-series were highly reliable (see SI S12).

fMRI preprocessing (NatStories)

The same preprocessing pipeline was used as for the localizer data in Study I and Study II. Additional preprocessing of data was performed using the CONN toolbox (Whitfield-Gabrieli & Nieto-Castañón, 2012; https://www.nitrc.org/projects/conn) with default parameters, unless specified otherwise. BOLD signal time courses were extracted from WM and CSF. Five temporal principal components were obtained from each, as well as voxel-wise averages. These were regressed out of each voxel's time course, along with additional noise regressors, specifically, six motion parameters estimated during offline motion correction (three translations and three rotations) and their first temporal derivatives, and artifact time points (based on global signal outliers and motion).

Brain encoding models

As in Study I, our main analyses were based on the training and evaluation of encoding models predicting brain activity from MNNLM-based embeddings. The general procedure that allowed us to derive TR-level embeddings was left unaltered with respect to Study I, and involved obtaining speech transcriptions with Whisper-timestamped, computing the embeddings, and temporally aligning the embeddings with the time-series. Differently from Study I, we here considered a single layer for each MNNLM: the layer that we identified in Study I as the best-performing layer in the ACROSS condition. All the languages we considered in Study II were present in the models' pre-training data.

We fitted several encoding models from different data sources (one for each model × data source combination) and transferred them zero-shot to the new data we collected. In the case of sentence-level datasets (*Pereira2018*, *Tuckute2024*), the embedding representations for the complete sentences were obtained by averaging over all the words in the sentence; thus, each sentence contributed to a single training example for the encoding models. In the case of the naturalistic passages (Study I data, *NatStories*), we aligned the embeddings with the fMRI time-series employing the same procedure as in Study I, averaging the embedding representations over the relevant TRs. All the technical aspects of the encoding models (e.g., penalty estimation, normalization), were left unaltered from Study I. As in the case of Study I, statistical significance was calculated by comparing the encoding performance with a randomized baseline model where the correspondence between embeddings and response variable had been randomly shuffled. Encoding scores were aggregated first across passages, then across languages.

Similarity in the language materials

Unlike Study I, where the language materials consisted of passages from the same book and thus were highly homogeneous, Study II involved the transfer of encoding models across very different kinds of linguistic stimuli. To assess the similarity between the linguistic materials used for training the encoding models (Study I materials, *NatStories*, *Pereira2018*, *Tuckute2024*) and those used for testing them (Study II materials), we considered both the similarity in the word embeddings produced by the MNNLMs in our sample and the word overlap across the stimuli. To calculate embedding-based similarity, for each training dataset, we computed the cosine similarity between each word embedding from the dataset's materials and each word

embedding in the test dataset. Similarity values were then averaged across words, sentences/passages, languages, layers, and models (**Figure 1C**). To measure word overlap, we calculated the number of unique words shared between each training dataset and the test dataset ($W_{train} \cap W_{test}$) divided by the number of unique words included in both datasets ($W_{train} \cup W_{test}$). Their ratio, which corresponds to the Jaccard index, was employed as a metric of lexical overlap.

Data and code availability

Code and data to reproduce this study are available at https://osf.io/uc6y7/.

References

Adebara, I., & Abdul-Mageed, M. (2022). Towards Afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3814–3841. Dublin, Ireland: Association for Computational Linguistics.

Antonello, R., & Huth, A. (2024). Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiology of Language, 5*(1), 64-79.

Antonello, R., Vaidya, A., & Huth, A. (2024). Scaling laws for language encoding models in fMRI. *Advances in Neural Information Processing Systems, 36*.

Assem, M., Glasser, M. F., Van Essen, D. C., & Duncan, J. (2020a). A domain-general cognitive core defined in multimodally parcellated human cortex. *Cerebral Cortex, 30*(8), 4361-4380.

Assem, M., Blank, I. A., Mineroff, Z. A., Ademoğlu, A., & Fedorenko, E. (2020b). Activity in the fronto-parietal multiple-demand network robustly associated with individual differences in working memory and fluid intelligence. *Cortex, 131*, 1–16.

Aw, K. L., & Toneva, M. (2023). Training Language Models to Summarize Narratives Improves Brain Alignment. In *Eleventh International Conference on Learning Representations.*

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., & Glass, J. (2020). On the linguistic representational power of neural machine translation models. *Computational Linguistics, 46*(1), 1-52.

Berlin, B., & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution.* Berkeley & Los Angeles University of California Press.

Blank, I., Kanwisher, N., & Fedorenko, E. (2014). A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of neurophysiology, 112*(5), 1105-1118.

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in cognitive sciences, 26*(12), 1153-1170.

Buchweitz, A., Shinkareva, S. V., Mason, R. A., Mitchell, T. M., & Just, M. A. (2012). Identifying bilingual semantic neural representations across languages. *Brain and language, 120*(3), 282–289.

Campbell, L. (2008). Ethnologue: Languages of the world.

Caucheteux, C., Gramfort, A., & King, J. R. (2021). Disentangling syntax and semantics in the brain with deep networks. In *International conference on machine learning* (pp. 1336-1348). PMLR.

Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. *Communications biology*, *5*(1), 134.

Carroll, L. (1865). *Alice's adventures in wonderland (1st)*. Macmillan.

Chang, T. A., Arnett, C., Tu, Z., & Bergen, B. K. (2024). Goldfish: Monolingual Language Models for 350 Languages. *arXiv preprint arXiv:2408.10441*.

Chee, M. W., Caplan, D., Soon, C. S., Sriram, N., Tan, E. W., Thiel, T., & Weekes, B. (1999). Processing of visually presented sentences in Mandarin and English studied with fMRI. *Neuron*, *23*(1), 127-137.

Chen, C., Gong, X., Tseng, C., Klein, D., Gallant, J., & Deniz, F. (2024). Bilingual language processing relies on shared semantic representations that are modulated by each language. *bioRxiv*, 2024-06.

Chen, X., Affourtit, J., Ryskin, R., Regev, T. I., Norman-Haignere, S., Jouravlev, O., ... & Fedorenko, E. (2021). The human language system does not support music processing (p. 2021.06. 01.446439). *bioRxiv*.

Chi, E. A., Hewitt, J., & Manning, C. D. (2020). Finding Universal Grammatical Relations in Multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Chi, Z., Dong, L., Zheng, B., Huang, S., Mao, X. L., Huang, H. Y., & Wei, F. (2021a). Improving Pretrained Cross-Lingual Language Models via Self-Labeled Word Alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 3418-3430).

Chi, Z., Dong, L., Wei, F., Yang, N., Singhal, S., Wang, W., ... & Zhou, M. (2021b). InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3576-3588).

Clifton, C., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith Rayner's 40-year legacy. *Journal of Memory and Language*, *86*, 1-19.

Collins, C., & Kayne, R. (2011). Syntactic structures of the world's languages.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.

Conwell, C., McMahon, E., Vinken, K., Prince, J. S., Alvarez, G., Konkle, T., ... & Livingstone, M. (2024). Is the visual cortex really "language-aligned"? Perspectives from Model-to-Brain Comparisons in Human and Monkeys on the Natural Scenes Dataset. *Journal of Vision*, *24*(10), 1288-1288.

Correia, J. M., Formisano, E., Valente, G., Hausfeld, L., Jansma, B., & Bonte, M. (2014). Brain-based translation: fMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. *Journal of Neuroscience, 34*(1), 332– 338.

Correia, J. M., Jansma, B., Hausfeld, L., Kikkert, S., & Bonte, M. (2015). Eeg decoding of spoken words in bilingual listeners: From words to language invariant semantic-conceptual representations. *Frontiers in psychology*, 6, 71.

NLLB Team. (2024). Scaling neural machine translation to 200 languages. Nature, 630(8018), 841.

de Varda, A. G., Berzak, Y., Fedorenko, E., & Levy, R. (in preparation). A Comparison between Surprisal and Language Model Embeddings as Predictors of Behavioral and Neural Responses during Language Processing.

de Varda, A. G., & Marelli, M. (2023). Data-driven cross-lingual syntax: An agreement study with massively multilingual models. *Computational Linguistics, 49*(2), 261-299.

de Varda, A. G., & Marelli, M. (2024). The Emergence of Semantic Units in Massively Multilingual Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 15910-15921).

Dehghani, M., Boghrati, R., Man, K., Hoover, J., Gimbel, S. I., Vaswani, A., Zevin, J. D., Immordino-Yang, M. H., Gordon, A. S., Damasio, A., et al. (2017). Decoding the neural representation of story meanings across languages. *Human brain mapping, 38*(12), 6096–6106.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737, 10*.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences, 14*(4), 172–179.

Duncan, J., Assem, M., & Shashidhara, S. (2020). Integrated intelligence from distributed brain activity. *Trends in Cognitive Sciences, 24*(10), 838–852.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences, 32*(5), 429–448.

Everett, D. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current anthropology, 46*(4), 621-646.

Fedorenko, E., Hsieh, P. J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *Journal of neurophysiology, 104*(2), 1177-1194.

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences, 110*(41), 16616-16621.

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024a). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 1-24.

Fedorenko, E., Piantadosi, S. T., & Gibson, E. A. (2024b). Language is primarily a tool for communication rather than thought. *Nature, 630*(8017), 575-586.

Fedorenko, E., & Shain, C. (2021). Similarity of computations across domains does not imply shared implementation: the case of language comprehension. *Current Directions in Psychological Science, 30*(6), 526-534.

Feghhi, E., Hadidi, N., Song, B., Blank, I. A., & Kao, J. C. (2024). What Are Large Language Models Mapping to in the Brain? A Case Against Over-Reliance on Brain Scores. *arXiv preprint arXiv:2406.01538.*

Fisher, R. A. (1921). On the" probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1*, 3–32.

Forgey, M. (2024). *Quantifying the lack of cross-linguistic diversity in fMRI research* (Master's thesis). Universitat Pompeu Fabra, Barcelona, Spain.

Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition, 108*(3), 819-824.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., & Fedorenko, E. (2018). The Natural Stories Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences, 112*(33), 10336-10341.

Gauthier, J., & Levy, R. (2019). Linking artificial and human neural representations of language. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 529-539).

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... & Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences, 114*(40), 10785-10790.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences, 23*(5), 389-407.

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length?. *Cognitive Science, 34*(2), 286-310.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... & Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature neuroscience, 25*(3), 369-380.

Günther, F., Petilli, M. A., Vergallito, A., & Marelli, M. (2020). Images of the unseen: Extrapolating visual representations for abstract and concrete words in a data-driven computational model. *Psychological Research*, 1–21.

Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (2005). The world atlas of language structures. *OUP Oxford.*

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science, 303*(5664), 1634-1640.

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences, 119*(32), e2201968119.

Hirano, Y., Stefanovic, B., & Silva, A. C. (2011). Spatiotemporal evolution of the functional magnetic resonance imaging response to ultrashort stimuli. *Journal of Neuroscience, 31*(4), 1440-1447.

He, P., Gao, J., & Chen, W. (2022). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *The Eleventh International Conference on Learning Representations.*

Honey, C. J., Thompson, C. R., Lerner, Y., & Hasson, U. (2012). Not lost in translation: neural responses shared across languages. *Journal of Neuroscience, 32*(44), 15277-15283.

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2024a). Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 1-21.

Hosseini, E. A., Casto, C. C., Zaslavsky, N., Conwell, C., Richardson, M., & Fedorenko, E. (2024b). Universality of representation in biological and artificial neural networks. *bioRxiv*, 2024-12.

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Ivanova, A. A., Mineroff, Z., Zimmerer, V., Kanwisher, N., Varley, R., & Fedorenko, E. (2021). The language network is recruited but not required for nonverbal event semantics. *Neurobiology of Language, 2*(2), 176-201.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language?. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics.*

Jeong, H., Sugiura, M., Sassa, Y., Haji, T., Usui, N., Taira, M., Horie, K., Sato, S., & Kawashima, R. (2007). Effect of syntactic similarity on cortical activation during second language processing: A comparison of English and Japanese among native Korean trilinguals. *Human Brain Mapping, 28*(3), 194–204.

Johns, B. T., & Jones, M. N. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science, 4*(1), 103–120

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282-6293).

Kauf, C., Tuckute, G., Levy, R., Andreas, J., & Fedorenko, E. (2023). Lexical semantic content, not syntactic structure, is the main contributor to ANN-brain similarity of fMRI responses in the language network. *Neurobiology of Language*, *5*(1), 7-42.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*(1), 109-128.

Klein, S., & Tsarfaty, R. (2020). Getting the## life out of living: How adequate are word-pieces for modelling complex morphology?. In *Proceedings of the 17th SIGMORPHON workshop on computational research in phonetics, phonology, and morphology* (pp. 204-209).

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*(1), 417-446.

Kuperman, V., Schroeder, S., & Gnetov, D. (2024). Word length and frequency effects on text reading are highly similar in 12 alphabetic languages. *Journal of Memory and Language*, *135*, 104497.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177.

Levshina, N., & Moran, S. (2021). Efficiency in human languages: Corpus evidence for universal principles. *Linguistics Vanguard*, *7*(s3), 20200081.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., et al. (2021). Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668.*

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., & Levin, L. (2017). Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 8–14.

Lopopolo, A., Van den Bosch, A., Petersson, K. M., & Willems, R. M. (2021). Distinguishing syntactic operations in the brain: Dependency and phrase-structure parsing. *Neurobiology of Language*, *2*(1), 152-175.

Louradour, J. (2023). Whisper-timestamped.

Roads, B. D., & Love, B. C. (2020). Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, *2*(1), 76-82.

Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., O'grady, L., ... & Levinson, S. C. (2018). Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences*, *115*(45), 11369-11376.

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., ... & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*(8), 1014-1019.

Malik-Moraleda, S., Taliaferro, M., Shannon, S., Jhingan, N., Swords, S., Peterson, D. J., ... & Fedorenko, E. (in press). Constructed languages are processed by the same brain mechanisms as natural languages. *Proceedings of the National Academy of Sciences*

Millet, J., Caucheteux, C., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., & King, J. R. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, 35, 33428-33443.

Moran, S., McCloy, D., & Wright, R. (2014). PHOIBLE online.

Nakagi, Y., Matsuyama, T., Koide-Majima, N., Yamaguchi, H., Kubo, R., Nishimoto, S., & Takagi, Y. (2024, November). Unveiling Multi-level and Multi-modal Semantic Representations in the Human Brain using Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 20313-20338)*.

Nieto-Castanon, A. (2020). FMRI minimal preprocessing pipeline. In *Handbook of functional connectivity Magnetic Resonance Imaging methods in CONN* (pp. 3–16). Hilbert Press. doi:10.56441/hilbertpress.2207.6599

Niu, J., Lu, W., & Penn, G. (2022). Does BERT rediscover a classical NLP pipeline?. In *Proceedings of the 29th International Conference on Computational Linguistics (pp. 3143-3153)*.

Paunov, A. M., Blank, I. A., Jouravlev, O., Mineroff, Z., Gallée, J., & Fedorenko, E. (2022). Differential tracking of linguistic vs. mental state content in naturalistic stimuli by language and theory of mind (ToM) brain networks. *Neurobiology of Language*, *3*(3), 413-440.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of machine Learning research*, *12*, 2825-2830.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., ... & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, *9*(1), 963.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526-3529.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *International Conference on Machine Learning*, 28492–28518.

Sánchez, A., Carreiras, M., & Paz-Alonso, P. M. (2023). Word frequency and reading demands modulate brain activation in the inferior frontal gyrus. *Scientific Reports*, *13*(1), 17217.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108.*

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: a defense of functional localizers. *NeuroImage*, *30*(4), 1088-1096.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Shain, C., Blank, I. A., Fedorenko, E., Gibson, E., & Schuler, W. (2022). Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *Journal of Neuroscience, 42*(39), 7412-7430.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia, 138*, 107307.

Sheikh, U. A., Carreiras, M., & Soto, D. (2019). Decoding the meaning of unconsciously processed words using fMRI-based MVPA. *NeuroImage, 191*, 430–440.

Sheikh, U. A., Carreiras, M., & Soto, D. (2021). Neurocognitive mechanisms supporting the generalization of concepts across languages. *Neuropsychologia, 153*, 107740.

Shliazhko, O., Fenogenova, A., Tikhonova, M., Mikhailov, V., Kozlova, A., & Shavrina, T. (2022). mGPT: Few-Shot Learners Go Multilingual. *arXiv e-prints, arXiv-2204.*

Silva, A. B., Liu, J. R., Metzger, S. L., Bhaya-Grossman, I., Dougherty, M. E., Seaton, M. P., ... & Chang, E. F. (2024). A bilingual speech neuroprosthesis driven by cortical articulatory representations shared between languages. *Nature Biomedical Engineering*, 1-15.

Speer, R. (2022). rspeer/wordfreq: v3.0 (v3.0.2). *Zenodo*. https://doi.org/10.5281/zenodo.7199437

Stanczak, K., Ponti, E., Hennigen, L. T., Cotterell, R., & Augenstein, I. (2022). Same Neurons, Different Languages: Probing Morphosyntax in Multilingual Pre-trained Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1589-1598).

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams Jr, R. M. (1949). The American soldier: Adjustment during army life. *Studies in social psychology in World War II*, vol. 1.

Sueoka, Y., Paunov, A., Tanner, A., Blank, I. A., Ivanova, A., & Fedorenko, E. (2024). The Language Network Reliably "Tracks" Naturalistic Meaningful Nonverbal Stimuli. *Neurobiology of Language, 5*(2), 385-408.

Sun, J., Zhang, X., & Moens, M. F. (2023). Tuning In to Neural Encoding: Linking Human Brain and Artificial Supervised Representations of Language. *arXiv preprint* arXiv:2310.04460.

Tang, J., Du, M., Vo, V., Lal, V., & Huth, A. (2024). Brain encoding models based on multimodal transformers can transfer across language and vision. *Advances in Neural Information Processing Systems, 36.*

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., ... & Fedorenko, E. (2024a). Driving and suppressing the human language network using large language models. *Nature Human Behaviour, 8*(3), 544-561.

Tuckute, G., Kanwisher, N., & Fedorenko, E. (2024b). Language in brains, minds, and machines. *Annual Review of Neuroscience, 47*.

Van de Putte, E., De Baene, W., Brass, M., & Duyck, W. (2017). Neural overlap of l1 and l2 semantic representations in speech: A decoding approach. *NeuroImage, 162*, 106–116.

Van de Putte, E., De Baene, W., Price, C. J., & Duyck, W. (2018). Neural overlap of l1 and l2 semantic representations across visual and auditory modalities: A decoding approach. *Neuropsychologia, 113*, 68–77.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems, 33*, 5776-5788.

Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., ... & Fedorenko, E. (2021). Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *Cerebral Cortex, 31*(9), 4006-4023.

Whitfield-Gabrieli, S., & Nieto-Castanon, A. (2012). Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain connectivity, 2*(3), 125-141.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., . . . Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Association for Computational Linguistics.

Xu, M., Li, D., & Li, P. (2021). Brain decoding in multiple languages: Can cross-language brain decoding work? *Brain and Language, 215*, 104922.

Xu, X., Wu, C., Rosenman, S., Lal, V., Che, W., & Duan, N. (2023). Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 9, pp. 10637-10647).

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). MT5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 483–498.

Yamashita, M., Kubo, R., & Nishimoto, S. (2023). Cortical representations of languages during natural dialogue. *bioRxiv, 2023-08*.

Yang, Y., Wang, J., Bailer, C., Cherkassky, V., & Just, M. A. (2017a). Commonalities and differences in the neural representations of English, Portuguese, and Mandarin sentences: When knowledge of the brain-language mappings for two languages is better than one. *Brain and language, 175*, 77–85.

Yang, Y., Wang, J., Bailer, C., Cherkassky, V., & Just, M. A. (2017b). Commonality of neural representations of sentences across languages: Predicting brain activation during Portuguese sentence comprehension using an English-based model of brain function. *NeuroImage, 146*, 658–666.

Zinszer, B. D., Anderson, A. J., Kang, O., Wheatley, T., & Raizada, R. D. (2016). Semantic structural alignment of neural representational spaces enables translation between English and Chinese words. *Journal of Cognitive Neuroscience, 28*(11), 1749–1759.

Zipf, George K. (1949). *Human Behavior and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley

Supplementary information

S1. Related work

A few lines of past research bear relevance to the current work. For example, one approach to evaluating the similarity of representations across languages has been to train a decoder model on brain activity in response to one language, and then to apply this model to brain responses to a different language (see Xu et al., 2021 for a review). For example, Buchweitz et al. (2012) employed Multivoxel Pattern Analysis (MVPA; Haxby et al., 2001) to develop a decoder model based on English and then assessed its effectiveness on translation-equivalent words in Portuguese, reporting an above-chance performance (see also Correia et al., 2014, 2015; Sheikh et al., 2019, 2021; Zinszer et al., 2016). Another approach had relied on inter-subject correlations during naturalistic language comprehension (Hasson et al., 2004). For example, Honey et al. (2012) demonstrated that fMRI time-series of English and Russian speakers in response to the same narrative (presented in the two languages) were similar across languages in several temporal, parietal, and frontal cortical areas. These sets of studies converge in showing that the language-processing brain areas respond in a sufficiently similar way to different languages to support generalization across linguistic boundaries.

Focusing on higher-level discourse representations, Dehghani et al. (2017) showed that vector representation of narratives (i.e., embeddings reflecting their high-level meaning) can predict brain responses in the default mode network, and that the story representation supporting the mapping could be switched across languages; however, differently from our procedure, the (decoding) models were not transferred across languages but rather re-trained with story representations from other languages, given that the story representations could not be projected onto the same space. This study therefore leaves open the question of whether the general mapping between linguistic and neural representations is actually shared across languages. Most similarly to our approach, Chen et al. (2024) showed that voxelwise encoding models based on contextual (mBERT) and non-contextual (fastText) word embeddings can be transferred across languages in Chinese-English bilinguals.

These previous studies provide valuable insights into how the brain represents information in different languages, but they suffer from some limitations. First, most past studies have used single words presented in isolation (Buchweitz et al., 2012; Correia et al., 2014, 2015; Sheikh et al., 2019, 2021; Van de Putte et al., 2018; Zinszer et al., 2016), and only a few experiments have used sentence stimuli (Chen et al., 2024; Dehghani et al., 2017; Honey et al., 2012; Yang et al., 2017a, 2017b)—the preferred stimulus of the language network, which responds only weakly to isolated words (e.g., see Fedorenko et al., 2024a for a review). Moreover, most studies have used bilingual speakers and presented the same individuals with linguistic stimuli in both of their languages (Buchweitz et al., 2012; Chen et al., 2024; Correia et al., 2014, 2015; Sheikh et al., 2019, 2021; Van de Putte et al., 2017, 2018; but see Dehghani et al., 2017; Honey et al., 2012; Yang et al., 2017a, 2017b; Zinszer et al., 2016). Especially for single-word stimuli, successful cross-lingual decoding might simply reflect the association of translation equivalents in bilingual speakers, as opposed to genuine similarity in neural representations across languages (see Xu et al., 2021 for a discussion of this limitation). Finally, prior research predominantly utilized pattern classification methods (e.g.,

MVPA), which are powerful techniques to detect representational regularities across languages but cannot generate predictions for novel stimuli beyond those presented during the training procedure (a similar limitation characterizes the inter-subject correlation approach), whereas encoding models based on contextualized word embeddings can predict brain responses to arbitrary stimuli. Besides these limitations, past studies have at most considered three languages (Dehghani et al., 2017; Yang et al., 2017a). Although the number of languages included in our current study (21 languages across 7 language families) is still only a tiny fraction of the world's ~7,000 languages, we come a little closer to representing the rich variety that characterizes the human languages.

S2. Results in individual languages (Study I)

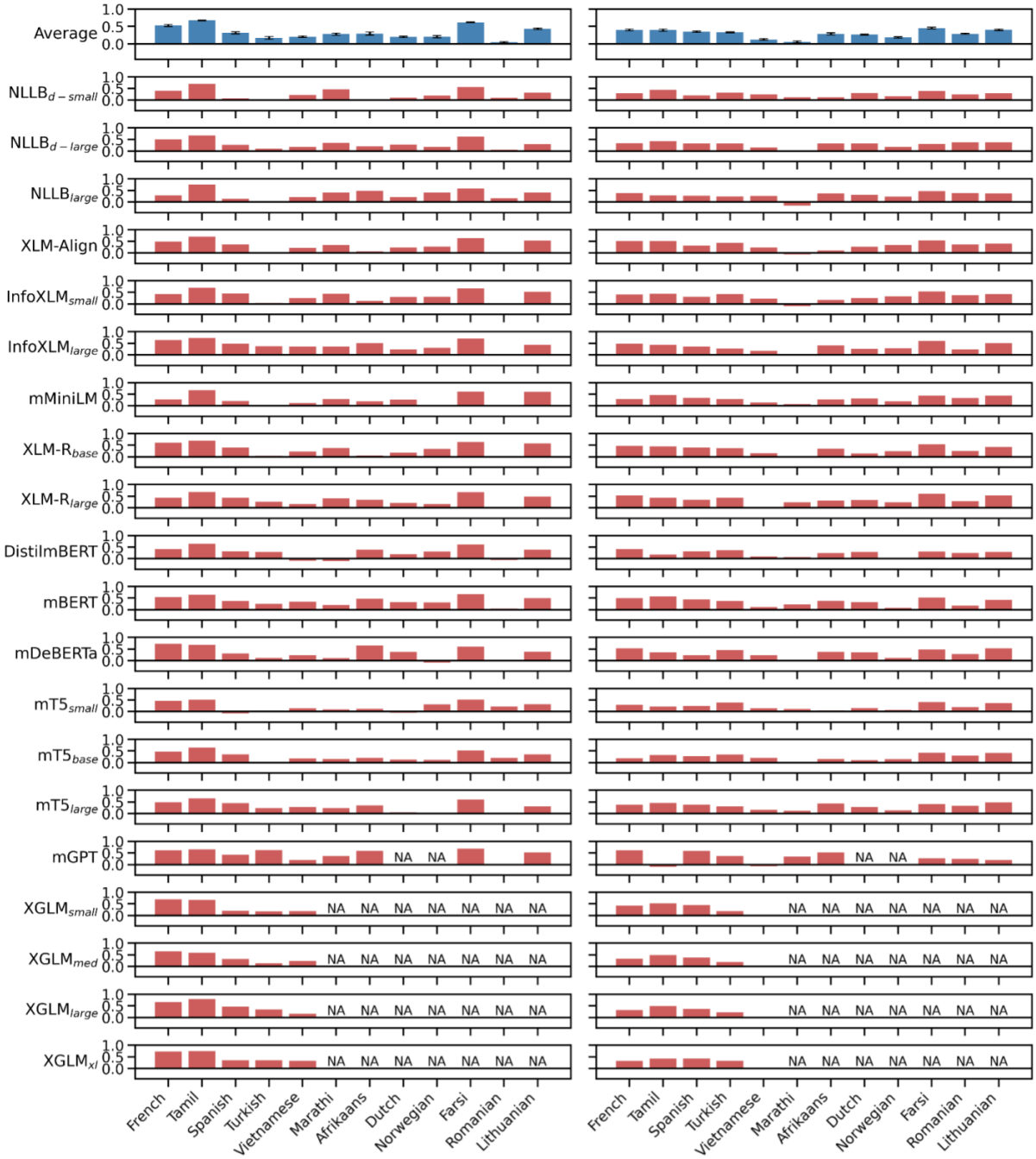
In the main text, we aggregated the encoding results across languages to emphasize the cross-lingual consistencies in encoding performance. This approach was driven by the primary objective of our study, which was to uncover language-general patterns and trends that hold across a variety of languages, and by the sheer volume of the dataset (20 language models, each comprising several layers ranging from 6 to 48, and data spanning 12 languages). This section presents a more granular view of our results. Here, we report the best-layer encoding performance for each language and model combination, as well as the results averaging across language models, both for the WITHIN (**Supplementary Figure 1**, left) and the ACROSS conditions (**Supplementary Figure 1**, right).

S3. Results in individual languages (Study II)

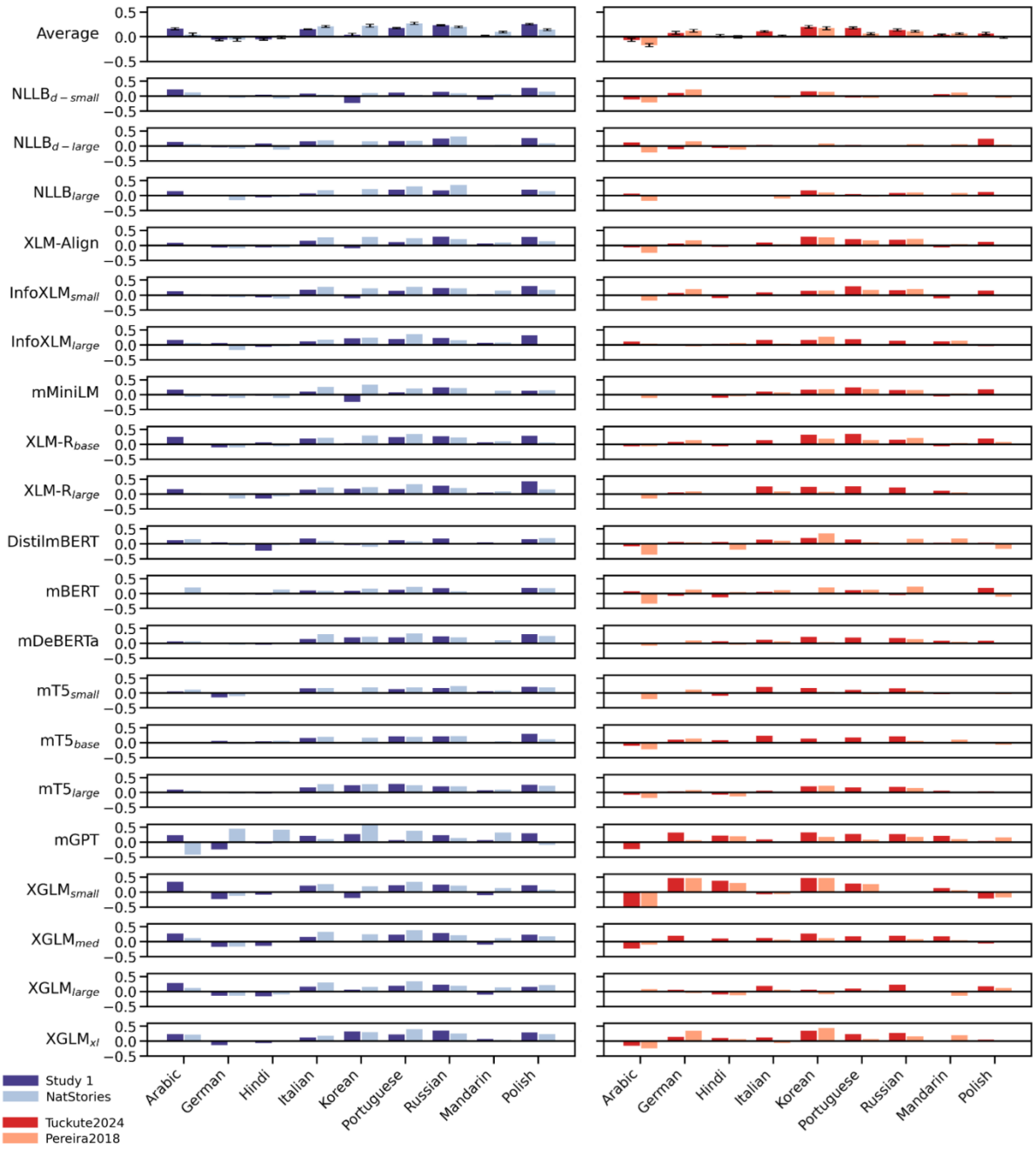
As for Study I, we report the transfer performance for each language and model combination, both for the encoding models trained on passage/story data (**Supplementary Figure 2**, left) and on sentence data (**Supplementary Figure 2**, right). Encoding performance is evaluated considering a single layer for each model, i.e., the one that was associated with the strongest encoding performance in Study I (ACROSS condition). Differently from Study I, where mGPT and XGLM were not evaluated in all languages (because some of them were not represented in their pre-training data), in Study II all models were evaluated in all the languages considered.

S4. WITHIN encoding performance using data from Study II

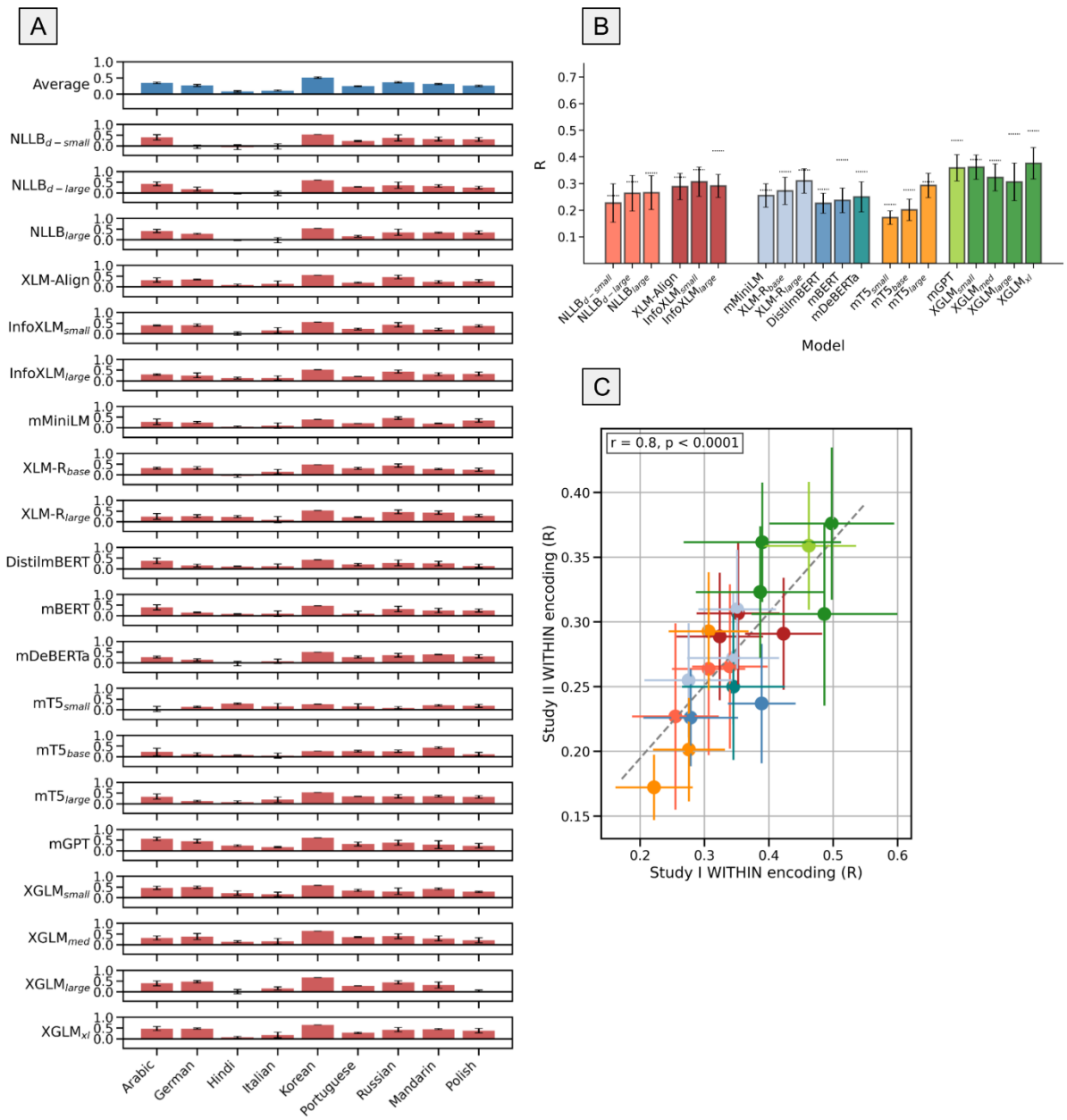
To evaluate whether language representations from MNNLMs could predict brain responses in the 9 additional languages we considered for Study II, we replicated the WITHIN encoding analysis (Results, "Study I – Language models predict brain responses in diverse languages" and Methods, "Study I – WITHIN encoding models") using the newly collected fMRI data. We kept all the details concerning model fitting (e.g., penalty estimation, normalization, cross-validation approach) identical to Study I with one exception: instead of evaluating all layers for each model, we only considered a single layer per model, namely the layer which obtained the best performance in Study I. Since in Study II we collected brain responses to three passages in each language (some of which were discarded due to low data quality, see Methods, "Study II –



Supplementary Figure 1. Best-layer encoding performance scores for each model and language, for the WITHIN (left) and ACROSS (right) condition. The rows on the top report the language-wise encoding performance averaged across models, with the error bars indicating the standard error of the mean across models. Note that XGLM encoding scores are not available for Marathi, Afrikaans, Dutch, Norwegian, Farsi, Romanian, and Lithuanian, and mGPT encoding results are not available for Dutch and Norwegian.



Supplementary Figure 2. Transfer performance of the encoding models trained on sentences (left; Study I: dark blue, *NatStories*: light blue) and continuous passage data (right; *Tuckute2024*: dark red; *Pereira2018*: pink) obtained in each individual language considered in Study II. The rows on the top report the language-wise encoding performance averaged across models, with the error bars indicating the standard error of the mean across models.



Supplementary Figure 3: WITHIN encoding performance obtained on the fMRI data from Study II. *A*: Encoding scores for each model and language, averaged across passages. The error bars indicate the standard error of the mean of the encoding performance across passages. The row on top reports the encoding performance averaged across models, with the error bars indicating the standard error of the mean across models. *B:* Model-wise summary of the encoding results. The horizontal, dotted lines indicate the performance obtained by the models in Study I (WITHIN condition). *C:* Relationship between each model's WITHIN encoding performance obtained on the data from Study I (*x*-axis) and Study II (*y*-axis). The error bars report the standard error of the mean of the encoding performance along the two axes.

Time-series extraction"), we averaged encoding performance scores across passages, excluding those that did not pass the selection criteria.

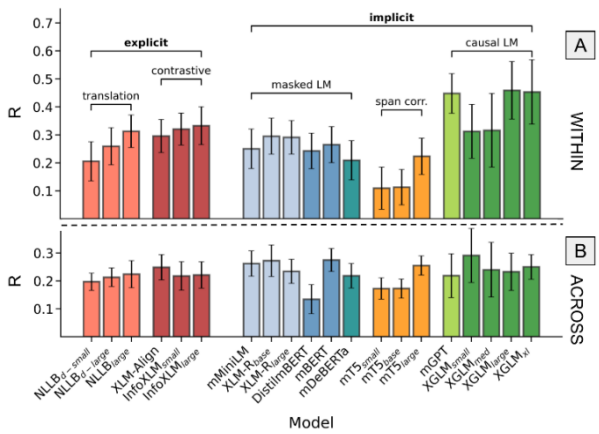
Confirming our observations from Study I, our results showed that the MNNLMs we considered were able to predict fMRI responses in most languages (**Supplementary Figure 3A**). Aggregating the results across languages, all MNNLMs were significantly predictive of brain responses (all $p < 0.001$;

Supplementary Figure 3B). The encoding performance obtained by the various models on the Study II data was significantly correlated with the encoding scores obtained in Study I (r = 0.80, p < 0.001, **Supplementary Figure 3C**), indicating that inter-model differences in brain predictivity are highly reliable.

S5. Performance in the median layer

In the main body of our paper, we presented the results obtained by the best-performing layer of each neural language model. This approach highlights the peak capabilities of each model, but it is also important to consider some indices of the central tendency of the models' performance. To this end, we report the results of the median-performing layer of each model. This analysis is relevant given that models have a varying number of layers, and larger models with more layers could potentially show a wider range of performance levels (and thus, higher *maxima*). Focusing on the median layer mitigates the impact of this variability.

The encoding performance scores for the median layers of each model are detailed in **Supplementary Figure 4**. This figure indicates that although there is a small decrease in performance compared to the best-layer results, the correlation between the reconstructed and the original fMRI responses remains positive and significant for all the models under consideration, both in the WITHIN (all p < 0.0036; **Supplementary Figure 4A**) and in the ACROSS condition (all p < 0.0002; **Supplementary Figure 4B**).



Supplementary Figure 4. Encoding results obtained by the median-performing layer in each network under consideration, both in the WITHIN (A) and in the ACROSS (B) condition. The scores are averaged across languages. The error bars indicate the standard error of the encoding performance across languages.

S6. Word frequency and length baseline

To assess the extent to which brain responses could be captured by surface linguistic properties, we fitted simple encoding models where brain activity was predicted by word-level frequency and length, averaged over TRs. Word length was defined as the number of characters in each word, whereas frequency (Zipf-transformed) was obtained with the Python package wordfreq (version 3.0), a library for looking up word frequencies in many languages (Speer, 2022). All other details concerning pre-processing, normalization, model fitting, etc. were left unaltered with respect to the main approach detailed in the Methods ("Study I – Encoding models").

For Study I, brain responses were predicted both within a single language, with sequential 10-fold cross-validation (as in the WITHIN condition), and by transferring the frequency-length models to novel languages (as in the ACROSS condition). Results were averaged across languages. Afrikaans and Marathi

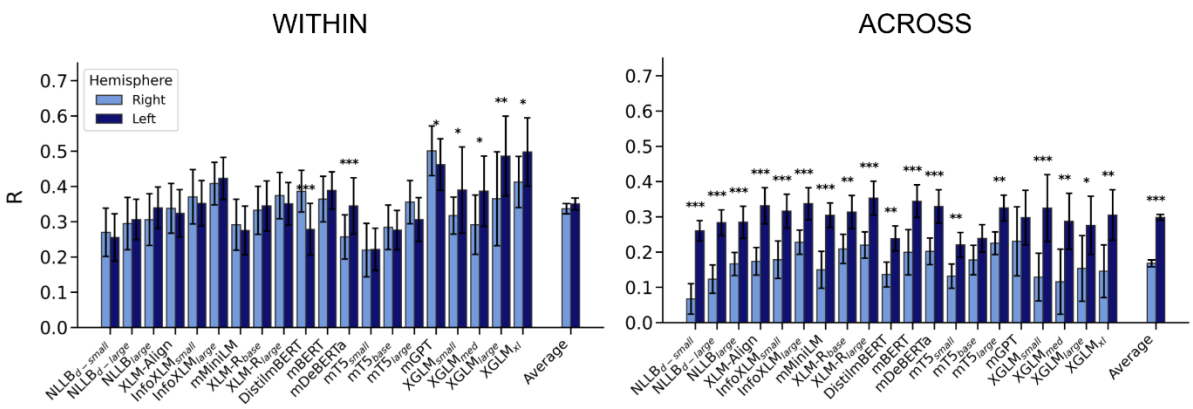
were discarded from the analyses because frequency estimates for these two languages were not available through the wordfreq package. In the WITHIN condition, the model based on frequency and length was not predictive of fMRI responses (r = -0.03, SE = 0.04); in the ACROSS condition, the frequency-length model performance showed less than half of the performance obtained by the worst encoding model (r = 0.09, SE = 0.02).

For Study II, we trained the encoding models based on frequency and length on the training data (Study I data, *NatStories*, *Pereira2018*, *Tuckute2024*), and tested them on the newly collected data, following the general procedure we used for the standard encoding analyses performed for Study II (Methods, "Study II – Brain encoding models"). Results were averaged across passages and languages. The models based on frequency and length achieved little to no predictivity across all four training datasets (Study I data: r = 0.06, SE = 0.03; *NatStories*: r = 0.004, SE = 0.02; *Pereira2018*: r = -0.05, SE = 0.03; *Tuckute2024*: r = 0.03, SE = 0.03).

S7. Spatial specificity: right-hemisphere language network encoding

To evaluate the spatial specificity of the encoding results, we tested whether encoding models based on MNNLMs could predict brain responses equally well in both the left-hemisphere and right-hemisphere language network. It is well-established that language processing predominantly occurs in the left hemisphere for most right-handed individuals (Fedorenko et al., 2024a). This analysis tests whether encoding performance mirrors this left-hemispheric bias.

In the WITHIN setting, the encoding performance averaged across models in the right-hemisphere language areas is r = 0.34, compared to r = 0.35 in the left-hemispheric language network, showing only a slight 3.86% average decrease (**Supplementary Figure 5**, left). This small difference is primarily driven by the best-performing models, particularly those in the XGLM family. However, the results are mixed for other models; for instance, DistilmBERT and mGPT show better performance in the right hemisphere compared to the left. Interestingly, in the ACROSS setting, the difference between hemispheres becomes more pronounced (**Supplementary Figure 5**, right). The encoding performance averaged across models in the right-hemisphere language areas is r = 0.17, whereas in the left hemisphere, it is r = 0.30, indicating a 43.71% decrease in performance. This difference is significant for most individual models (all but $mT5_{base}$ and mGPT) and when aggregating across models ($p < 0.001$), in line with the left hemisphere's dominant role in processing linguistic information across languages (see also Malik-Moraleda, Ayyash, et al., 2022). We remark that the hemispheric asymmetry is substantially more pronounced in the ACROSS condition, highlighting the low performance of cross-lingual transfer in the right hemisphere.

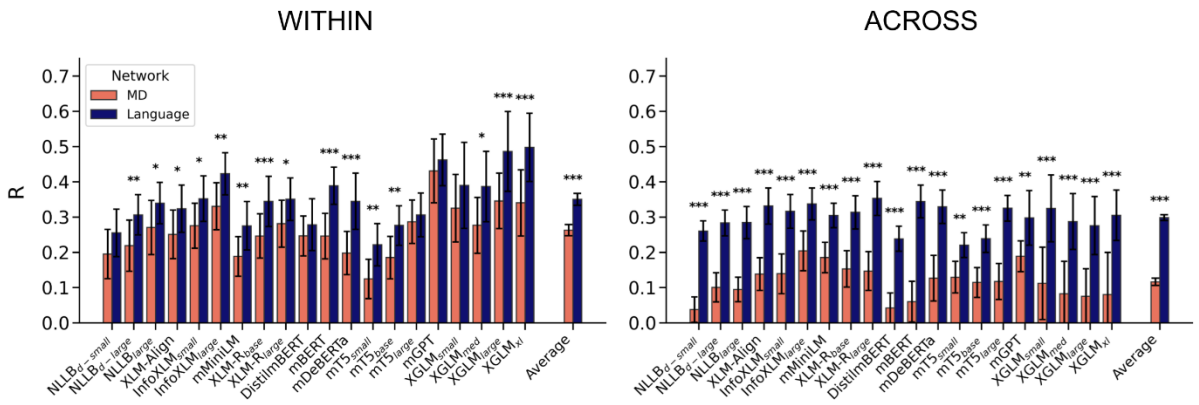


Supplementary Figure 5. Best-layer encoding performance in the right (light blue) and left (blue) hemispheric regions of the language network in the WITHIN (left) and ACROSS (right) condition. The error bars indicate the standard error of the mean across languages. The two additional bars on the right of each subplot ("Average") indicate the encoding performance averaged across models, with the error bars reporting the standard error of the mean across models.

S8. Spatial specificity: multiple demand network encoding

To further evaluate the spatial specificity of the encoding results, we tested whether encoding models based on MNNLMs could predict brain responses in another high-level cognitive network. In particular, we analyzed fMRI BOLD signals from the multiple demand (MD) network, a set of functionally connected bilateral brain areas implicated in executive functions and fluid intelligence (Duncan, 2010; Fedorenko et al., 2013; Assem et al., 2020a). The MD network was functionally defined using a commonly used localizer task based on a spatial working memory task (e.g., Assem et al. 2020b), following the approach used for the language network. The parcels for the MD network were derived from a large independent set of participants and are available from https://www.evlab.mit.edu/resources-all/download-parcels.

In the WITHIN setting, the encoding performance in the MD network averaged across models is r = 0.26, compared to r = 0.35 in the left-hemisphere language network, which represents a 24.91% decrease (**Supplementary Figure 6**, left). This difference is significant for 15 out of the 20 models, and the pattern remains consistent across all models. In the ACROSS setting, the asymmetry is more pronounced. The encoding performance averaged across models is r = 0.12, compared to r = 0.30 in the language network, which represents a 61.03% decrease (**Supplementary Figure 6**, right). This difference is significant for all models considered.



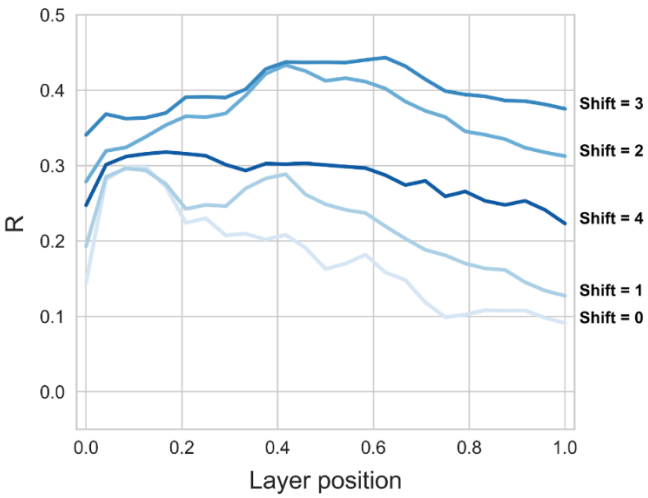
Supplementary Figure 6. Best-layer encoding performance in the MD (red) and language network (blue) obtained by each model in the WITHIN (left) and ACROSS (right) condition. The error bars indicate the standard error of the mean across languages. The two additional bars on the right of each subplot ("Average") indicate the encoding performance averaged across models, with the error bars reporting the standard error of the mean across models.

S9. Optimal TR shift

The hemodynamic response measured with fMRI peaks several seconds after the presentation of a stimulus (e.g., Hirano, Stefanovic, & Silva, 2011). Previous work in computational language neuroscience has addressed the issue by either (a) implementing an arbitrary shift corresponding to 2 TRs (Schrimpf et al., 2021), or (b) concatenating embedding representations of the stimuli corresponding to the current and previous temporal resolution intervals to estimate the hemodynamic response function directly from the embeddings (e.g., Aw & Toneva, 2023). However, the approach in (a) does not ensure that the assumed temporal delay is optimal, and the approach in (b) increases the dimensionality of the predictors, raising possible concerns about over-fitting especially in settings where the amount of data is limited. To overcome these limitations, we have identified in a separate dataset the optimal temporal shift for the specific task at hand, i.e., MNNLM-based encoding of brain responses in a naturalistic listening task. We have conducted these analyses on the *NatStories* fMRI corpus (see Methods, "NatStories"). Our procedure was analogous to what we reported in the Methods section; we averaged the fMRI time-series across voxels for each fROI, subsequently across fROIs within the language network, and lastly across participants. We obtained word-by-word speech transcriptions with Whisper-timestamped, and used $\text{XGLM}_{\text{small}}$ to derive word embeddings for the obtained texts, which were then aligned with the fMRI responses by averaging the embeddings over the TR intervals. As a final step, we fitted independent encoding models for each temporal shift to identify the optimal delay. We tested five different delays: 0, 1, 2, 3, and 4 TR intervals, equivalent to 0, 2, 4, 6, and 8 seconds, respectively. For instance, with a delay of 1 TR interval, we used the embeddings corresponding to the ith TR as predictors and the fMRI response corresponding to the $i+1^{\text{th}}$ image as a response. Similarly, with a delay of 2 TR intervals, the embeddings for the ith TR were aligned with the fMRI response of the $i+2^{\text{th}}$ image, and so forth. This systematic evaluation allowed us to determine the optimal temporal alignment between the neural language model embeddings and the fMRI responses, ensuring that our encoding models were accurately capturing the temporal dynamics of the hemodynamic response. The encoding models were

fitted with cross-validation, where the linear encoding models were iteratively fitted in eight out of nine stories, and evaluated on the held-out story.

Performance of the encoding models increased with the temporal delay up to a shift of 3 TR intervals (see **Supplementary Figure 7**). Specifically, we observed a progressive improvement in the model's ability to predict fMRI responses as the temporal delay increased from 0 to 3 TRs. However, performance decreased with a shift of 4 TR intervals, supporting the notion that the optimal temporal delay for capturing the hemodynamic response in our task is around 6 seconds. Therefore, for the main analyses of this paper, we adopted a temporal shift of 3 TR intervals (6 seconds).



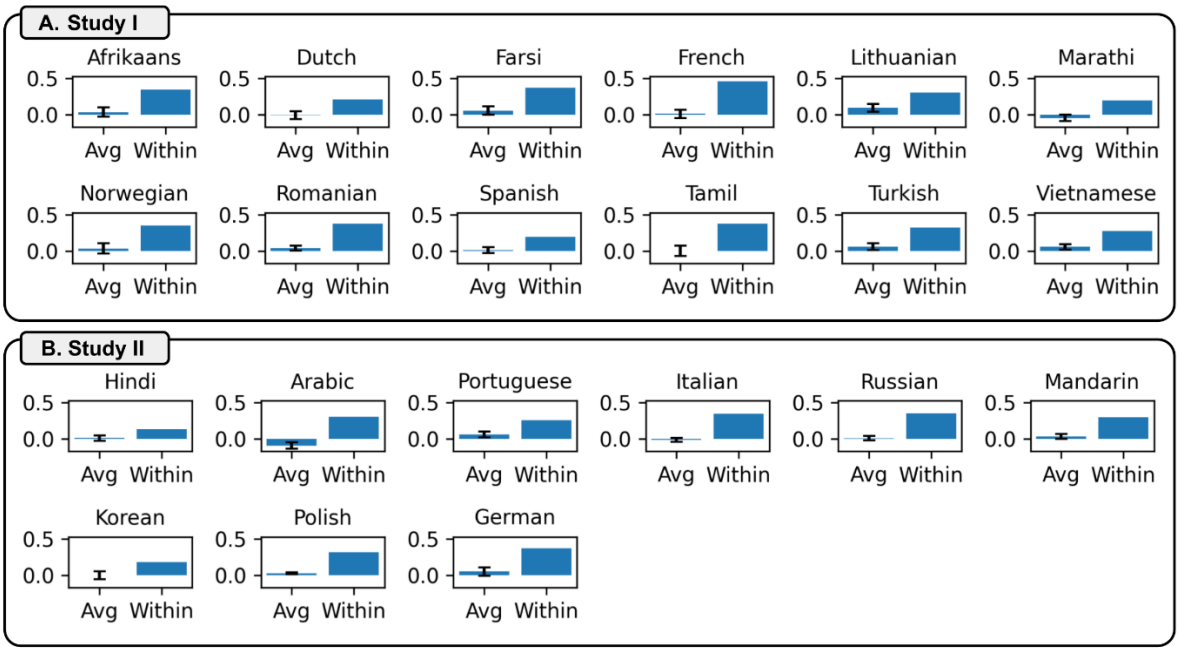
Supplementary Figure 7. Layer-wise encoding performance as obtained by $XGLM_{small}$ on held-out passages in the *NatStories* fMRI data. We evaluated five different TR shifts to account for the lag in the hemodynamic response, corresponding to 0, 2, … 8 seconds. We found that the optimal encoding performance was obtained with a 3-TR shift, corresponding to a lag of 6 seconds.

S10. Time-series correlation

To pre-select the fMRI data for our analyses, we examined the reliability of BOLD responses to the linguistic stimuli. In particular, we calculated the Pearson correlation between the time-series of the two/three participants in each language (Methods, "Study I – Time-series extraction" and "Study II – Time-series extraction"). In Study I (where two participants listened to a single passage for each language), the languages where the time-series of the two participants did not show a significant correlation were excluded from our sample; in Study II (where three participants participant listened to three passages for each language), we used this criterion to select participants and passages for each language. As a sanity check, we additionally compared these within-language correlations to the correlations of participants' fMRI responses across different languages (for Study II, we performed this analysis for the different passages, restricting to the combinations of languages and passages that were used in the critical analyses, and averaged the correlations across passages). This analysis helps rule out the possibility that the results that we obtained in the ACROSS condition are due to the correlation in the fMRI time-series across languages.

Supplementary Figure 8 summarizes the results for Study I (**Supplementary Figure 8A**) and Study II (**Supplementary Figure 8B**). The correlations within each language (the 'Within' bars) are consistently higher than the average correlation observed between each participant's responses in one

language and all other participants across different languages (the ‘Avg’ bars); the latter correlations cluster around zero.



Supplementary Figure 8. Comparison of the correlation coefficients between participant pairs within and across languages, both for Study I (A, top) and Study II (B, bottom). The ‘Within’ bars represent the correlation between the time-series of two participants in the same language. The ‘Avg’ bars show the average correlation between each participant’s fMRI responses in one language and those of all other participants across different languages. Error bars indicate the standard error of the mean of these correlations across languages. In the case of Study II (B, bottom), we averaged the correlations across passages. The plot shows that the time-series of two participants listening to the same passage are correlated, but the correlation across languages is low.

S11. Context contamination

A critical aspect often overlooked in computational language neuroscience is the potential for context contamination in transformer-based encoding models. This section presents an approach to evaluate the extent of such contamination.

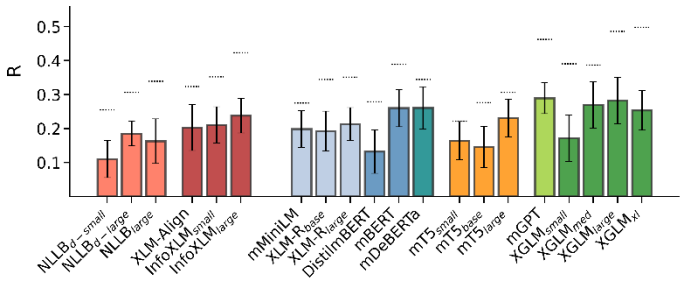
Context contamination might arise when self-attention mechanisms in language models retain information from previous data folds in a cross-validation setting. If the input text is processed in a single pass, the embeddings generated by the model may carry over contextual information from one fold to the next. This carryover can confound results, leading to an overestimation of a model’s true encoding capacity, as it may partially rely on residual information from preceding folds rather than on the current input. Note that, in our approach, context contamination might arise only in the WITHIN setting, as the cross-lingual encoding models are trained and evaluated on different passages in different languages.

To address this potential confound, we refitted our encoding models with a modification: resetting the context after each fold when obtaining the models’ embeddings. This reset aims to emulate a “clean slate” for each new data fold, ensuring that the generated embeddings are a product of just the current fold’s

input and creating a more stringent testing scenario for the encoding models. It's important to note that, in the context of fMRI data collection, participants were exposed to a continuous stream of language stimuli, allowing for a natural accumulation of context. However, our models, with the reset applied, will not have access to such prior input.

In these analyses, we recomputed the contextualized word embeddings from the various models considered in the study separately for each fold. In the case of bidirectional models, a causal setup was simulated as in the main analyses. Consistently with the main analyses, we refitted the encoding models with 10-fold cross-validation and sequential data splits. Encoding performance was measured as the Pearson correlation between the reconstructed fMRI response (i.e., the concatenation of the predictions obtained in each fold) and the recorded time-series.

The results obtained by resetting the models' context at the fold boundaries are reported in **Supplementary Figure 9**. The figure displays a contained decline in performance with respect to what we reported in the main analyses (horizontal dotted lines, see also **Figure 2A**). Indeed, the encoding performance averaged across models is r = 0.21, compared to r = 0.34 obtained with the full context (40% decrease). With the available data, it is difficult to ascertain whether this drop in encoding performance is a consequence of context contamination or rather a result of a mismatch between how the models process the passages (with a constrained context size) and how humans do (with full access to the passage). However, the fact that the difference in performance is relatively contained indicates that neither of these factors plays a decisive role.



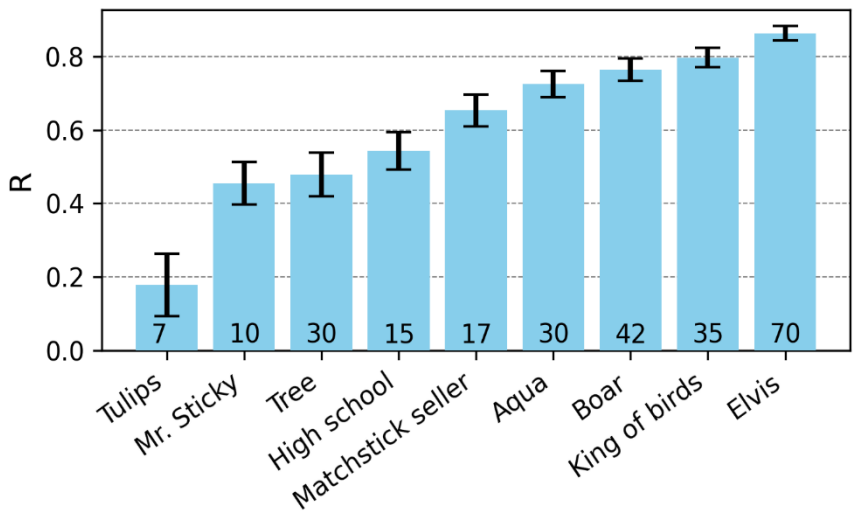
Supplementary Figure 9. Best-layer WITHIN encoding performance obtained after resetting the models' context across folds. The scores are averaged across languages. The error bars report the standard error of the encoding performance across languages. The horizontal, dotted lines on top of the bars indicate the performance obtained by the various models in the standard setting, without the context being reset at fold boundaries (thus corresponding to the results reported in Figure 2A).

S12. Data reliability

To get a sense of the quality of the fMRI data employed across Study I and Study II, we conducted a reliability analysis. The results of the analysis are reported in **Figure 1C**.

In the case of Study I, we measured the correlations between the time-series of the two participants in each of the 12 languages considered (Methods, "Study I – Time-series extraction") and averaged them. For Study II, we adopted the same approach, but for the passages where the time-series were reliable for three participants, we averaged the correlations across the three possible participant pairs. The final reliability estimate was obtained by first averaging across participants, then across passages (only considering the passages included in the final sample, see Methods, "Study II – Time-series extraction"), and lastly across languages.

For the additional datasets considered for Study II (*Pereira2018*, *Tuckute2024*, and *NatStories*), which included a larger number of participants, we calculated the split-half reliability of the responses. In the case of *NatStories*, for each passage, we randomly assigned the participants to two different groups, averaged their time-series within the two groups, and calculated their correlation. This procedure was repeated 1,000 times to obtain more stable estimates. This analysis showed that the majority of the passages obtained a high split-half reliability score (see **Supplementary Figure 10**). An aggregated reliability score was obtained as a weighted average of the reliability scores of each passage, with the weights being proportional to the number of TR-level responses for each passage (ranging from 151 to 213). A similar procedure was adopted for *Pereira2018* and *Tuckute2024*, with the key difference that for these two datasets each datapoint corresponded to a response to a single sentence. *Pereira2018* includes data from two studies (Experiment 2 and Experiment 3), with different numbers of participants (9 and 6). Reliability was estimated independently in each dataset and then aggregated with a weighted average, with the weights being proportional to the number of sentences in each experiment (384 and 243).

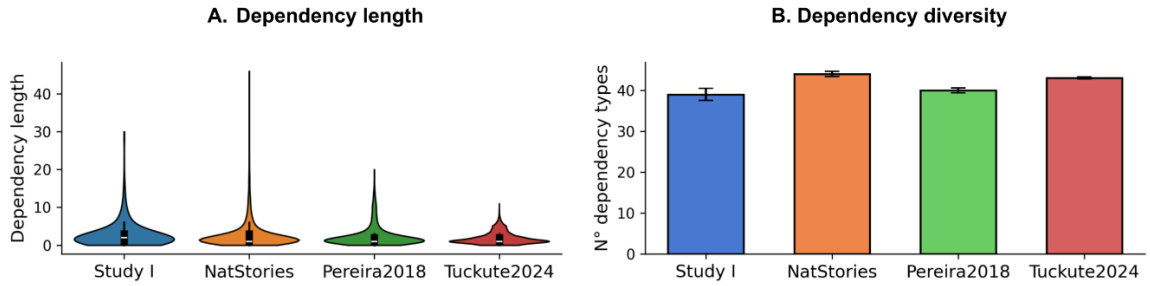


Supplementary Figure 10. Split-half reliability for each passage in the *NatStories* fMRI corpus. At the base of each bar, we report the number of participants for each passage.

S13. Syntactic complexity

To measure variation in syntactic complexity across the different datasets employed in Study II (Study I data, *NatStories*, *Pereira2018*, and *Tuckute2024*), we used two complementary metrics: dependency length and dependency diversity. Dependency length measures the linear distance between tokens and their syntactic heads, whereas dependency diversity quantifies the variety of dependency relations within each dataset, reflecting the range of grammatical constructions.

Sentences in all datasets were tokenized and parsed using the SpaCy dependency parser (en_core_web_sm). For each sentence, dependency length was calculated as the absolute difference between the index of a token and its syntactic head. To quantify the diversity of syntactic structures, dependency relations (e.g., *nsubj*, *dobj*, *prep*) were extracted for all tokens in each dataset. The number of unique dependency types in each dataset was calculated as a proxy for syntactic diversity.



Supplementary Figure 11. *A:* Distribution of dependency length across the four datasets employed in Study II. *B:* Number of unique dependency types across the four datasets. The error bars indicate the bootstrapped standard deviation as calculated across 1,000 samples with replacement.

This analysis reveals differences in both syntactic depth and diversity across datasets. The *NatStories* datasets includes syntactic dependencies that are both longer, especially in their extremes (**Supplementary Figure 11A**) and more diverse (**Supplementary Figure 11B**). The Study I dataset comprises syntactic dependencies that are intermediate in length, but with low variability in dependency types. The sentences in the *Pereira2018* and *Tuckute2024* datasets are characterized by short dependencies; furthermore, the *Pereira2018* dataset includes comparatively fewer unique dependency relations.

S14. Typological factors and transfer performance

In a follow-up analysis, we attempted to identify whether typological factors influence the quality of the transfer of an encoding model from a given language to another. To do so, we restricted our focus to XLMRlarge's 15th layer (the best-performing layer of the best-performing model in the ACROSS condition). Our procedure was based on fitting an encoding model in an individual language $l_i \in L$, and transferring it zero-shot to all remaining languages $l_j \in L \setminus \{l_i\}$ (cf. our main approach in Study I, where we fitted encoding models in $l_j \in L \setminus \{l_i\}$ languages and transferred them to the held-out language $l_i \in L$; Methods, "Study I – Brain encoding models"). The encoding performance was once again evaluated as the Pearson correlation between the predicted and the actual brain response in all $l_j \in L \setminus \{l_i\}$. Note that this approach gives, for each language pair (l_i, l_j), a first transfer score obtained by transferring the encoding model learned on l_i to l_j, and a second score obtained by transferring the model from l_j to l_j. Since these two scores were highly correlated across the language pairs ($r = 0.6902$, $p < 0.001$), we decided to average them to derive a single transferability metric.

Once obtained the pairwise transferability scores, we evaluated whether they correlated with several measurements of similarity between the two languages. In particular, we considered:

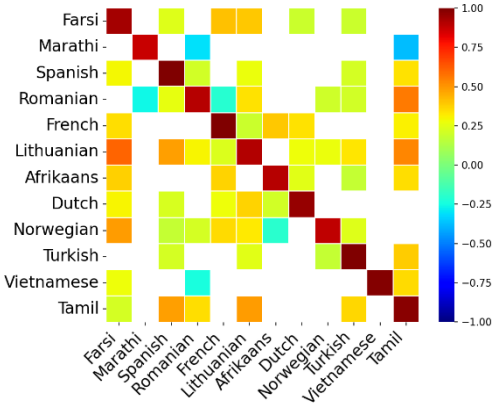
1. **Genetic distance.** A distance metric based on the Glottolog's language tree. The metric calculates the distance between two languages by counting the number of upward steps required on the tree until both languages converge at a common node. This number is then divided by the total number of branches from the first language to the tree's root.
2. **Geographical distance.** The distance on the earth surface between the geographical locations where two languages are spoken.

3. **Syntactic distance.** This metric measures the pairwise distance between two languages as the cosine distance between feature vectors describing the languages' syntax. These features encode attributes such as word order or the presence of specific constructions. The syntactic features are derived from the World Atlas of Language Structures (WALS; Haspelmath et al., 2005), the Syntactic Structures of the World's Languages (SSWL; Collins & Kayne, 2011), and Ethnologue (Campbell, 2008).
4. **Phonological distance.** Similarly to syntactic distance, phonological distance is computed over phonological feature vectors based on WALS and Ethnologue characterizing the languages.
5. **Phonetic distance.** This distance metric measures the distance between two languages as the cosine distance between the phonetic feature vectors derived from PHOIBLE (Moran, McCloy, & Wright, 2014).

All the language features were accessed from the URIEL knowledge base, queried with the lang2vec package (Littell et al., 2017).

Supplementary Figure 12 reports the results of the pairwise transfer. The correlational analyses relating the linguistic distance metrics with the transferability scores showed no meaningful pattern of association between language similarity and cross-lingual encoding transfer: indeed, transferability was not significantly associated with genetic (r = 0.0272, p = n.s.), geographical (r = 0.1273, p = n.s.), syntactic (r = -0.1916, p = n.s.), phonological (r = 0.0009, p = n.s.), or phonetic distance (r = 0.0109, p = n.s.).

The absence of a reliable association between encoding transferability and language similarity, as measured with various metrics, might seem surprising at first, and is in contrast with previous findings showing slight transfer improvements for syntactically similar languages (Jeong et al., 2007). However, we suggest that the absence of such a relationship is in line with the type of linguistic information that is responsible for the alignment between computational language models and brain responses. It has been shown that lexical semantic content, as opposed to syntactic structure, is the main contributor to the alignment between representations from language models and fMRI recordings (Caucheteux et al., 2021; Gauthier & Levy, 2019; Kauf, Tuckute et al., 2023; see Discussion). Under this view, transfer performance should not be affected by typological differences across languages, which manly deal with form (as opposed to compositional meaning). Yet, we remark that single-language results are expected to be noisy, since individual encoding models trained and evaluated on single-language data are subject to variability originating from factors such as the limited number of participants, differences in data quality across languages, and the relatively small dataset sizes. This variability is in contrast with model-level analyses, where averaging across multiple languages helps smooth out language-level fluctuations and produces more robust estimates. Thus, the noise in single-language analyses could make it challenging to detect subtle patterns of association between typological similarity and transfer performance. Furthermore, these results should be interpreted with caution since they represent a null result, which lacks the evidence weight of a significant finding and cannot definitively rule out the existence of small effects that are impossible to detect with the current statistical power.

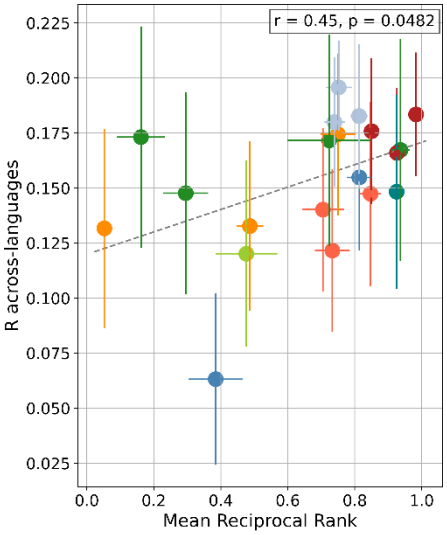


Supplementary Figure 12: Pairwise transfer performance obtained with XLMR$_{\text{large}}$ (layer 15). The rows and columns indicate the language on which the encoding model is trained and tested, respectively; each cell in the heatmap reports the correlation between the predicted and the actual response. Only the significant correlations are displayed ($p < 0.05$, uncorrected).

S15. Cross-lingual alignment of representation spaces supports successful transfer to new languages (replication with Study II data)

In the main text of the article ("Next-word prediction ability explains inter-model differences in within-language encoding performance, and cross-lingual alignment of representation spaces supports successful transfer to new languages"), we showed that the degree to which representations are aligned across languages within each model predicts the models' transfer performance to new languages. To replicate the analysis with the new data from Study II, we followed the same approach used in the original analysis. We began with a set of semantically diverse English sentences (Pereira et al., 2018, Experiment 2 and 3 materials) and translated them into the nine additional languages collected for Study II using the Google Translator API. For each language pair, we obtained embeddings for these sentences from each of the tested MNNLMs, and for each model and language pair, we constructed a cosine distance matrix between the embeddings of all sentences and computed the mean reciprocal rank (MRR) of the correct translation. Finally, we averaged the MRR scores across all language pairs to derive a single alignment metric for each model.

Transfer performance was assessed by replicating the approach employed in Study I (ACROSS condition) to the data we collected in the context of Study II: for each passage, we trained encoding models in $l_j \in L \setminus \{l_i\}$ languages and transferred them to the held-out language $l_i \in L$ (Methods, "Study I – Brain encoding models"). We then averaged the encoding results across languages and passages, and correlated the encoding scores with the MRR scores. We discarded the data from the passages where fMRI time-series were not reliable, as assessed with their correlations across participants (Methods, "Study II – Time-series extraction"). **Supplementary Figure 13** shows that the results we obtained with the data from Study I generalize to the data we collected for Study II, with the correlation between cross-lingual alignment and transfer scores being positive and significant ($r = 0.45$, $p < 0.05$).



Supplementary Figure 13: Relationship between the models' cross-lingual alignment and encoding performance in the transfer, when evaluated on the Study II data. Cross-lingual alignment (*x*-axis) is measured as the mean reciprocal rank of the embedding distance of correct translations compared to unrelated sentences in an independent dataset (Pereira et al., 2018, Experiment 2 and 3 materials). Higher values indicate stronger cross-lingual alignment. The error bars indicate the language-wise standard error of the mean on the two axes. The gray, dashed line indicates the best linear fit.

S16. Language-level analyses

In the paper, we showed that the models' next-word prediction abilities and the alignment of their representations across languages accounted for inter-model differences in encoding performance ("Next-word prediction ability explains inter-model differences in within-language encoding performance, and cross-lingual alignment of representation spaces supports successful transfer to new languages"). To assess whether those features additionally predicted differences in encoding performance across languages, we repeated those analysis following an identical procedure, but averaging across models instead of across languages. With this approach, for each language *li*, we obtained one single datapoint for encoding performance, perplexity (next-word prediction ability), and MRR (cross-lingual representational alignment) by averaging those values across the models that supported *li* (20 for French, Tamil, Spanish, Turkish, Vietnamese; 16 for Marathi, Afrikaans, Farsi, Romanian, Lithuanian; 15 for Dutch and Norwegian).

Differences in encoding performance across languages could not be reliably predicted by perplexity (WITHIN: r = 0.1426, p = n.s.; ACROSS: r = 0.1681, p = n.s.) nor MRR scores (WITHIN: r = -0.3135, p = n.s.; ACROSS: r = 0.0965, p = n.s.). We attribute this null result to the noise in the encoding scores obtained in each individual language, possibly stemming from the different reliability of brain data for specific languages and the relatively low number of responses in each language.

S17. Language resources

Between Study I and Study II, our experiments considered a total of 22 languages. In this section, we present an overview of the resources available for those languages, both in the context of NLP (focusing on the availability of textual data) and neuroscience research (focusing on the number of publications in language neuroscience). For the former, we considered as an index of resource availability the number of tokens in a given language, using the mC4 corpus as a reference (Xue et al., 2021). For the latter, we considered the estimates provided by Malik-Moraleda, Ayyash, et al. (2022). We report this information in **Supplementary Table 1**, which shows that our language sample is highly variable in terms of both NLP resources (with languages contributing from 2B up to 713B tokens to mC4) and coverage in neuroscience research,

encompassing both understudied (<10 papers; 8 languages), somewhat studied (>10 but <100 papers; 7 languages), and well-studied languages (>100 papers; 6 languages).

Not surprisingly, we found that languages that are underrepresented in the language technologies tend to be understudied in language neuroscience as well (Spearman's $r = 0.50$, $p = 0.0201$), highlighting a systemic imbalance in the representation of languages across fields.

Language	Study	Tokens (B)	fMRI studies	Language	Study	Tokens (B)	fMRI studies
English	N/A	2733	Well studied	Romanian	I	52	Understudied
French	I	318	Well studied	Lithuanian	I	11	Understudied
Tamil	I	3	Understudied	Hindi	II	24	Somewhat studied
Spanish	I	433	Well studied	Korean	II	26	Well studied
Turkish	I	71	Understudied	Arabic	II	57	Somewhat studied
Vietnamese	I	116	Understudied	Polish	II	130	Somewhat studied
Marathi	I	14	Understudied	Russian	II	713	Somewhat studied
Afrikaans	I	2	Understudied	Portuguese	II	146	Somewhat studied
Dutch	I	73	Well studied	Mandarin	II	39	Well studied
Norwegian	I	27	Somewhat studied	German	II	347	Well studied
Farsi	I	52	Understudied	Italian	II	162	Somewhat studied

Supplementary Table 1: Summary of the resource available for the 21 languages considered across Study I and Study II. The column "Tokens (B)" indicates the number of tokens (billions) in the target language in mC4, a large-scale multilingual corpus. The column "fMRI studies" reports how much the language in question had been studied with fMRI experiments until 2022, as indicated by Malik-Moraleda, Ayyash, et al. (2022) (well studied: >100 papers per language; somewhat studied: >10 but <100 papers; understudied: <10 papers) We also report information for English for reference, even though it was not included in our sample.