Cortical language areas are coupled via a soft hierarchy of model-based linguistic features

Ahmad Samara[1], Zaid Zada[2,3], Tamara Vanderwal[1,4], Uri Hasson[2,3], Samuel A. Nastase[2]

[1] University of British Columbia, Vancouver, BC, Canada
[2] Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA
[3] Department of Psychology, Princeton University, Princeton, NJ, USA
[4] BC Children's Hospital Research Institute, Vancouver, BC, Canada

Corresponding authors: Samuel A. Nastase (snastase@princeton.edu), Tamara Vanderwal (tamara.vanderwal@ubc.ca).

Keywords: encoding models; functional connectivity; language network; large language models (LLMs); narrative comprehension; naturalistic paradigm

Abstract

Natural language comprehension is a complex task that relies on coordinated activity across a network of cortical regions. In this study, we propose that regions of the language network are coupled to one another through subspaces of shared linguistic features. To test this idea, we developed a model-based connectivity framework to quantify stimulus-driven, feature-specific functional connectivity *between* language areas during natural language comprehension. Using fMRI data acquired while subjects listened to spoken narratives, we tested three types of features extracted from a unified neural network model for speech and language: low-level acoustic embeddings, mid-level speech embeddings, and high-level language embeddings. Our modeling framework enabled us to quantify the stimulus features that drive connectivity between regions: early auditory areas were coupled to intermediate language areas via lower-level acoustic and speech features; in contrast, higher-order language and default-mode regions were predominantly coupled through more abstract language features. We observed a clear progression of feature-specific connectivity from early auditory to lateral temporal areas, advancing from acoustic connectivity to speech- and finally to language-driven connectivity. These findings suggest that regions of the language network are coupled through feature-specific communication channels to facilitate efficient and context-sensitive language processing.

Introduction

Language is a fundamental part of everyday human behavior that allows us to communicate complex ideas from one person to another. Our ability to understand language is remarkably efficient given the complexity of the task. As we listen to spoken language, we rapidly convert acoustic signals into words, link words into complex grammatical structures, and integrate all of these patterns into a holistic understanding of the discourse or situation (Christiansen & Chater, 2016). In most cases, we do this effortlessly. Language comprehension emerges from the coordinated activity of a number of different brain areas. A large body of research on the neurobiology of language has identified a highly interconnected network of cortical regions that are selectively engaged during language processing (Hickok & Poeppel, 2007; Friederici, 2011; Price, 2012; Fedorenko et al., 2024). How these language regions coordinate with one another to support efficient language comprehension, however, remains unclear.

Regions of the language network are both structurally interconnected (Catani et al., 2005; Duffau, 2008; Saur et al., 2008; Friederici, 2009; Turken & Dronkers, 2010; Dick et al., 2014) and functionally integrated (Hampson et al., 2002; Lee et al., 2012; Tomasi & Volkow, 2012; Blank et al., 2014; Tie et al., 2014; Zhu et al., 2014; McAvoy et al., 2016; Kong et al., 2021; Du et al., 2024; Salvo et al., 2025). Recent work, for example, has shown that language areas are functionally interconnected even during rest and non-linguistic tasks (Braga et al., 2020; Shain & Fedorenko, 2025). However, traditional within-subject functional connectivity (WSFC) metrics cannot distinguish between intrinsic and extrinsic (i.e., stimulus-driven) co-fluctuations between regions. A between-subject approach is becoming increasingly common in naturalistic neuroscience, whereby intersubject correlation (ISC) analyses isolate stimulus-evoked activity within a brain region (Hasson et al., 2004; Nastase et al., 2019). Following the logic of ISC, intersubject functional connectivity (ISFC) has been used to isolate stimulus-driven connectivity between regions in response to naturalistic stimuli such as spoken narratives (Simony et al., 2016). However, ISFC captures the stimulus-driven components of functional connectivity in a data-driven fashion that is agnostic to the content of the stimulus shared between regions: any features of the stimulus can drive connectivity between any two regions. ISFC can tell us *where* and *how much* connectivity is driven by the stimulus, but not *which* stimulus features are driving the connectivity.

How can we begin to unravel *what* linguistic features are shared across different language regions? A growing body of recent work indicates that different kinds of linguistic structures (e.g., syntax and semantics) appear to be co-localized across language areas (Fedorenko et al., 2012, 2016; 2020; Wehbe et al., 2014; Blank et al., 2016; Nelson et al., 2017; Caucheteux et al., 2021; Reddy & Wehbe, 2021; Toneva et al., 2022; Kumar, Sumers et al., 2024; Shain et al., 2024), and most language areas display similar response profiles to a variety of language tasks (Fedorenko et al., 2024). These observations create a certain tension: surely, different language areas contribute differently to the overall circuit, but why do we observe such overwhelming functional similarity across regions? From a computational perspective, one possible explanation for this tension is that different areas of the language network may interact through a “communication subspace” of shared features (Semedo et al., 2019). For example, in visual areas, one region is coupled to another via fluctuations along a subset of dimensions in the population-level activity space (Semedo et al., 2019; Kohn et al., 2020;

MacDowell et al., 2025). Perhaps different regions of the language network rely on a shared, multidimensional embedding space of linguistic features to efficiently coordinate their contributions to the network as a whole. Interestingly, modern large language models (LLMs) appear to rely on a similar geometric mechanism of connectivity: the circuits across layers of an LLM interact with one another by gradually refining linguistic representations as they proceed through a high-dimensional embedding space shared across layers (Elhage et al., 2021). This so-called "residual stream" serves as a communication channel across layers and is thought to be critical to the capacity of LLMs to capture the rich, content-specific structure of natural language.

In this study, we hypothesized that different regions of the language network are coupled to one another via a multidimensional space of linguistic features. This hypothesis is inspired by modern neural network models for speech and language, which integrate phonemic, syntactic, and semantic features into a unified neural population code (e.g., Radford et al., 2023), and construct increasingly refined representations by modifying a high-dimensional embedding space that is shared from layer to layer (Elhage et al., 2021). In a similar way, we hypothesize that two regions of the language network may harmonize their contributions to language comprehension via a shared subspace of linguistic features. Our hypothesis yields two predictions. First, functional connectivity between one language region and another should result from moment-to-moment, stimulus-driven covariation along a shared subset of linguistic features. Second, as we proceed along the cortical processing hierarchy, connectivity should be driven by increasingly abstract linguistic features. Given the mixed selectivity (or "polysemanticity") of neural population codes (Fusi et al., 2016; Elhage et al., 2022; Bricken et al., 2023), we expect some low- and mid-level features to be retained even as other features become increasingly complex; we refer to this as a "soft hierarchy".

To test these predictions, we developed a novel model-based framework for quantifying stimulus-driven, feature-specific co-fluctuations in neural activity between one region and another. This model-driven framework provides a theoretical advance over content-agnostic metrics of functional connectivity (both WSFC and ISFC) by allowing us to test explicit, feature-specific models of the functional connectivity between language areas. To this end, we decomposed two spoken stories into low-level acoustic, mid-level speech, and high-level language features based on embeddings extracted from the Whisper speech and language model (Radford et al., 2023). We tested these model features against naturalistic fMRI data acquired while subjects listened to the same spoken stories. Our findings reveal a soft processing hierarchy where language areas are coupled along shared acoustic, speech, and language features, and connectivity from lower- to higher-order areas is driven by progressively more refined linguistic features.

Results

To investigate how regions of the language network coordinate their activity during naturalistic language comprehension, we developed a model-based framework that quantifies stimulus-driven, feature-specific functional connectivity between brain areas. Before reporting our core results, we first develop the theoretical motivation for our approach and validate our models within each brain region.

ISC analysis (Hasson et al., 2004; Nastase et al., 2019) has been used, initially, to isolate the stimulus-driven component of neural activity within a given brain region (**Fig. 1a**). In separate subjects (scanned at separate times), the stimulus is the only shared factor that could drive shared activity. However, ISC analysis is data-driven and content-agnostic: it can tell us *where* and *how much* activity is driven by the stimulus, but it cannot determine *what* stimulus features drive neural activity (Zada et al., 2024). To quantify *what* stimulus features are driving activity in a given brain region, we use parcel-wise encoding models to quantify which explicit linguistic features are encoded in brain activity (Wu et al., 2006; Nalesaris et al., 2011; Huth et al., 2016; de Heer et al., 2017; Dupré la Tour, Visconti di Oleggio Castello et al., 2025; **Fig. 1b**).

To quantify the linguistic features of the stimulus, we utilized a state-of-the-art transformer-based neural network model for speech recognition, known as Whisper (Radford et al., 2023). We elected to use Whisper because it is a unified acoustic-to-speech-to-language model that learns how to map the acoustic signals of natural speech into language representations useful for natural language tasks like next-word prediction and transcription (**Fig. 1c**). The “encoder” component of the model learns to extract linguistic features from acoustic inputs, whereas the “decoder” component extracts linguistic features from text inputs. We extracted three types of linguistic features from Whisper: (1) low-level “acoustic” embeddings from the input to the transformer stack of the encoder: (2) mid-level “speech” embeddings from the final layer of the encoder; and (3) high-level, more abstract “language” embeddings from a late-intermediate layer of the decoder. These three sets of embeddings capture increasingly abstract and contextual features of spoken language that the model uses to perform natural language tasks (Goldstein et al., 2025). All three embeddings were of the same dimensionality (1024 dimensions).

We evaluated our models against fMRI data comprising N = 46 participants, each of whom listened to two different ~13-minute spoken stories (Nastase et al., 2021). To reduce computational demands, we first reduced the voxel-level time series into 1,000 parcel-level time series derived from a functional atlas (Schaefer et al., 2018). We used banded ridge regression to fit parcel-wise encoding models jointly across all three sets of features (acoustic, speech, and language embeddings; Dupré la Tour et al., 2022). This allows all three types of linguistic features to compete for variance in the neural activity fairly. To test the alignment of linguistic features with human brain activity, we generated predictions from each of the three feature bands for the left-out data and computed the correlation between the model-predicted and actual brain activity at each parcel. We also computed the joint model performance, which corresponds to the sum of the predictions for each of the three feature bands. To more closely match the formulation of ISC, we estimated encoding models within each subject and evaluated their performance across subjects by computing the correlation of model-based predictions from one subject with the average actual time series across all other subjects. While this intersubject encoding approach differs from much prior work (e.g., Huth et al., 2016; Schrimpf et al., 2021; Goldstein et al., 2022, 2025; cf. Van Uden, Nastase et al., 2018; Nastase et al., 2020b; Zada et al., 2024), the focus of our analyses is ultimately on what features are shared between regions (not between subjects). To further ensure the generalizability of our findings, we estimated all models within one story and evaluated their performance in predicting the other story (and vice versa; **Fig. S1**).

Overall, this modeling framework allows us to quantify how well the model features linearly align with human neural activity during complex, naturalistic language comprehension.

Modeling the soft hierarchy of linguistic features across the language network

We first localized language areas involved in processing the story stimuli using a conventional ISC analysis, which identifies parcels where neural activity is synchronized to the stimulus (**Fig. 1a**). ISC analysis revealed a large-scale cortical network for spoken narrative comprehension comprising low-level auditory areas, language areas, and higher-level default-mode areas (**Fig. 1d**) (Nastase et al., 2021). Next, we evaluated the joint encoding model performance combined across all three feature bands from the Whisper model using our intersubject encoding approach (**Fig. 1b**). The resulting joint model predicted neural activity across much of the language network, including parcels in frontotemporal language regions, as well as in posterior medial cortex (PMC), superior frontal language area (SFL), dorsomedial prefrontal cortex (dmPFC; **Fig. 1e,g**). The joint model performance map captures a subset of regions identified by the ISC analysis, suggesting that ISC identifies certain stimulus features of spoken stories that are not represented in the embeddings extracted from the Whisper model.

We then evaluated predictions derived separately from each of the three feature bands. By visualizing the top-performing feature at every parcel (**Fig. 1f**), we identified a coarse processing hierarchy where acoustic features dominate in superior temporal auditory areas, speech features are more sparsely distributed along temporal cortex and higher-level areas, and the language features dominate in lateral temporal areas (both anterior and posterior) and inferior frontal gyrus (IFG). The acoustic model performance was largely confined to the early auditory cortex (EAC) and superior temporal gyrus (STG), with punctate clusters in middle frontal gyrus (MFG; **Fig. 1g**). Speech model performance extended from EAC and STG anterolaterally to superior temporal sulcus (STS), and included portions of right IFG (**Fig. 1g**). Language embeddings predicted a broader array of language regions, including parcels in both anterior and posterior lateral temporal cortex, IFG, as well as PMC, SFL, dmPFC (**Fig. 1g**). These results are generally consistent with prior work using encoding models to map linguistic features onto cortical activity (e.g., Wehbe et al., 2014; Huth et al., 2016; de Heer et al., 2017; Goldstein et al., 2025). In line with the notion of a soft hierarchy, we found significant overlap among the feature-specific model performance maps, suggesting that many cortical areas encode mixed acoustic, speech, and language features (**Fig. 1g**).

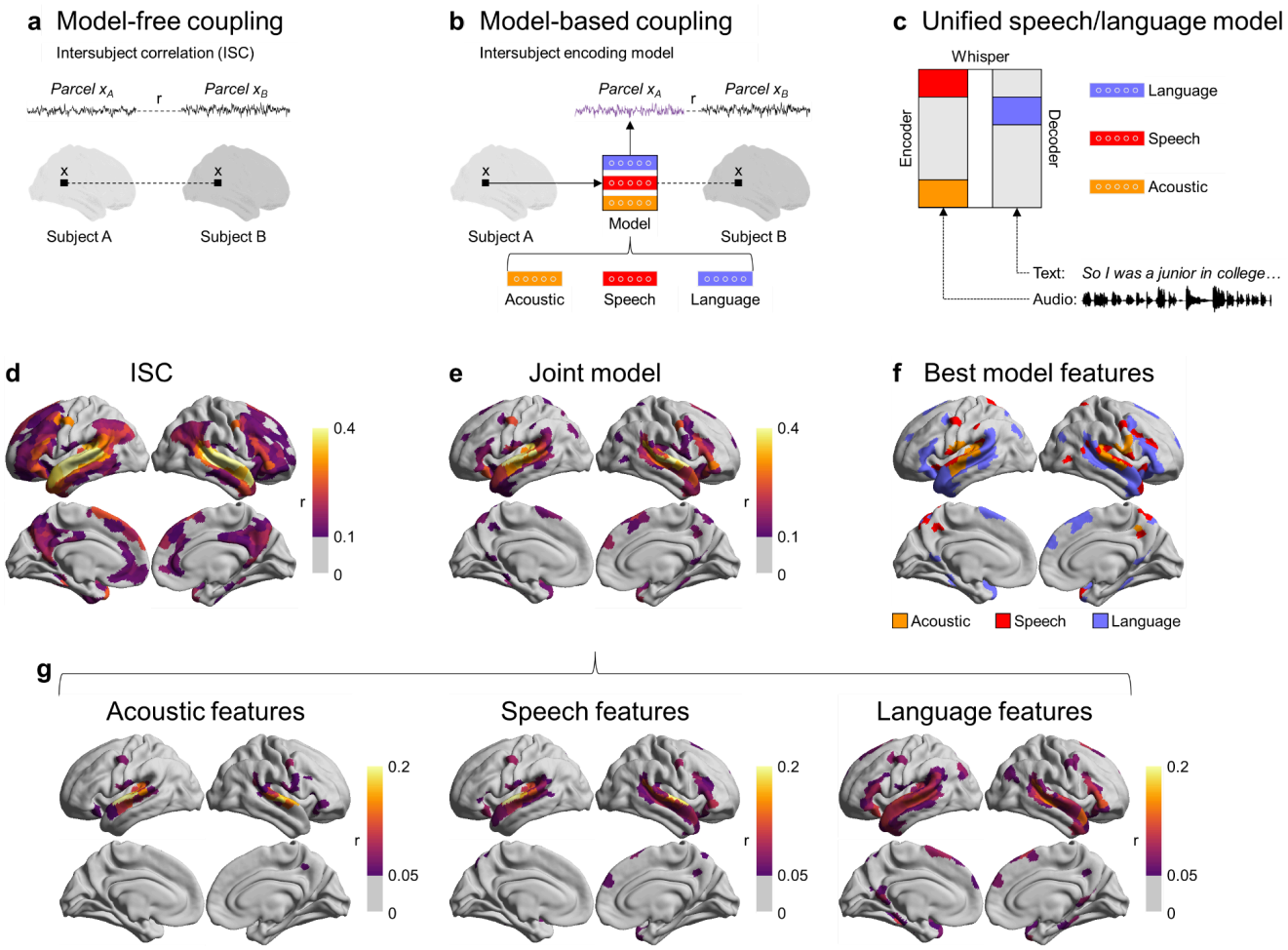

Fig. 1. Modeling stimulus-driven, feature-specific neural activity during natural language comprehension. **(a)** Intersubject correlation is a data-driven, model-free method for quantifying the stimulus-driven component of neural activity *within* a given region. **(b)** To quantify the activity driven by specific stimulus features, we construct parcel-wise encoding models using explicit linguistic features extracted from a computational model. Encoding models are evaluated within a given brain region by correlating model-predicted activity with the actual activity in a left-out subject or group of subjects (in the same way as ISC analysis). **(c)** We extract three types of linguistic features from a unified transformer-based speech and language model called Whisper: acoustic (orange), speech (red), and language (blue) feature representations (i.e., embeddings) of the stimulus. The schematic shows a more detailed version of the model depicted in **(b)**. For more details on the encoding model approach, refer to **Fig. S1**. Dashed lines indicate correlation; solid lines indicate model input/output. **(d)** We computed ISC within 1,000 cortical parcels for two story-listening fMRI datasets. Parcels with significant ISC ($p < .05$) were further thresholded at 0.10 for visualization. Encoding models were estimated across all three feature bands using banded ridge regression. **(e)** Performance for the joint model (i.e., encoding model fit jointly using all three feature bands) and thresholded in the same way **(e)**. **(f)** The best-performing feature band was identified for each parcel. **(g)** The joint model performance can be decomposed into feature-specific performance values.

Modeling stimulus-driven, feature-specific connectivity between language areas

Next we turned to the core question of this manuscript: How can we quantify the stimulus-derived linguistic features driving connectivity *between regions* of the language network? First, by extending the same logic as ISC, we can use ISFC analysis to quantify the stimulus-driven *connectivity* between

brain areas, effectively filtering out idiosyncratic noise and the intrinsic fluctuations that play a large role in traditional WSFC (Simony et al., 2016; **Fig. 2a**). ISFC analysis yields a parcel-by-parcel matrix of stimulus-driven connectivity values between pairs of brain regions. The diagonal of this matrix corresponds to the within-region ISC values (**Fig. 1a**). Similar to both ISC analysis and WSFC analysis, ISFC analysis is a data-driven method that indicates *where* and *how much* stimulus-driven connectivity exists between two regions; however, it does not reveal *what* features of the stimulus drive that connectivity.

To quantify *what* stimulus features are shared between different brain regions, we again used explicit linguistic embeddings extracted from the Whisper model. We use the same parcel-wise encoding models trained within each parcel from the previous section. In our model-based connectivity analysis, we now evaluate these models in terms of how well their predictions generalize to *other parcels* in the same fashion as the ISFC analysis. We generate model-based predictions for one parcel, then correlate these predicted time series with the average actual time series of the remaining subjects across all pairs of parcels (**Fig. 2b**). This model-based functional connectivity analysis results in a parcel-by-parcel matrix of feature-specific connectivity values between pairs of parcels (Toneva et al., 2022; Meschke et al., 2023; Zada et al., 2024). In this case, the diagonal of the connectivity matrix corresponds to the within-parcel intersubject encoding model performance (**Fig. 1b**). This analysis effectively filters the connectivity between regions based on what can be captured by an explicit feature space.

We first visualized the conventional ISFC matrix (**Fig. 2c**) and the joint model-based connectivity matrix based on the combination of all three sets of linguistic features (**Fig. 2d**). The joint model connectivity matrix appears to recapitulate some but not all of the stimulus-driven connectivity structure in the ISFC matrix, albeit with lower correlation values. We also constructed model-based connectivity matrices based on predictions derived from each of the three types of linguistic features. When examining the best-performing feature band at each connection, we see relatively focal connectivity for the acoustic embeddings, sparse connectivity for the speech embeddings, and more widespread connectivity for the language embeddings (**Fig. 2e**). The acoustic model best captured shared connectivity within EAC and between EAC and STG/S parcels. In contrast, the language model outperformed the other two models in capturing shared connectivity within and between STG/S and IFG/S parcels. Model performance was comparatively lower in the temporoparietal junction (TPJ) and PMC regions, with mixed results across models. Overlapping patterns of connectivity across matrices for the three types of stimulus features suggest that many edges may be captured by overlapping sets of features (**Fig. 2f**). To simplify visualizing the connectivity matrices, we focused on a subset of 280 cortical parcels spanning 23 regions of interest (**Fig. 2g**) associated with language comprehension based on the ISC results (**Fig. 1d**). These model-based functional connectivity matrices serve as the basis for all of the subsequent analyses in the manuscript.

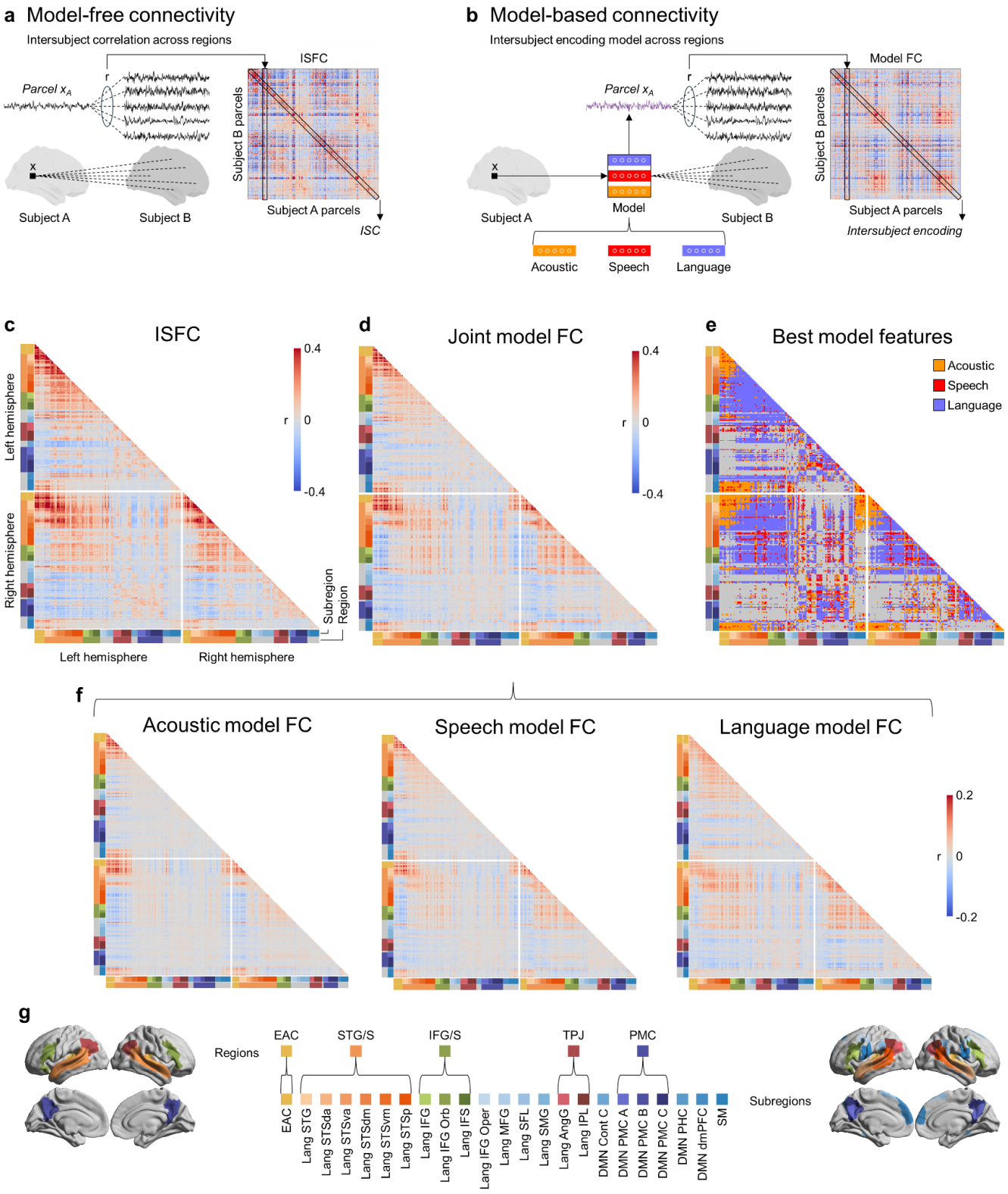

Fig. 2. Modeling stimulus-driven, feature-specific functional connectivity during natural language comprehension. (a) Following the logic of ISC, intersubject functional connectivity (ISFC) quantifies the stimulus-driven connectivity between brain areas in a data-driven, model-free fashion. The diagonal of the ISFC matrix corresponds to the within-parcel ISC values. **(b)** To quantify feature-specific model-based connectivity, we evaluate encoding models based on the three sets of linguistic features extracted from Whisper *across* pairs

of parcels. The diagonal of the resulting model-based connectivity matrix corresponds to the within-parcel intersubject encoding performance values. Dashed lines indicate correlation; solid lines going into and out of the model indicate model input/output. **(c)** A parcel-by-parcel ISFC matrix was computed for all pairs of parcels within language areas. The joint model-based connectivity matrix **(d)**, as well as feature-specific model connectivity (model FC) matrices **(f)** were computed in the same way as the ISFC matrix for each subject and were averaged across subjects for visualization. Feature-specific ISFC matrices were computed based on acoustic, speech, and language embeddings extracted from the Whisper model. **(e)** The best-performing model at each edge was chosen as the model with the correlation value closest to the original ISFC value at that edge. The best model map is thresholded to include only edges that are positive in the original ISFC matrix. **(g)** All connectivity matrices are visualized for a set of 280 parcels spanning functionally defined language regions (see *Regions of Interest* section under *Methods* for details on how these regions were defined).

Language regions are coupled via distinct and overlapping subspaces of linguistic features

For a more quantitative assessment of the encoding space underlying the shared network structure during language comprehension, the original ISFC matrices served as an index of the reliable, stimulus-driven connectivity between cortical areas. In contrast, the model connectivity matrices allowed us to isolate elements of this connectivity driven by specific stimulus features. We quantified the feature-specific model-based connectivity values within and between eight regions of interest ranging from low- to high-level areas for spoken narrative comprehension: EAC, STG/S, IFG/S, TPJ angular gyrus (TPJ AngG), TPJ inferior parietal lobule (TPJ IPL), PCM A, PMC B, and PMC C (**Fig. 3a**). We first examined the average local connectivity across all edges connecting parcels *within* a given region (**Fig. 3b**). We found that model connectivity among parcels within EAC was best predicted by acoustic (A) features, followed by speech (S) features, then language (L) features (A vs. S: $t = 5.40$, $p_{FDR} < .001$; S vs. L: $t = 7.70$, $p_{FDR} < .001$; A vs. L: $t = 9.39$, $p_{FDR} < .001$). The opposite was true for connectivity among parcels within STG (A vs. S: $t = -10.84$, $p_{FDR} < .001$; S vs. L: $t = -9.09$, $p_{FDR} < .001$; A vs. L: $t = -15.15$, $p_{FDR} < .001$), IFG (A vs. S: $t = -11.54$, $p_{FDR} < .001$; S vs. L: $t = -7.27$, $p_{FDR} < .001$; A vs. L: $t = -11.39$, $p_{FDR} < .001$), and appears to be the case in several other higher-level areas for language and narrative comprehension: the language embeddings captured the largest proportion of connectivity, followed by the speech embeddings, then the acoustic embeddings. This suggests that nearby parcels within a given language region tend to co-fluctuate along a shared set of linguistic features. Early auditory areas are most strongly coupled along acoustic and speech features. In contrast, parcels within higher-level areas, including language areas and default-mode areas, are coupled according to higher-level language features.

We next examined feature-specific model-based connectivity *across* different regions of the language network (**Fig. 3c**). We found that the acoustic embedding best captured the connectivity between EAC and STG/S regions (A vs. S: $t = 7.87$, $p_{FDR} < .001$; S vs. L: $t = 1.40$, $p_{FDR} = 0.179$; A vs. L: $t = 7.16$, $p_{FDR} < .001$). The edges connecting STG/S and IFG, on the other hand, were dominated by the language embeddings, with speech embeddings capturing a large but secondary portion of variance (A vs. S: $t = -17.47$, $p_{FDR} < .001$; S vs. L: $t = -10.59$, $p_{FDR} < .001$; A vs. L: $t = -19.25$, $p_{FDR} < .001$). We observed a similar profile for the edges connecting STG/S and PMC B (A vs. S: $t = -4.72$, $p_{FDR} < .001$; S vs. L: $t = -3.77$, $p_{FDR} < .001$; A vs. L: $t = -6.03$, $p_{FDR} < .001$). IFG/S was also coupled with PMC B, but similarly along all three sets of features (A vs. S: $t = -4.57$, $p_{FDR} < .001$; S vs. L: $t = 0.53$, $p_{FDR} = 0.599$; A vs. L: $t = -1.82$, $p_{FDR} = .086$). The language embeddings captured the most connectivity between default mode

areas, such as TPJ and PMC C (A vs. S: t = -4.05, $p_{FDR} < .001$; S vs. L: t = -7.26, $p_{FDR} < .001$; A vs. L: t = -7.40, $p_{FDR} < .001$). We also observed negative model connectivity values between the EAC and IFG/S, which may be due to systematic differences in the time course of activity between these regions. In some cases, for example, between neighboring PMC B and PMC C in DMN, model connectivity was negligible, despite markedly stronger ISFC. Overall, these results suggest that edges connecting different language areas, including discontiguous and relatively distant cortical regions, are coupled along a subset of shared linguistic features. In line with a soft hierarchy, many connections appear to be driven by overlapping sets of linguistic features. For example, connectivity among STG/S regions is driven by acoustic, speech, and linguistic features (**Fig. 3b**). Furthermore, there are clear trends in which lower-level areas are coupled along acoustic features, and higher-level areas are coupled along language (and to a lesser extent, speech) features.

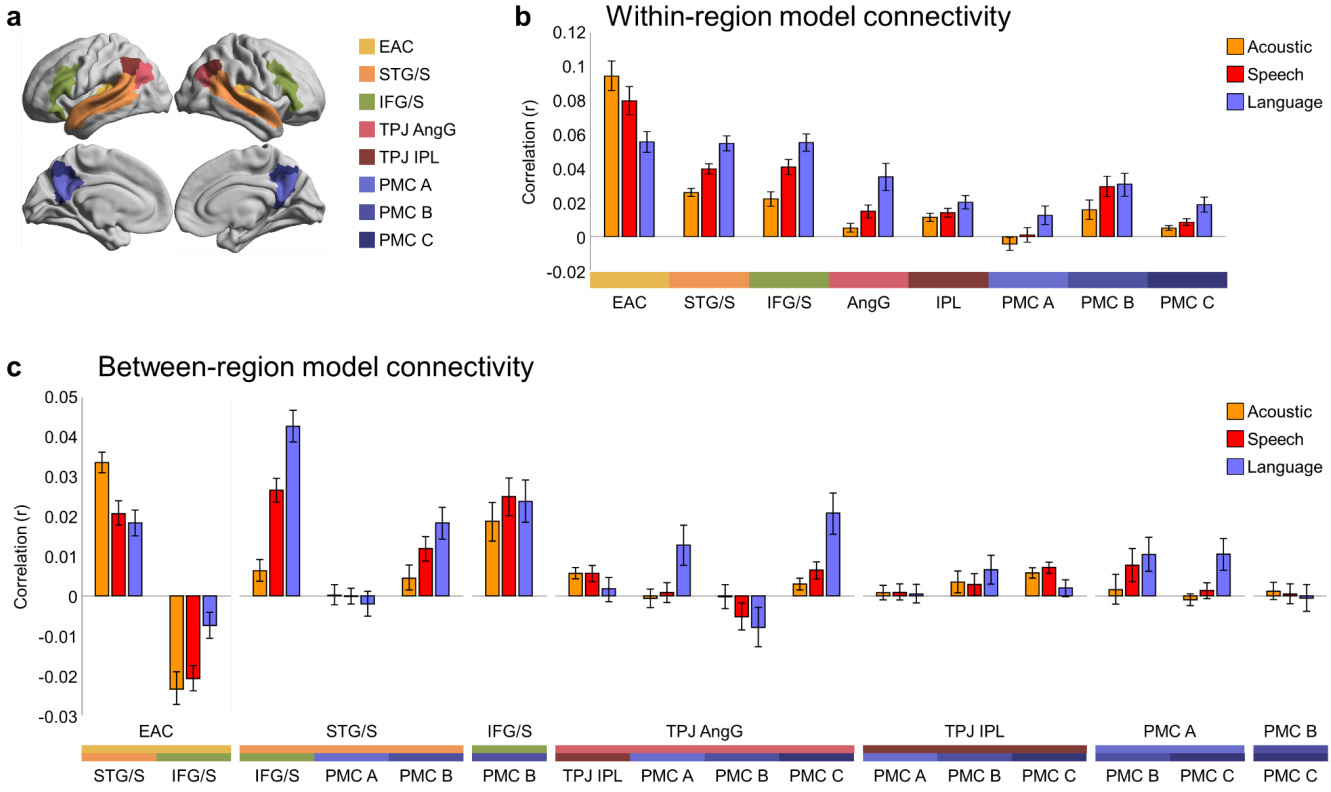

Fig. 3. Feature-specific model-based connectivity within and between language regions. (a) We focused on eight language-related regions (comprising 207 parcels) that previously showed strong ISC during story listening: early auditory cortex (EAC), superior temporal gyrus and sulcus (STG/S), inferior frontal gyrus and sulcus (IFG/S), temporoparietal junction angular gyrus and inferior parietal lobule (TPJ AngG and TPJ IPL), and posterior medial cortex A, B, and C (PMC A, PMC B, and PMC C). **(b)** Feature-specific model connectivity values were averaged across parcel pairs within each language region. **(c)** Feature-specific model connectivity values were averaged across parcel pairs connecting different language regions. Between-region model connectivity results are shown only for region pairs with positive ISFC values. Error bars indicate bootstrap 95% confidence intervals. See **Fig. S2** for the same results plotted against the ISFC and joint model performance values.

To better visualize the cortical extent of feature-specific connectivity between cortical language regions, we created seed-based model connectivity maps of EAC, STG/S, and IFG/S. For each seed, we computed model connectivity for each feature band across all parcels and identified parcels with significant connectivity (using a one-sample t-test across subjects). This analysis revealed overlapping maps of model connectivity captured by each of the different feature bands for each seed (**Fig. 4**). Models trained in EAC yielded significant feature-specific model connectivity maps constrained to perisylvian areas, with the acoustic embeddings driving the marginally wider-spread connectivity. For models trained in STG, connectivity based on acoustic embeddings was tightly localized to the middle STG, whereas speech embeddings yielded connectivity extending along the STG/S, and language embeddings expanded connectivity in the STG/S and IFG/S. For models trained in IFG/S, only weak encoding was found for acoustic embeddings, whereas the speech and language embeddings yielded increasingly widespread connectivity in frontotemporal language areas. This suggests that language areas are coupled through multiple, overlapping sets of linguistic features. However, higher-level linguistic features (speech and language embeddings) are increasingly dominant in linking higher-order regions.

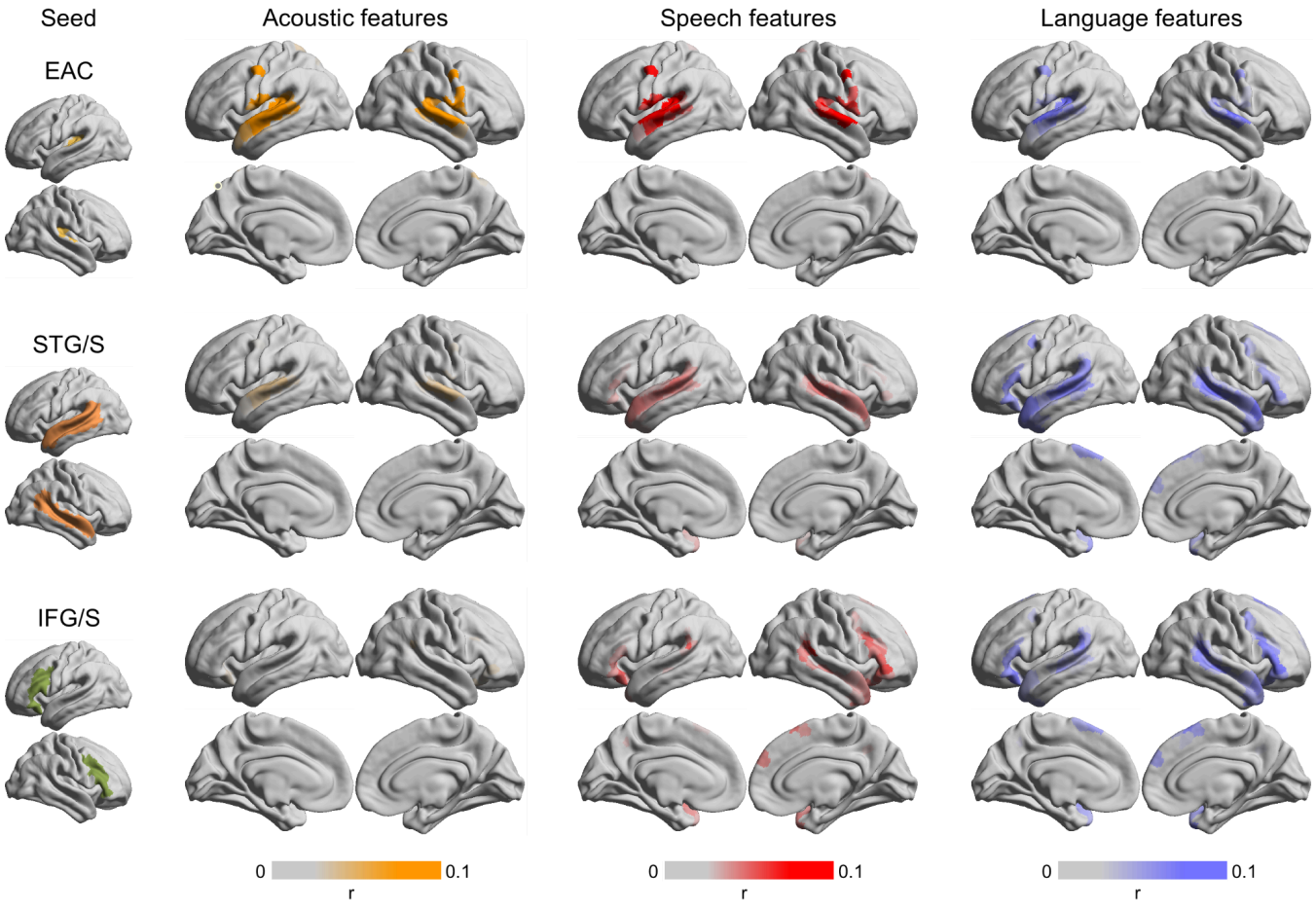

Fig. 4. Seed-based feature-specific connectivity maps. Seed-based model connectivity maps were computed separately for the three feature bands in three seed regions: early auditory cortex (EAC), superior temporal gyrus and sulcus (STG/S), and inferior frontal gyrus and sulcus (IFG/S). For each feature band and seed, we first computed the parcel-pair connectivity between parcels within the seed and all parcels with significant encoding

based on the joint model (i.e., the same regions as in **Fig. 1e**). Then, we averaged the resulting connectivity values across the seed parcels.

Transition from acoustic- to language-based connectivity in the temporal cortex

We next zoomed in on a subset of regions for speech comprehension for which we have clear hypotheses about which linguistic features are shared across regions. We hypothesized that EAC would be coupled to neighboring STG regions primarily according to acoustic and speech features, while in more lateral STG/S regions, acoustic features of connectivity would recede and be overtaken by language features. We examined the feature-specific model connectivity between seven smaller regions along three pathways extending from the EAC/STG towards the STSva, STSvm, and STSp (**Fig. 4**).

We found that the model connectivity, calculated by averaging the connectivity values across all edges linking a pair of regions, progressed systematically across feature bands as we moved further along these pathways. Specifically, model connectivity between EAC and STG was most strongly driven by acoustic features, followed closely by speech features, and then language features (A vs. S: $t = 6.69$, $p_{FDR} < .001$; S vs. L: $t = 10.00$, $p_{FDR} < .001$; A vs. L: $t = 12.66$, $p_{FDR} < .001$). At the transition from STG to STSda and STSdm, acoustic connectivity relatively decreased, and all three feature bands contributed more similarly to connectivity, though some comparisons remained significant (A vs. S for STG-STSda: $t = 4.95$, $p_{FDR} < .001$; S vs. L for STG-STSda: $t = -2.87$, $p_{FDR} = .007$; A vs. L for STG-STSda: $t = 1.63$, $p_{FDR} = 0.119$; A vs. S for STG-STSdm: $t = 1.49$, $p_{FDR} = 0.143$; S vs. L for STG-STSdm: $t = -5.13$, $p_{FDR} < .001$; A vs. L for STG-STSdm: $t = -3.32$, $p_{FDR} = .002$). Finally, as we moved from STSda and STSdm to STSva and STSvm, the pattern fully reversed, such that the language features captured the largest portion of connectivity, followed by the speech embeddings, then the acoustic embeddings (A vs. S for STSda-STSva: $t = -9.10$, $p_{FDR} < .001$; S vs. L for STSda-STSva: $t = -5.15$, $p_{FDR} < .001$; A vs. L for STSda-STSva: $t = -9.79$, $p_{FDR} < .001$; A vs. S for STSdm-STSvm: $t = -14.40$, $p_{FDR} < .001$; S vs. L for STSdm-STSvm: $t = -12.63$, $p_{FDR} < .001$; A vs. L for STSdm-STSvm: $t = -19.41$, $p_{FDR} < .001$). These findings demonstrate a clear transition in the subspace of linguistic features linking EAC to increasingly lateral language areas in STG/S from predominantly acoustic to predominantly higher-level language features. In all cases, however, the model-based connectivity did not fully capture stimulus-driven connectivity quantified using ISFC (**Fig. S3**).

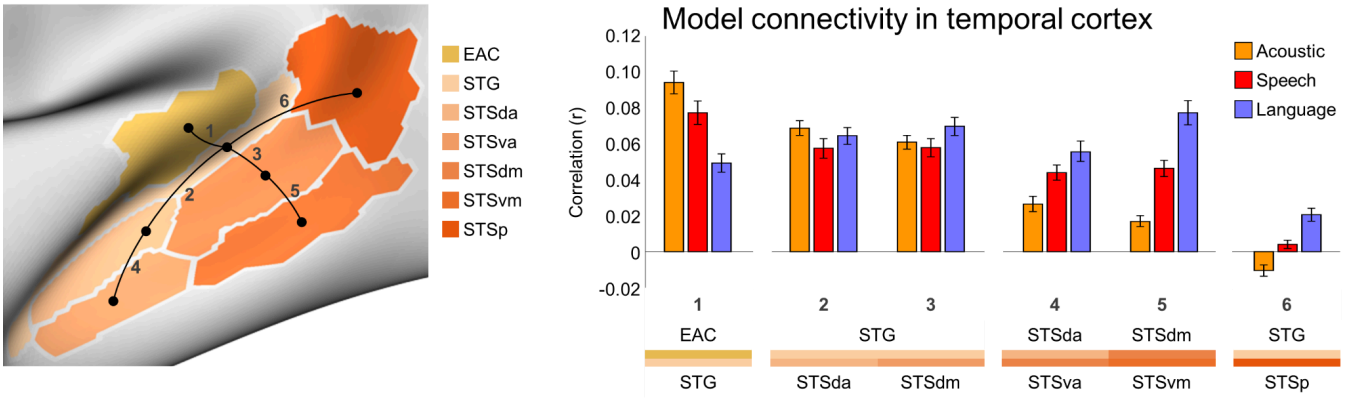

Fig. 5. Model connectivity along superior temporal pathways. Feature-specific model connectivity was computed across parcel pairs along pathways linking early auditory cortex (EAC), superior temporal gyrus (STG), and superior temporal sulcus (STS) spanning the following regions: EAC, STG, dorsal anterior, ventral anterior, dorsal mid, ventral mid, and posterior superior temporal sulcus (STSda, STSva, STSdm, STSvm, STSp). Model connectivity showed progressive transition across feature bands as we moved further along these pathways: EAC-STG connectivity was most strongly driven by acoustic features, followed by speech features, and then language features; connectivity of STG to STSda and STSdm relatively decreased for the acoustic feature band, and all three feature bands contributed more similarly; and the pattern of connectivity from STSda to STSva and STSdm to STSvm was reversed, where language features captured the largest portion of connectivity, followed by the speech embeddings, then the acoustic embeddings. Error bars indicate bootstrap 95% confidence intervals.

Model-based connectivity recapitulates large-scale cortical network configuration

To assess how well model connectivity captures larger-scale patterns of connectivity, we systematically examined the correlation between the connectivity patterns from feature-specific model connectivity matrices and the corresponding connectivity patterns quantified using ISFC (**Fig. 5**). This analysis was performed at three spatial scales. First, when considering the whole language network, we found that the language embeddings better captured the ISFCs, followed by speech embeddings, then acoustic embeddings (A vs. S: $t = -10.16$, $p_{FDR} < .001$; S vs. L: $t = -6.78$, $p_{FDR} < .001$; A vs. L: $t = -12.19$, $p_{FDR} < .001$) (**Fig. 5a**). As a control, we examined the correlation between model connectivity patterns and ISFC for all parcel pairs across the rest of the cortex (excluding language areas), which resulted in much lower correlations (A in language network vs. rest of brain: $t = 27.32$, $p_{FDR} < .001$; S in language network vs. rest of brain: $t = 31.47$, $p_{FDR} < .001$; L in language network vs. rest of brain: $t = 35.62$, $p_{FDR} < .001$).

Second, we examined the correlation between model connectivity and ISFC for the connectivity profiles of EAC, language (STG/S and IFG/S), and default-mode (TPJ and PMC) regions to all language network regions (**Fig. 5b**). Model connectivity for the auditory features was most similar to the ISFC profile for EAC (A vs. S for EAC regions: $t = 2.00$, $p_{FDR} = .054$; S vs. L for EAC regions: $t = 4.44$, $p_{FDR} < .001$; A vs. L for EAC regions: $t = 6.21$, $p_{FDR} < .001$), whereas model connectivity for language features was most similar to ISFC for language and default-mode profiles (A vs. S for language regions: $t = -10.89$, $p_{FDR} < .001$; S vs. L for language regions: $t = -6.96$, $p_{FDR} < .001$; A vs. L for language regions: t

= -13.69, $p_{FDR} < .001$; A vs. S for default-mode regions: t = -10.64, $p_{FDR} < .001$; S vs. L for default-mode regions: t = -7.95, $p_{FDR} < .001$; A vs. L for default-mode regions: t = -12.74, $p_{FDR} < .001$).

Third, we focused on connectivity patterns between the three groups of regions involved in spoken narrative comprehension (**Fig. 5c**). The acoustic and speech features best recapitulated ISFC connectivity between EAC and downstream frontotemporal language areas (A vs. S: t = 2.45, p_{FDR} = .020; S vs. L: t = 7.62, $p_{FDR} < .001$; A vs. L: t = 8.82, $p_{FDR} < .001$), whereas the language features (followed by the speech features) best recapitulated ISFCs between language and default-mode areas (A vs. S: t = -13.30, $p_{FDR} < .001$; S vs. L: t = -7.48, $p_{FDR} < .001$; A vs. L: t = -14.71, $p_{FDR} < .001$). Interestingly, the correspondence between model connectivity and ISFC was fairly similar for acoustic and language features, with speech features slightly (but significantly) outperforming the other features for the connections between EAC and default-mode areas (A vs. S: t = -4.86, $p_{FDR} < .001$; S vs. L: t = 3.11, p_{FDR} = .004; A vs. L: t = -1.03, p_{FDR} = 0.311). These regions may be weakly coupled, or they may be indirectly coupled through the intermediate language areas.

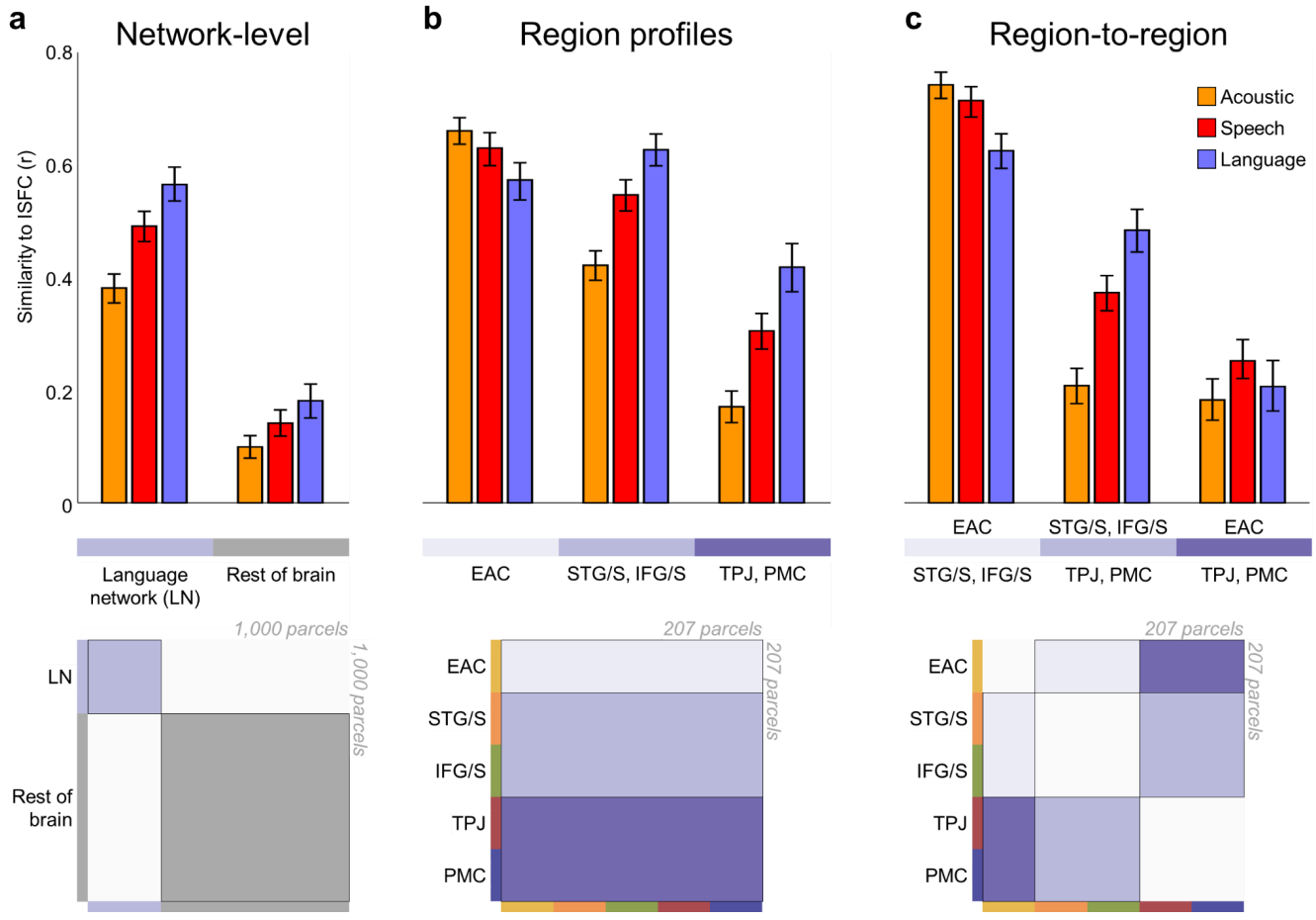

Fig. 6. The feature-specific model connectivity recapitulates the intersubject functional connectivity network configuration. The similarity between model connectivity and intersubject functional connectivity (ISFC) was quantified as the correlation between vectorized model connectivity and ISFC (sub)matrices. The highlighted (purple) parts in the schematic parcel-by-parcel connectivity matrices at the bottom indicate the

edges contributing to each group of bars. **(a)** Similarity between patterns of feature-specific model connectivity and ISFC was computed within the language network level for acoustic, speech, and language embeddings. As a control, similarity between model connectivity and ISFC was computed within the rest of the cortex. **(b)** Similarity between model connectivity and ISFC was computed for three broad groups of regions associated with spoken narrative comprehension across their respective connectivity profiles: early auditory cortex (EAC), language (i.e., superior temporal gyrus and sulcus [STG/S] and inferior frontal gyrus and sulcus [IFG/S], and default-mode (i.e., TPJ and posterior medial cortex [PMC]) regions. **(c)** Finally, we examined the similarity of model connectivity and ISFC between groups of low-, mid, and high-level regions associated with spoken narrative comprehension. Error bars indicate bootstrap 95% confidence intervals.

Discussion

This study provides evidence that cortical language regions are coupled through distinct and overlapping subspaces of linguistic features. By using a model-based functional connectivity framework, we demonstrated that acoustic, speech, and language embeddings derived from a unified speech and language model can predict connectivity between regions. Our findings revealed that different subsets of linguistic features capture distinct aspects of overall network connectivity. Acoustic features, and to a lesser extent speech features, drive connectivity between early auditory and intermediate language areas. In contrast, language features, and secondarily speech features, drive connectivity between language areas and higher-level default-mode areas. When examining more anatomically specific connectivity in superior temporal cortex, we observed a systematic transition from acoustic to language features linking neighboring temporal regions, corresponding to a processing hierarchy progressing from perceptual to more abstract features of spoken language. Overall, these results reveal a soft processing hierarchy: regions of the language network are coupled via a mixture of acoustic, speech, and language features, but increasingly abstract linguistic features drive connectivity among higher-order cortical regions.

Over the last two decades, ISC analysis (Hasson et al., 2004; Nastase et al., 2019) has advanced the field of neuroimaging by enabling researchers to map stimulus-driven neural responses to naturalistic stimuli, like spoken language, even in the absence of an explicit model of such complex stimuli (e.g., Stephens et al., 2010; Lerner et al., 2011). In recent years, both methodological and computational advances have enabled the testing of more meaningful, neuroscientifically relevant feature representations of complex naturalistic stimuli at scale (Nastase et al., 2020a), particularly in the language domain. First, voxelwise encoding models provide a powerful framework for isolating *feature-specific* components of stimulus-driven brain activity in naturalistic paradigms (Wu et al., 2006; Naselaris et al., 2011; Wehbe et al., 2014; Huth et al., 2016, Dupré la Tour, Visconti di Oleggio Castello et al., 2025). Second, with recent advances in deep learning (e.g., Brown et al., 2020; Radford et al., 2023), we now have explicit, computational models that can accommodate the richness of real-world speech and language. Combining these advances provides a qualitative leap beyond model-free, data-driven methods like ISC, allowing researchers to test explicit models of the neural activity supporting speech, language, and communication in completely natural contexts (Schrimpf et al., 2021; Caucheteux et al., 2022; Goldstein et al., 2022, 2025; Tuckute et al., 2024; Zada et al., 2024).

While ISC captures stimulus-driven neural activity within a given brain region, ISFC provides a theoretical extension that enables us to quantify stimulus-driven functional connectivity *between* regions (Simony et al., 2016). Following the same logic, we evaluated our encoding models *across regions* to identify *feature-specific* functional connectivity (Toneva et al., 2022). This model-based connectivity framework allowed us to quantify which features of speech and language drive the co-fluctuations in neural activity between regions of the language network. In the current work, we used both ISC and ISFC to quantify the upper limit of reliable, time-locked, stimulus-driven variability in the connectivity between regions (or activity within regions; i.e., a "noise ceiling"), then use encoding models to quantify how much of this stimulus-driven connectivity can be captured by specific features from a neural network model for speech. This approach is conceptually similar to the model connectivity method introduced by Meschke et al. (2023), which quantifies the similarity of encoding model weights across brain regions.

Our findings suggest that different areas of the language network are coupled to one another via a multidimensional space of shared linguistic features. This geometric notion of functional connectivity is conceptually similar to the "communication subspace" observed in visual areas by Semedo and colleagues (2019), where one population of neurons is coupled to another population through moment-to-moment co-fluctuations along a subset of dimensions in the overall activity space (Kohn et al., 2020; MacDowell et al., 2025). Geometric computations of this kind may also underlie both motor control and more abstract cognitive control (Vyas et al., 2020; Panichello & Buschman, 2021; Churchland & Shenoy, 2024); for example, preparatory motor activity can be maintained in an output-null subspace of the overall activity space to avoid prematurely triggering a motor output, then rotated into the output-potent subspace to initiate the action (Kaufman et al., 2014); similar population-level dynamics have been observed in working memory and attention tasks (Panichello & Buschman, 2021).

Closely related ideas have begun to emerge from efforts to understand information processing in large language models. The layers of a large language model are coupled to one another via a high-dimensional embedding space—the residual stream (Elhage et al., 2021). The attention heads (i.e., a circuit that allows the model to integrate information across words) within each layer process language by effectively reading and writing into subspaces of the residual stream. For example, an attention head in one layer can effectively communicate contextual information to an attention head in a downstream layer by modifying a subset of features in the current activity vector. In LLMs, this high-capacity residual stream is critical for encoding the rich contextual inflections of natural language; it allows the model to populate information from prior words into the current activity pattern (Desbordes et al., 2023; Muller et al., 2024). The individual "regions" of the model (i.e., the attention heads) make relatively small token-by-token adjustments to the overall activity of the shared residual stream. Different structures of language—such phonemic, syntactic, and semantic structures—are fused into a unified embedding space, and refined layer by layer. Integrating the contributions of different regions into a shared feature space may provide a computational explanation for why brain regions of the language network appear to have very similar functional tuning (Fedorenko et al., 2024).

We encountered several challenges in pursuing the core questions of this work. First, we observed negative ISFC and model connectivity between certain pairs of regions, such as the connections

between the EAC and DMN regions (TPJ and PMC), and the connections from TPJ to STG/S and IFG/S. For the sake of interpretational simplicity, this manuscript focuses on modeling pairs of regions with positive ISFC. We also encountered some surprising cases where the model connectivity diverged from the ISFC. For example, EAC and IFG/S were positively correlated in the ISFC matrix, but we found negative model connectivity across all three feature bands. Some of these observations could be attributed to lags in information processing between regions (Chang et al., 2022), the flexible timing incorporated into the fitting of the encoding models (Huth et al., 2016), and/or task-positive and task-negative network dynamics (Fox et al., 2005). Future work could pursue more detailed analyses of temporal interactions between regions, shedding further light on these observations. Finally, we resorted to relatively coarse parcel-level encoding models to reduce the computational burden and facilitate model evaluation across subjects. This choice is in tension with a body of work demonstrating that both the functional topography and functional tuning of language regions are quite variable across individual brains (Fedorenko et al., 2010; Huth et al., 2016; Mahowald & Fedorenko, 2016; Braga et al., 2020). Future work could use hyperalignment methods to obtain a finer-grained correspondence across individuals (Haxby et al., 2011, 2020; Van Uden, Nastase et al., 2018; Nastase et al., 2020b; Bhattacharjee et al., 2025).

Finally, throughout our results, we observed a gap between feature-specific model connectivity and the full stimulus-driven connectivity quantified using ISFC, even when combining acoustic, speech, and language features (**Figs. S2**, **S3**). This raises a question: what stimulus features are driving reliable connectivity during story listening that are not yet captured by our models? This gap could be partially due to the quality of the linguistic embeddings used to predict brain activity and connectivity. Embeddings from more advanced, and ideally more brain-like, language models may improve encoding performance and reduce the gap. The current pattern of results reveals one possible path forward. While the current selection of linguistic features captures a sizable proportion of ISFC in frontotemporal language areas, the gap is most pronounced in connections to higher-level default-mode areas (**Fig. S3c**). This suggests that current language models do not fully capture the transformation from linguistic representations to the more abstract narrative- or event-level representations thought to be encoded in the DMN (Baldassano et al., 2017; Chen et al., 2017; Yeshurun et al., 2021). As more human-like models emerge, our model-based connectivity framework provides a means to evaluate these models and reduce the gap.

Methods

fMRI data. This study used story-listening fMRI data from the openly available Narratives collection (Nastase et al., 2021). We used two story datasets acquired from the same 46 participants (32 females, mean age 23.33 ± 7.55). The two story stimuli are “I Knew You Were Black” (534 TRs) and “The Man Who Forgot Ray Bradbury” (558 TRs). Both stories were recorded in front of live audiences with occasional laughter, applause, and other audience reactions. “I Knew You Were Black” is an autobiographical account written and narrated by Carol Daniel which explores the intersection between her job on the radio and her identity as a Black woman. “The Man Who Forgot Ray Bradbury”, written and narrated by Neil Gaiman, is a story that explores themes of memory, forgetfulness, and

language at individual and collective levels. Functional data were acquired on a 3T Siemens Magnetom Prisma with a 1.5 s TR and 2.5 mm isotropic voxels. Refer to the data descriptor for more acquisition details (Nastase et al., 2021).

fMRI preprocessing. fMRI data were minimally preprocessed using fMRIPrep v20.0.5 (Esteban et al., 2019) including realignment, susceptibility distortion correction, spatial normalization, and resampling to the *fsaverage6* surface template (Fischl et al., 1999), as described in the data descriptor (Nastase et al., 2021). Confound regression was performed with the following nuisance variables: six head motion parameters, five principal components from both white matter and cerebrospinal fluid masks (aCompCor; Behzadi et al., 2007), cosine detrending variables, and two stimulus confounds tracking the number of words per TR, and whether a TR has words or silence. To reduce computational demands and facilitate intersubject analyses, vertex-wise time series were averaged within 1,000 parcels covering the entire cortex based on the functional atlas derived from resting-state functional connectivity (Schaefer et al., 2018).

Regions of interest. Across both hemispheres, 280 parcels were assigned to 46 language-related regions of interest (ROIs) consisting of 23 homotopic pairs (**Fig. 2g**), defined based on four methods: functionally defined language regions (Fedorenko et al., 2010), language localizer tasks (Lipkin et al., 2022), the NeuroSynth activation map for “language” (Yarkoni et al., 2011), and intersubject correlations from 345 subjects listening to spoken stories (Nastase et al., 2021). These ROIs were selected to capture as comprehensively as possible the full cortical hierarchy for spoken language, including early auditory cortex (EAC), all core language ROIs, default-mode areas associated with event representation and narrative comprehension, as well as speech articulation areas. The procedure for defining these regions is described in detail by Zada and colleagues (2025). The 46 ROIs consisted of left and right pairs for the following regions: EAC, superior temporal gyrus (STG), dorsal anterior, ventral anterior, dorsal mid, ventral mid, and posterior superior temporal sulcus (STSda, STSva, STSdm, STSvm, STSp), inferior frontal gyrus (IFG), orbital inferior frontal gyrus (IFG orb), inferior frontal sulcus (IFS), opercular inferior frontal gyrus (IFG oper), middle frontal gyrus (MFG), superior frontal language area (SFL), supramarginal gyrus (SMG), temporoparietal junction angular gyrus and inferior parietal lobule (TPJ AngG, TPJ IPL), control C (Cont C), posterior medial cortex A, B and C (PMC A, PMC B, PMC C), parahippocampal cortex (PHC), dorsomedial prefrontal cortex (dmPFC), and sensorimotor cortex (SM). To more easily summarize our results, we also grouped 15 of these ROIs to define the following five broad, anatomically-contiguous language-related regions: EAC, superior temporal gyrus and sulcus (STG/S: STG, STSda, STSva, STSdm, STSvm, STSp), inferior frontal gyrus and sulcus (IFG/S: IFG, IFG orb, IFS, IFG oper), temporoparietal junction (TPJ: TPJ AngG, TPJ IPL), and posterior medial cortex (PMC: PMC A, PMC B, PMC C).

Intersubject correlation and connectivity. Intersubject correlation (ISC) was computed by correlating parcel time series in each subject with the average time series across all other subjects for the corresponding parcel (i.e., leave-one-out ISC; Nastase et al., 2019). Intersubject functional connectivity (ISFC) was computed for each subject and story separately as the pairwise correlations of the subject’s parcel time series and the average parcel time series of all other subjects across all pairs of parcels. ISFC matrices were symmetrized by averaging the upper and lower off-diagonal triangles of

each ISFC matrix. ISFC matrices were then averaged across the two stories. The diagonal of the ISFC matrix corresponds to the within-parcel ISC values.

Following the logic of ISC, ISFC captures stimulus-driven connectivity, because the stimulus is the only source of variance that is time-locked across subjects. Whereas traditional within-subject functional connectivity (WSFC) can be driven by intrinsic fluctuations with idiosyncratic, subject-specific timing, ISFC isolates the stimulus-driven component of connectivity (Simony et al., 2016; Simony & Chang, 2020). That said, data-driven methods like ISC, ISFC, and WSFC do not tell us what stimulus features are driving activity and/or connectivity; for example, ISC in early auditory areas may be driven by low-level acoustic features, whereas ISC in lateral temporal language areas may be driven by higher-level linguistic features. To more precisely quantify what is driving the connectivity between regions, we need to test explicit models of different stimulus features.

Stimulus feature extraction. For the two spoken story stimuli, we extracted three types of word-level embeddings from Whisper ("openai/whisper-medium.en" from the HuggingFace library), a multimodal, transformer-based speech-to-text large language model (Radford et al., 2023). Whisper is a deep neural network using a full transformer architecture composed of separate encoder and decoder stacks. The encoder stack takes as input the speech waveform in a spectrogram format. The decoder stack takes as input text tokens corresponding to the words (or sub-words) in the audio transcript. For every word in the story, we extracted (up to) a 30-second audio segment preceding the current up until after the current word is articulated. At the same time, we extracted the words uttered in the 30-second segment from the transcript—again, ending in the current word. Then, we extracted the spectrogram from the audio, and split words into tokens. We fed this input to both the encoder and decoder in a full forward pass through the model. From the network's activations we collected three embeddings for each word: 1) an "acoustic" embedding from the activations just prior to the first transformer layer of the encoder stack; 2) a "speech" embedding from the activations after the last layer of the encoder stack; and 3) a contextual "language" embedding from the activations of the 20th layer of the decoder stack (out of 24 layers). We use the term "acoustic" to denote that these activations are closest to the audio input; we use the term "speech" embeddings because it is the final representation of the audio—and the one that is referenced by each layer of the decoder; and we use the term "language" for the contextual word embeddings from the decoder stack because this are most similar to the embeddings extracted from typical text-based large language models (Goldstein et al., 2025). All three types of embeddings are 1,024-dimensional vectors. Timing information from the transcripts (i.e., word onsets and offsets) were used to average word-level embeddings within each corresponding fMRI TR for use in the encoding models.

Intersubject encoding models. We used encoding models to quantify to what extent different linguistic features are encoded in the activity of a given brain region. The two stories were used alternately for training and testing encoding models. Banded ridge regression was used to estimate parcel-wise encoding models in the train story using all three sets of embeddings (i.e., feature spaces) jointly in three separate "bands" to allow these features to fairly compete for variance in the brain activity. Each feature band was assigned its own regularization parameter based on random search across 20 log-spaced parameters in the range [1, 10^{19}], using five-fold nested cross-validation within each training story. All encoding models were trained at the level of parcel time series. Encoding

models of this kind quantify the average feature tuning of neural populations within each parcel. Model-predicted BOLD activity was generated for the test story based on the joint model weights, as well as for the weights at each of the three feature bands separately. Encoding models were trained within each subject, but were evaluated by correlating the model-based predictions with the actual activity at a given parcel averaged across the remaining subjects, to more closely match the formulation of the ISC analysis. In this way, all encoding models were forced to generalize both across stories and across subjects.

Intersubject model-based connectivity. We then evaluated the encoding models fit within each parcel (from the preceding section) *across* parcels, following the logic of ISFC: the subject's model-predicted time series was correlated with the average actual time series of all other subjects across all pairs of parcels. We refer to this analysis as intersubject model-based functional connectivity (see Toneva et al., 2022, and Meschke et al., 2023, for related ideas). We computed a separate model-based connectivity matrix for each feature-band, as well as for the joint model. Model-based connectivity matrices were symmetrized in the same way as ISFC. When summarizing our results, we always average model-based connectivity values (correlations) among pairs of parcels, instead of averaging predicted or actual time series across parcels. ISFC conceptually serves as a noise ceiling for stimulus-driven connectivity that is reliable across subjects. In implementation, however, given that encoding models are more flexible in accounting for hemodynamic lags, model-based connectivity may numerically diverge from ISFC values.

In summarizing model connectivity results, we focused on eight language-related regions comprising 207 parcels that have been shown to exhibit strong ISC during story listening: EAC, STG/S, IFG/S, TPJ AngG, TPJ IPL, PMC A, PMC B, and PMC C. Within-region (model) connectivity was summarized for a given region by averaging the correlation values between all pairs of parcels within that region separately for each subject and feature band. Between-region (model) connectivity between a given pair of regions was summarized by averaging the correlation values between all pairs of parcels connecting the two regions. For visualization and interpretational simplicity, we opted not to evaluate model performance across regions (i.e., model connectivity) for pairs of regions with negative ISFC values.

Statistical testing. We assessed the statistical significance of whole-brain ISC and encoding model performance at each parcel using a one-sample *t*-test across subjects (**Fig. 1**). The false discovery rate (FDR) was controlled at $p < .05$ across 1,000 parcels. To determine whether model-based connectivity values differed significantly from one another, we performed paired *t*-tests ($df = 45$) between the model performance values for different feature bands. FDR was controlled at $p < .05$ among the comparisons under consideration. For visualization, we generated error bars by bootstrapping subject-level values (i.e., resampling subjects with replacement) 1,000 times for each mean value.

Data, Materials, and Software Availability. The fMRI data used here are openly available as part of the Narratives collection (Nastase et al., 2021): https://doi.org/10.18112/openneuro.ds002345.v1.1.4; https://datasets.datalad.org/?dir=/labs/hasson/narratives; https://fcon_1000.projects.nitrc.org/indi/retro/Narratives.html. The code used to perform the core analyses of this study is available at https://github.com/zaidzada/narrative-enc.

Acknowledgments

We are grateful for the use of data from the freely-available Narratives collection. We would like to acknowledge funding sources: UBC Friedman Award for Scholars in Health and BC Children's Hospital Research Institute Doctoral Studentship (AS); National Institutes of Health CRCNS grant R01DC022534 (ZZ, UH, SAN).

Author contributions:

Conceptualization: AS, ZZ, UH, SAN
Data curation: ZZ, SAN
Formal analysis: AS, ZZ
Funding acquisition: AS, TV, UH
Investigation: AS, ZZ
Methodology: AS, ZZ, SAN
Project administration: SAN
Software: AS, ZZ
Supervision: TV, UH, SAN
Visualization: AS
Writing – original draft: AS, ZZ, SAN
Writing – review & editing: AS, ZZ, TV, UH, SAN

Competing interests: Authors declare that they have no competing interests.

References

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709–721. https://doi.org/10.1016/j.neuron.2017.06.041

Bhattacharjee, A., Zada, Z., Wang, H., Aubrey, B., Doyle, W., Dugan, P., Friedman, D., Devinsky, O., Flinker, A., Ramadge, P. J., Hasson, U., Goldstein, A., & Nastase, S. A. (2025). Aligning brains into a shared space improves their alignment to large language models. *Nature Computational Science*. https://doi.org/10.1101/2024.06.04.597448

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Blank, I., Balewski, Z., Mahowald, K., & Fedorenko, E. (2016). Syntactic processing is distributed across the language system. *NeuroImage*, *127*, 307–323. https://doi.org/10.1016/j.neuroimage.2015.11.069

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. https://doi.org/10.1152/jn.00884.2013

Braga, R. M., DiNicola, L. M., Becker, H. C., & Buckner, R. L. (2020). Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. *Journal of Neurophysiology*, *124*(5), 1415–1448. https://doi.org/10.1152/jn.00753.2019

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., & Olah, C. (2023). Towards monosemanticity: decomposing language models with dictionary learning. *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features/index.html

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems: Vol. 33* (pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Catani, M., Jones, D. K., & Ffytche, D. H. (2005). Perisylvian language networks of the human brain. *Annals of Neurology*, *57*(1), 8–16. https://doi.org/10.1002/ana.20319

Caucheteux, C., Gramfort, A., & King, J.-R. (2021). Disentangling syntax and semantics in the brain with deep networks. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 1336–1348). PMLR. https://proceedings.mlr.press/v139/caucheteux21a.html

Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134. https://doi.org/10.1038/s42003-022-03036-1

Chang, C. H., Nastase, S. A., & Hasson, U. (2022). Information flow across the cortical timescale hierarchy during narrative construction. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(51), e2209307119. https://doi.org/10.1073/pnas.2209307119

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*, 115–125. https://doi.org/10.1038/nn.4450

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: a fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62. https://doi.org/10.1017/S0140525X1500031X

Churchland, M. M., & Shenoy, K. V. (2024). Preparatory activity and the expansive null-space. *Nature Reviews Neuroscience*, *25*(4), 213–236. https://doi.org/10.1038/s41583-024-00796-z

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, *37*(27), 6539–6557. https://doi.org/10.1523/jneurosci.3267-16.2017

Desbordes, T., Lakretz, Y., Chanoine, V., Oquab, M., Badier, J. M., Trébuchon, A., Carron, R., Bénar, C.-G., Dehaene, S., & King, J. R. (2023). Dimensionality and ramping: Signatures of sentence integration in the dynamics of brains and deep language models. *Journal of Neuroscience*, *43*(29), 5350–5364. https://doi.org/10.1523/jneurosci.1163-22.2023

Dick, A. S., Bernal, B., & Tremblay, P. (2014). The language connectome: new pathways, new concepts. *The Neuroscientist*, *20*(5), 453–467. https://doi.org/10.1177/1073858413513502

Du, J., DiNicola, L. M., Angeli, P. A., Saadon-Grosman, N., Sun, W., Kaiser, S., Ladopoulou, J., Xue, A., Yeo, B. T. T., Eldaief, M. C., & Buckner, R. L. (2024). Organization of the human cerebral cortex estimated within individuals: networks, global topography, and function. *Journal of Neurophysiology*, *131*(6), 1014–1082. https://doi.org/10.1152/jn.00308.2023

Duffau, H. (2008). The anatomo-functional connectivity of language revisited: new insights provided by electrostimulation and tractography. *Neuropsychologia*, *46*(4), 927–934. https://doi.org/10.1016/j.neuropsychologia.2007.10.025

Dupré la Tour, T., Eickenberg, M., Nunez-Elizalde, A. O., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, *264*, 119728. https://doi.org/10.1016/j.neuroimage.2022.119728

Dupré la Tour, T., Visconti di Oleggio Castello, M., & Gallant, J. L. (2025). The Voxelwise Encoding Model framework: a tutorial introduction to fitting encoding models to fMRI data. *Imaging Neuroscience*. https://doi.org/10.1162/imag_a_00575

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy models of superposition. *Transformer Circuits Thread*. https://transformer-circuits.pub/2022/toy_model/index.html

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., & Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*, 111–116. https://doi.org/10.1038/s41592-018-0235-4

Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, *203*, 104348. https://doi.org/10.1016/j.cognition.2020.104348

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. https://doi.org/10.1152/jn.00032.2010

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, *25*(5), 289–312. https://doi.org/10.1038/s41583-024-00802-4

Fedorenko, E., Nieto-Castanon, A., & Kanwisher, N. (2012). Lexical and syntactic representations in the brain: an fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia*, *50*(4), 499–513. https://doi.org/10.1016/j.neuropsychologia.2011.09.014

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(41), E6256–E6262. https://doi.org/10.1073/pnas.1612132113

Fischl, B., Sereno, M. I., Tootell, R. B., & Dale, A. M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, *8*(4), 272–284. https://doi.org/10.1002/(SICI)1097-0193(1999)8:4%3C272::AID-HBM10%3E3.0.CO;2-4

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(27), 9673–9678. https://doi.org/10.1073/pnas.0504136102

Friederici, A. D. (2009). Pathways to language: fiber tracts in the human brain. *Trends in Cognitive Sciences*, *13*(4), 175–181. https://doi.org/10.1016/j.tics.2009.01.001

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological Reviews*, *91*(4), 1357–1392. https://doi.org/10.1152/physrev.00006.2011

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010

Goldstein, A., Wang, H., Niekerken, L., Zada, Z., Aubrey, B., Sheffer, T., Nastase, S. A., Gazula, H., Schain, M., Singh, A., Rao, A., Choe, G., Kim, C., Doyle, W., Friedman, D., Devore, S., Dugan, P., Hassidim, A., Brenner, M., Matias, Y., Devinsky, O., Flinker, A., & Hasson, U. (2025). A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature Human Behaviour*, *9*, 1041–1055. https://doi.org/10.1038/s41562-025-02105-9

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanual D., Cohen, A., Jensen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Lora, F., Flinker, A., Devore, S., Doyle, W., Dugan, P., Friedman, D., Hassidim, A., Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., & Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*, 369–380. https://doi.org/10.1038/s41593-022-01026-4

Hampson, M., Peterson, B. S., Skudlarski, P., Gatenby, J. C., & Gore, J. C. (2002). Detection of functional connectivity using temporal correlations in MR images. *Human Brain Mapping*, *15*(4), 247–262. https://doi.org/10.1002/hbm.10022

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, *303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., & Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, *72*(2), 404–416. https://doi.org/10.1016/j.neuron.2011.08.026

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020). Hyperalignment: modeling shared information encoded in idiosyncratic cortical topographies. *eLife*, *9*, e56601. https://doi.org/10.7554/eLife.56601

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. https://doi.org/10.1038/nature17637

Jolly, E., Sadhukha, S., & Chang, L. J. (2020). Custom-molded headcases have limited efficacy in reducing head motion during naturalistic fMRI experiments. *NeuroImage*, *222*, 117207.

Kaufman, M. T., Churchland, M. M., Ryu, S. I., & Shenoy, K. V. (2014). Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience*, *17*(3), 440–448. https://doi.org/10.1038/nn.3643

Kohn, A., Jasper, A. I., Semedo, J. D., Gokcen, E., Machens, C. K., & Yu, B. M. (2020). Principles of corticocortical communication: proposed schemes and design considerations. *Trends in Neurosciences*, *43*(9), 725–737. https://doi.org/10.1016/j.tins.2020.07.001

Kong, R., Yang, Q., Gordon, E., Xue, A., Yan, X., Orban, C., Zuo, X.-N., Spreng, N., Ge, T., Holmes, A., & Yeo, B. T. T. (2021). Individual-specific areal-level parcellations improve functional connectivity prediction of behavior. *Cerebral Cortex*, *31*(10), 4477–4500. https://doi.org/10.1093/cercor/bhab101

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2024). Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, *15*, 5523. https://doi.org/10.1038/s41467-024-49173-5

Lee, M. H., Hacker, C. D., Snyder, A. Z., Corbetta, M., Zhang, D., Leuthardt, E. C., & Shimony, J. S. (2012). Clustering of resting state networks. *PLOS One*, *7*(7), e40370. https://doi.org/10.1371/journal.pone.0040370

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906–2915. https://doi.org/10.1523/jneurosci.3684-10.2011

Lipkin, B., Tuckute, G., Affourtit, J., Small, H., Mineroff, Z., Kean, H., Jouravlev, O., Rakocevic, L., Pritchett, B., Siegelman, M., Hoeflin, C., Pongos, A., Blank, I. A., Struhl, M. K., Ivanova, A., Shannon, S., Sathe, A., Hoffman, M., Nieto-Castañón, A., & Fedorenko, E. (2022). Probabilistic atlas for the language network based on precision fMRI data from > 800 individuals. *Scientific Data*, *9*, 529. https://doi.org/10.1038/s41597-022-01645-3

MacDowell, C. J., Libby, A., Jahn, C. I., Tafazoli, S., Ardalan, A., & Buschman, T. J. (2025). Multiplexed subspaces route neural activity across brain-wide networks. *Nature Communications*, *16*, 3359. https://doi.org/10.1038/s41467-025-58698-2

Mahowald, K., & Fedorenko, E. (2016). Reliable individual-level neural markers of high-level language processing: a necessary precursor for relating neural variability to behavioral and genetic variability. *NeuroImage*, *139*, 74–93. https://doi.org/10.1016/j.neuroimage.2016.05.073

McAvoy, M., Mitra, A., Coalson, R. S., d'Avossa, G., Keidel, J. L., Petersen, S. E., & Raichle, M. E. (2016). Unmasking language lateralization in human brain intrinsic activity. *Cerebral Cortex*, *26*(4), 1733–1746. https://doi.org/10.1093/cercor/bhv007

Meschke, E. X., Castello, M. V. D. O., Tour, T. D. L., & Gallant, J. L. (2023). Model connectivity: leveraging the power of encoding models to overcome the limitations of functional connectivity. *bioRxiv*. https://doi.org/10.1101/2023.07.17.549356

Muller, L., Churchland, P. S., & Sejnowski, T. J. (2024). Transformers and cortical waves: encoders for pulling in context across time. *Trends in Neurosciences*, *47*(10), 788–802. https://doi.org/10.1016/j.tins.2024.08.006

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, *14*(6), 667–685. https://doi.org/10.1093/scan/nsz037

Nastase, S. A., Goldstein, A., & Hasson, U. (2020a). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*, 117254. https://doi.org/10.1016/j.neuroimage.2020.117254

Nastase, S. A., Liu, Y. F., Hillman, H., Norman, K. A., & Hasson, U. (2020b). Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space. *NeuroImage*, *217*, 116865. https://doi.org/10.1016/j.neuroimage.2020.116865

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C. J., Yeshurun, Y., Regev, M., Nguyen, M., Chang, C. H. C., Baldassano, C., Lositsky, O., Simony, E., Chow, M. A., Leong, Y. C., Brooks, P. P., Micciche, E., Choe, G., Goldstein, A., Vanderwal, T., Halchenko, Y. O., Norman, K. A., & Hasson, U. (2021). The “Narratives” fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, *8*, 250. https://doi.org/10.1038/s41597-021-01033-3

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C., & Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(18), E3669–E3678. https://doi.org/10.1073/pnas.1701590114

Noble, S., Curtiss, J., Pessoa, L., & Scheinost, D. (2024). The tip of the iceberg: a call to embrace anti-localizationism in human neuroscience research. *Imaging Neuroscience*, *2*, 1–10. https://doi.org/10.1162/imag_a_00138

Panichello, M. F., & Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature*, *592*(7855), 601–605. https://doi.org/10.1038/s41586-021-03390-w

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*(2), 816–847. https://doi.org/10.1016/j.neuroimage.2012.04.062

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning: Vol. 202* (pp. 28492–28518). PMLR. https://proceedings.mlr.press/v202/radford23a.html

Reddy, A. J., & Wehbe, L. (2021). Can fMRI reveal the representation of syntactic structure in the brain? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems* (Vol. 34, pp. 9843–9856). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2021/hash/51a472c08e21aef54ed749806e3e6490-Abstract.html

Salvo, J. J., Anderson, N. L., & Braga, R. M. (2025). Intrinsic functional connectivity delineates transmodal language functions. *Imaging Neuroscience*. https://doi.org/10.1162/imag.a.25

Saur, D., Kreher, B. W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M. S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., & Weiller, C. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(46), 18035–18040. https://doi.org/10.1073/pnas.0805234105

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X. N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, *28*(9), 3095–3114. https://doi.org/10.1093/cercor/bhx179

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M., & Kohn, A. (2019). Cortical areas interact through a communication subspace. *Neuron*, 102(1), 249–259. https://doi.org/10.1016/j.neuron.2019.01.026

Shain, C., & Fedorenko, E. (2025). A language network in the individualized functional connectomes of over 1,000 human brains doing arbitrary tasks. *bioRxiv*. https://doi.org/10.1101/2025.03.29.646067

Shain, C., Kean, H., Casto, C., Lipkin, B., Affourtit, J., Siegelman, M., Mollica, F., & Fedorenko, E. (2024). Distributed sensitivity to syntax and semantics throughout the language network. *Journal of Cognitive Neuroscience*, *36*(7), 1427–1471. https://doi.org/10.1162/jocn_a_02164

Simony, E., & Chang, C. (2020). Analysis of stimulus-induced brain dynamics during naturalistic paradigms. *NeuroImage*, *216*, 116461. https://doi.org/10.1016/j.neuroimage.2019.116461

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, *7*, 12141. https://doi.org/10.1038/ncomms12141

Stephens, G. J., Silbert, L. J., & Hasson, U. (2010). Speaker–listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(32), 14425–14430. https://doi.org/10.1073/pnas.1008662107

Tie, Y., Rigolo, L., Norton, I. H., Huang, R. Y., Wu, W., Orringer, D., Mukundan Jr., S., & Golby, A. J. (2014). Defining language networks from resting-state fMRI for surgical planning—a feasibility study. *Human Brain Mapping*, *35*(3), 1018–1030. https://doi.org/10.1002/hbm.22231

Tomasi, D., & Volkow, N. D. (2012). Resting functional connectivity of language networks: characterization and reproducibility. *Molecular Psychiatry*, *17*(8), 841–854. https://doi.org/10.1038/mp.2011.177

Toneva, M., Mitchell, T. M., & Wehbe, L. (2022). Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, *2*(11), 745–757. https://doi.org/10.1038/s43588-022-00354-6

Toneva, M., Williams, J., Bollu, A., Dann, C., & Wehbe, L. (2022). Same cause; different effects in the brain. In B. Schölkopf, C. Uhler, & K. Zhang (Eds.), *Proceedings of the First Conference on Causal Learning and Reasoning* (Vol. 177, pp. 787–825). PMLR. https://proceedings.mlr.press/v177/toneva22a.html

Tuckute, G., Kanwisher, N., & Fedorenko, E. (2024). Language in brains, minds, and machines. *Annual Review of Neuroscience*, *47*(2024), 277–301. https://doi.org/10.1146/annurev-neuro-120623-101142

Turken, A. U., & Dronkers, N. F. (2011). The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. *Frontiers in Systems Neuroscience*, *5*. https://doi.org/10.3389/fnsys.2011.00001

Van Uden, C. E., Nastase, S. A., Connolly, A. C., Feilong, M., Hansen, I., Gobbini, M. I., & Haxby, J. V. (2018). Modeling semantic encoding in a common neural representational space. *Frontiers in Neuroscience*, *12*, 437. https://doi.org/10.3389/fnins.2018.00437

Vyas, S., Golub, M. D., Sussillo, D., & Shenoy, K. V. (2020). Computation through neural population dynamics. *Annual Review of Neuroscience*, *43*(1), 249–275. https://doi.org/10.1146/annurev-neuro-092619-094115

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLOS One*, *9*(11), e112575. https://doi.org/10.1371/journal.pone.0112575

Wu, M. C. K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, *29*(1), 477–505. https://doi.org/10.1146/annurev.neuro.29.051605.113024

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635

Yeshurun, Y., Nguyen, M., & Hasson, U. (2021). The default mode network: where the idiosyncratic self meets the shared social world. *Nature Reviews Neuroscience*, *22*(3), 181–192. https://doi.org/10.1038/s41583-020-00420-w

Zada, Z., Goldstein, A. Y., Michelmann, S., Simony, E., Price, A., Hasenfratz, L., Barham, E., Zadbood, A., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Devore, S., Flinker, A., Devinsky, O., Hasson, U.*, & Nastase, S. A.* (2024). A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron*, *112*(18), 3211–3222. https://doi.org/10.1016/j.neuron.2024.06.025

Zada, Z., Nastase, S. A., Speer, S., Mwilambwe-Tshilobo, L., Tsoi, L., Burns, S., Falk, E., Hasson, U., & Tamir, D. (2025). Linguistic coupling between neural systems for speech production and comprehension during real-time dyadic conversations. *bioRxiv*. https://doi.org/10.1101/2025.02.14.638276

Zhu, L., Fan, Y., Zou, Q., Wang, J., Gao, J. H., & Niu, Z. (2014). Temporal reliability and lateralization of the resting-state language network. *PLOS One*, *9*(1), e85880. https://doi.org/10.1371/journal.pone.0085880

Supplementary Information

Cortical language areas are coupled via a soft hierarchy of model-based linguistic features

Ahmad Samara[1], Zaid Zada[2,3], Tamara Vanderwal[1,4], Uri Hasson[2,3], Samuel A. Nastase[2]

[1] University of British Columbia, Vancouver, BC, Canada
[2] Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA
[3] Department of Psychology, Princeton University, Princeton, NJ, USA
[4] BC Children's Hospital Research Institute, Vancouver, BC, Canada

Corresponding authors: Samuel A. Nastase (snastase@princeton.edu), Tamara Vanderwal (tamara.vanderwal@ubc.ca).

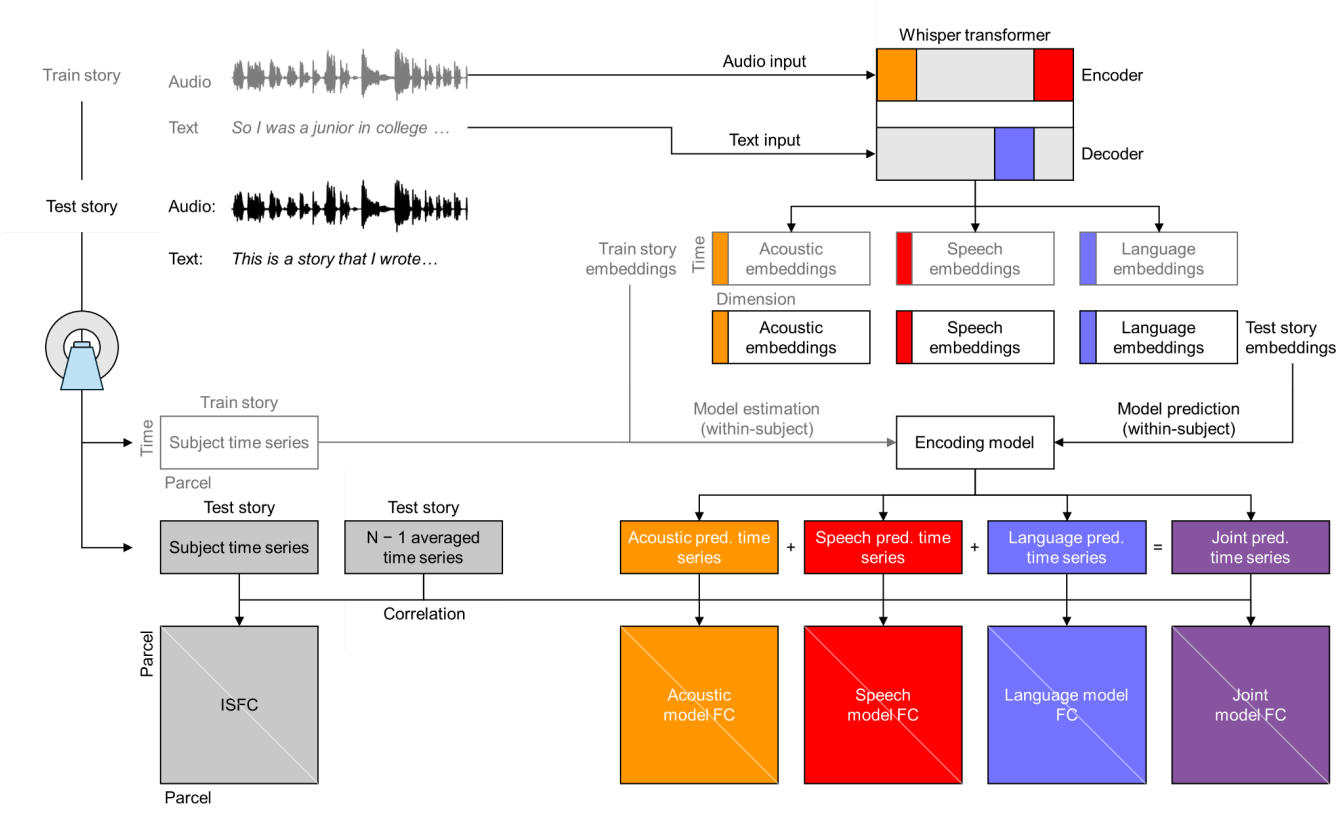

Fig. S1. Schematic workflow and encoding model evaluation. fMRI data were collected from 46 subjects while they listened to two spoken narratives: each story served as the train story and test story. The transcript and audio spectrograms of the train and test stories were fed into a large language model (Whisper) to extract distinct word-level embedding representations of the story as follows: acoustic embeddings were extracted from the input to the first transformer layer of the encoder stack; speech embeddings were extracted from the last encoder stack; and language embeddings were extracted layer 20 of the decoder stack. The train story embeddings were then used in combination with train story parcel-resolution fMRI time series data to estimate parcel-wise encoding models using banded ridge regression. Model-based predictions of parcel time series were generated based on the weights associated with each of the three feature bands, as well as the joint model weights. In parallel, intersubject function connectivity (ISFC) was computed as the pairwise correlation of a given subject's time series and the group-averaged time series of all other subjects for all pairs of brain parcels. Model-based connectivity was also computed following the logic of ISFC by correlating the subject's predicted time series and the group-average actual time series of all other subjects for all pairs of parcels.

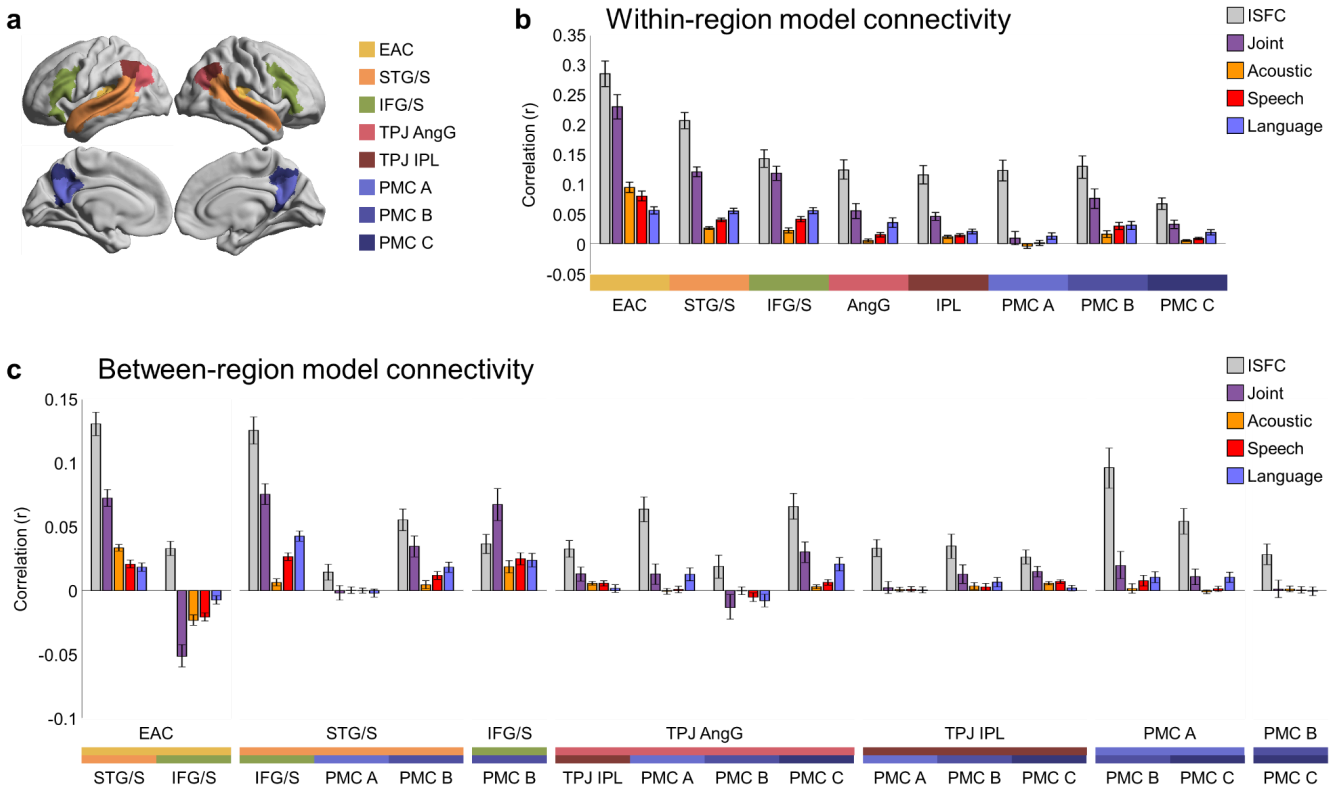

Fig. S2. Model-based connectivity within and between language regions in reference to intersubject functional connectivity, related to Fig. 3. (a) Focusing on the same eight language-related regions (early auditory cortex [EAC], superior temporal gyrus and sulcus, [STG/S], inferior frontal gyrus and sulcus [IFG/S], temporoparietal junction angular gyrus and inferior parietal lobule [TPJ AngG and TPJ IPL], and posterior medial cortex A, B and C [PMC A, PMC B, and PMC C]) from **Fig. 3**, we reproduced the feature-specific model connectivity results with the addition of the joint model connectivity and intersubject functional connectivity (ISFC) values for both the within **(b)** and between **(c)** language region. Similar to **Fig. 3**, between-region model connectivity results are shown only for region pairs with positive ISFC values. Error bars indicate bootstrap 95% confidence intervals.

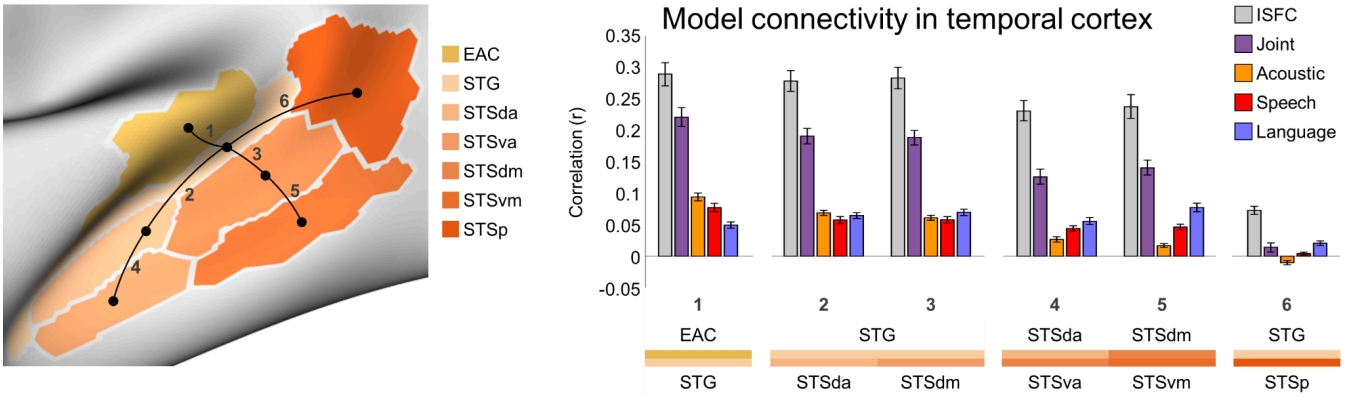

Fig. S3. Model connectivity along superior temporal pathways in reference to intersubject functional connectivity, related to Fig. 5. We reproduced model connectivity results across parcel pairs along pathways linking early auditory cortex (EAC), superior temporal gyrus (STG), and superior temporal sulcus (STS) with the addition of joint model connectivity and intersubject functional connectivity (ISFC) values for reference. Error bars indicate bootstrap 95% confidence intervals.