Neuron

Article

Linguistic coupling between neural systems for speech production and comprehension during real-time dyadic conversations

Highlights

- Everyday conversations require speakers to alternate between speaking and listening
- fMRI hyperscanning enabled us to study naturalistic, free-form dyadic conversations
- Linguistic encoding analyses reveal common neural systems for speaking and listening
- Conversation elicits unique neural processes not engaged by passive comprehension

Authors

Zaid Zada, Samuel A. Nastase, Sebastian Speer, ..., Emily Falk, Uri Hasson, Diana I. Tamir

Correspondence

zzada@princeton.edu

In brief

Zada et al. use fMRI hyperscanning to simultaneously record dyads engaging in free-form, interactive conversations. They find that speech production and comprehension rely on highly overlapping neural representations across the cortical language network. Brain-to-brain coupling is strongest in areas associated with social cognition.

Zada et al., 2026, Neuron *114*, 774–787
February 18, 2026
https://doi.org/10.1016/j.neuron.2025.11.004

CellPress

Neuron

Article

Linguistic coupling between neural systems for speech production and comprehension during real-time dyadic conversations

Zaid Zada,[1,2,10,*] Samuel A. Nastase,[3] Sebastian Speer,[1,2] Laetitia Mwilambwe-Tshilobo,[1,4,5] Lily Tsoi,[6] Shannon M. Burns,[7,8] Emily Falk,[4,5,9] Uri Hasson,[1,2] and Diana I. Tamir[1]

[1]Department of Psychology, Princeton University, Princeton, Princeton, NJ 08544, USA
[2]Princeton Neuroscience Institute, Princeton University, Princeton, Princeton, NJ 08544, USA
[3]Department of Psychology and Center for Computational Language Sciences, University of Southern California, Los Angeles, CA 90089, USA
[4]Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA 19104, USA
[5]Annenberg Public Policy Center, University of Pennsylvania, Philadelphia, PA 19104, USA
[6]Department of Psychology, Caldwell University, Caldwell, NJ 07006, USA
[7]Department of Psychological Science, Pomona College, Claremont, CA 91711, USA
[8]Department of Neuroscience, Pomona College, Claremont, CA 91711, USA
[9]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA
[10]Lead contact
*Correspondence: zzada@princeton.edu

https://doi.org/10.1016/j.neuron.2025.11.004

SUMMARY

The core use of human language is to send complex ideas from one mind to another. In everyday conversations, comprehension and production are intertwined, as speakers and listeners alternate roles. Nonetheless, the neural systems underlying these faculties are typically studied in isolation, using paradigms that cannot capture interactive communication. Here, we used fMRI hyperscanning to simultaneously record dyads engaged in real-time conversations. We used language model embeddings to quantify the degree to which production and comprehension systems rely on shared neural representations, both within and across brains. We found that both processes key into overlapping neural systems, with similar neural tuning for both processes, spanning the cortical language network. Speaker-listener coupling extended beyond the language network into areas associated with social cognition. Our results suggest that the neural systems for speech comprehension and production align with common linguistic features encoded in a broad cortical network for language and communication.

INTRODUCTION

Everyday language relies on two fundamental processes: production and comprehension (e.g., speaking and listening). While these may seem different and are often studied separately, there is evidence for shared representations and mechanisms between them.[1–6] However, to truly test the overlap between these neural systems, we need to explicitly model their linguistic representations and study dyads in real-time interactive conversation who engage in both speaking and listening. Recent advances allowed us to address these points by using large language models (LLMs) as explicit computational models of brain activity during real-time conversations using fMRI hyperscanning.

Prior studies on the similarity between production and comprehension neural processes have only indirectly tested their overlap. These studies typically use different tasks for speaking and listening and then compare whether the same brain region is activated for both. For example, Awad and colleagues[7] localized brain regions for comprehension using a contrast between listening to regular speech versus rotated speech and regions for production by contrasting activity during counting versus speaking. Contrast-based approaches may find the same region involved in both processes,[6] but this does not inform us about the similarity between their representations. Some studies used a two-brain approach, such as speaker-listener coupling, as a data-driven test of similarity.[3,8] However, intersubject correlation methods are fundamentally content-agnostic—they cannot tell us “what” is shared between brains or processes. Other studies use an adaptation paradigm to probe for syntax and semantics.[9,10] However, these methods do not require a generalized computational model in a predictive framework.[11] To quantify linguistic representations shared between communicators, we need an explicit model of linguistic features.[12] Here, we aimed to use explicit features to quantitatively compare the underlying representations across production and comprehension.

Encoding models have recently been used to quantify linguistic features encoded in neural activity during passive language comprehension[13–15] and, less commonly, language production.[16–18] LLMs have been shown to have similarity judgments comparable to humans[19,20] and have internal linguistic representations that are more similar to the human brain than any other model during language comprehension.[21–24] By leveraging the rich linguistic representations from LLMs, encoding models can identify the components of neural activity that encode linguistic features. These methods have begun to lend further support for the integrated view of neural representations for production and comprehension both within subjects[16–18] and across subjects[12] during real-time dialogue. For example, Zada and colleagues[12] used contextual embeddings from an LLM as an explicit, shared model mediating speaker-listener neural coupling during real-time conversations in electrocorticography (ECoG). They leveraged the temporal resolution of ECoG to identify the temporal dynamics of coupling between speaker and listener. While that study focused on how speaker and listener are aligned to shared linguistic space, here we aim to quantify the functional overlap between the neural systems for language production (speaking) and comprehension (listening). Furthermore, we use whole-brain fMRI to overcome the limited spatial coverage of ECoG, which in large part excluded the right hemisphere and midline areas associated with narrative comprehension and social cognition. Finally, most studies using LLM embeddings for encoding analyses have only been applied to comprehension, and those that include production only do so while recording from one subject.

In real-time dyadic conversations, production and comprehension are contemporaneous and interleaved. Production is often spontaneous, and comprehension must be proactive, as listeners must be ready to respond in a relevant way as they process the incoming speech.[25,26] Conversation is unique in that it requires coordination in the form of interactive alignment[27] and agreement on meaning through common ground.[28–30] Conversation is also arguably the most fundamental setting of language use. It is universal to human societies, does not require specialized skills (e.g., literacy) or technologies (e.g., telephones),[31] and allows people to go well beyond simple stimulus-response signaling to share and shape each other's representational thought through language. However, previous research has decoupled production and comprehension, using separate tasks and stimuli for each process, controlled paradigms (e.g., rehearsed speech and covert production), and isolated linguistic contexts. This raises the question of whether these paradigms fully capture the neural systems for real-time, interactive conversation.[32,33] Here, we are able to directly compare linguistic processing during active conversation versus "passive" listening, where only comprehension is necessary.

Conversational language is fundamentally a social process. Previous work has investigated production-comprehension coupling across subjects using a sequential, asynchronous protocol: first recording a subject speaking and then playing the speech back to multiple listeners at a later time.[3,8,34–39] Using content-agnostic measures, these studies find that during communication, the speaker's neural activity is coupled to that of the listeners in regions associated with both comprehension and production. Moreover, the strength of speaker-listener coupling is related to outcome measures of comprehension. Hyperscanning paradigms, where researchers simultaneously measure neural processes during dyadic social interactions using two MRI scanners, are uniquely suited to studying language usage in interactive social settings[25,40–46] and can inform the extent of shared neural mechanisms and speaker-listener alignment.[47,48] Here, we extend previous work[3,8] using asynchronous paradigms, where subjects passively listen to a recording of the speaker, to investigate the coupling of neural systems for production and comprehension in real-time, interactive conversations.

In this paper, we developed an fMRI hyperscanning paradigm to simultaneously measure whole-brain activity in dyadic pairs of subjects engaged in free-form, interactive conversations across a range of prompted topics. We used these data to answer five questions: first, we aimed to directly quantify the degree of overlap of linguistic representations across different brain areas associated with language, moving beyond identifying brain regions that are involved in both speech production and speech comprehension. By combining encoding models with word embeddings that broadly capture linguistic structure, we can formally quantify overlap between production and comprehension in terms of the functional *tuning* of linguistic features for a given brain area. Functional tuning refers to a set of encoding weights across model features indicating the extent to which different features predict a given voxel's activity.[49] We hypothesized that production and comprehension processes will substantially overlap, in terms of functional tuning, in mid- and high-level regions of the language network, with minimal overlap in early perceptual and motor regions (Figure 1B). Second, we also collected fMRI data where the same subjects simply listened to a 13-min story. This allowed us to test the similarity between comprehension while only listening and production and comprehension during conversation. We also predicted that some functional tuning would be shared whether one is actively listening in a conversation and preparing to respond or if they are passively listening to a narrative—but that interactive conversation relies on unique linguistic features not fully captured by story-listening encoding models. Third, we tested *what* features in the communication signal explained the neural coupling. Based on prior literature, we expected that the contextual word embeddings from LLMs would best capture a broad range of linguistic features encoded in neural activity, over other low-level features. Fourth, we used the real-time, dyadic conversations paradigm to compute model-based coupling between the speaker's and listener's brains. We expected to find speaker-listener coupling in language areas, but coupling was in fact strongest in areas associated with social cognition. Finally, we explored the spatial and temporal dynamics of speaker-listener coupling. These questions were uniquely testable in our dataset of simultaneous recordings of brain activity in two individuals engaged in real-time, interactive conversations. Capitalizing on the whole-brain coverage of fMRI and the rich linguistic structure captured by LLM embeddings, we gain valuable insights into how successful conversations depend on shared language representations between production and comprehension across various cortical regions.

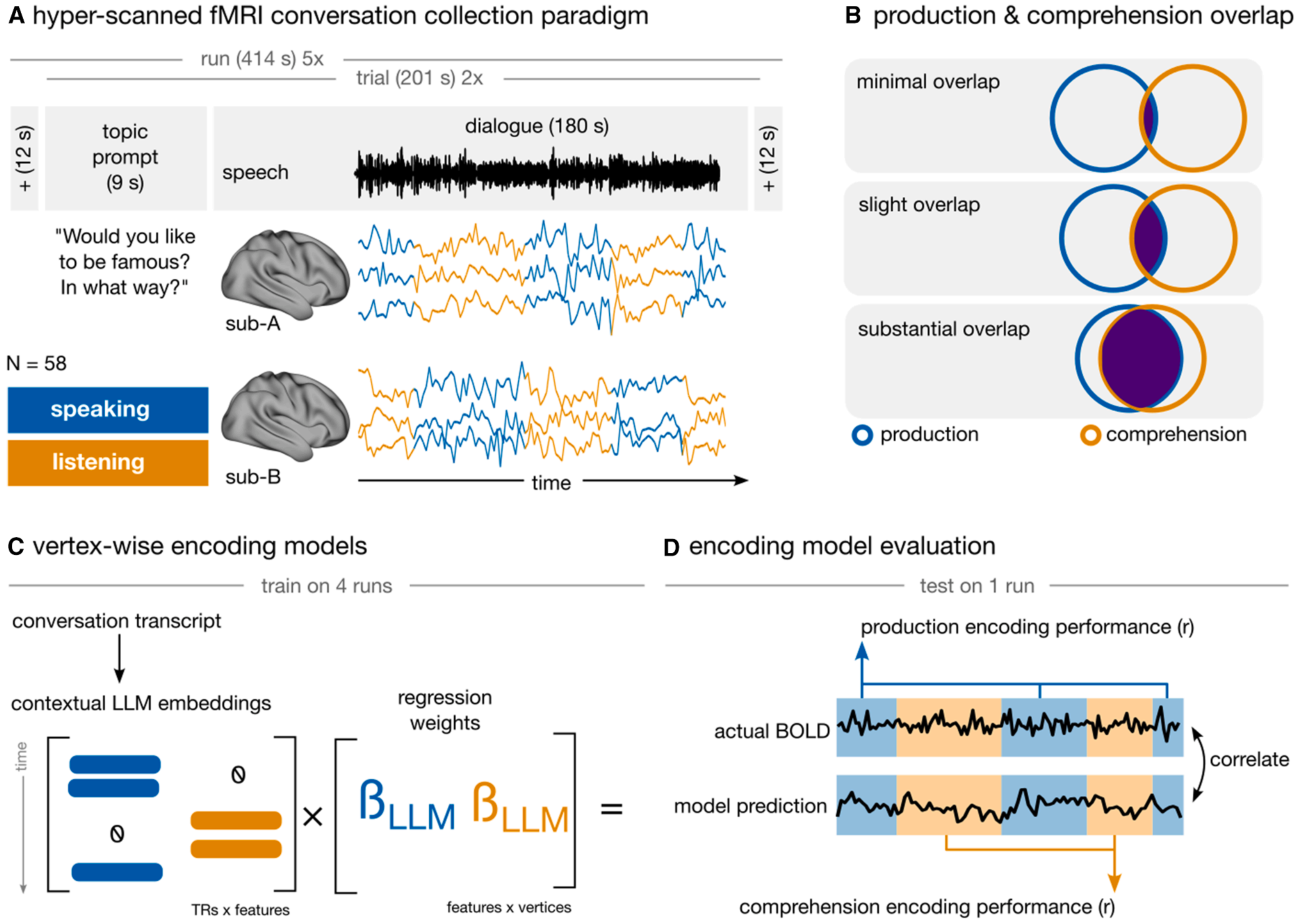

Figure 1. Data collection and modeling framework

(A) We collected fMRI data simultaneously from pairs of subjects as they engaged in interactive, prompted conversations.

(B) We aimed to quantify the overlap of production and comprehension processes using an explicit computational model.

(C) We extracted LLM word embeddings from each conversation transcript and used them as encoding model features to predict the subject's brain activity. We split the regressors into separate time series for production (blue) and comprehension (orange).

(D) Finally, we evaluated the performance of production and comprehension time points separately in a held-out test run of different conversations.

RESULTS

We aimed to model production and comprehension processing within and between brains during free-form, turn-based conversations. We used hyperscanning to collect simultaneous fMRI data in 30 dyads (60 subjects) as they freely discussed ten topics across five ~6 min runs (Figure 1A).[46] Topic prompts were presented as a starting point, but each dyad was free to pursue the discussion differently, resulting in 30 unique conversations (Table S1). To characterize the linguistic content in the blood-oxygen-level-dependent (BOLD) signal, we explicitly represented the language stimuli with several different feature spaces: confound variables (e.g., word rate), spectral acoustic features, phonemic articulatory features, and word embeddings extracted from GPT-2.[50] With reference to LLM embeddings, we use the terms "linguistic structure" and "linguistic features" inclusively to refer to the representations of grammar, word meaning, and context learned by LLMs in order to produce natural language. Then, we used banded ridge regression to estimate a linear mapping from the model features onto the BOLD activity at each vertex[14,51–53] (Figure 1C). To evaluate the models, we correlated the model-predicted and actual BOLD time series for left-out runs for each feature space and for production or comprehension time points separately (Figure 1D). Finally, we averaged the model performance correlations across subjects for all analyses. Statistically, we evaluated the average using a one-sample t-test, correcting for multiple comparisons over all ~75k cortical vertices. To summarize our results, we averaged the encoding performance across vertices within 11 regions of interest (ROIs) spanning an extended language network, from low-level auditory and articulatory areas to high-level semantic areas: early auditory cortex (EAC), posterior and anterior superior temporal gyrus (pSTG and aSTG), inferior and middle frontal gyri (IFG and MFG), somatomotor cortex (SM), supplementary motor area (SMA), frontal opercular (FOP), intraparietal sulcus (IPS), temporoparietal junction (TPJ), and posterior medial cortex (PMC).

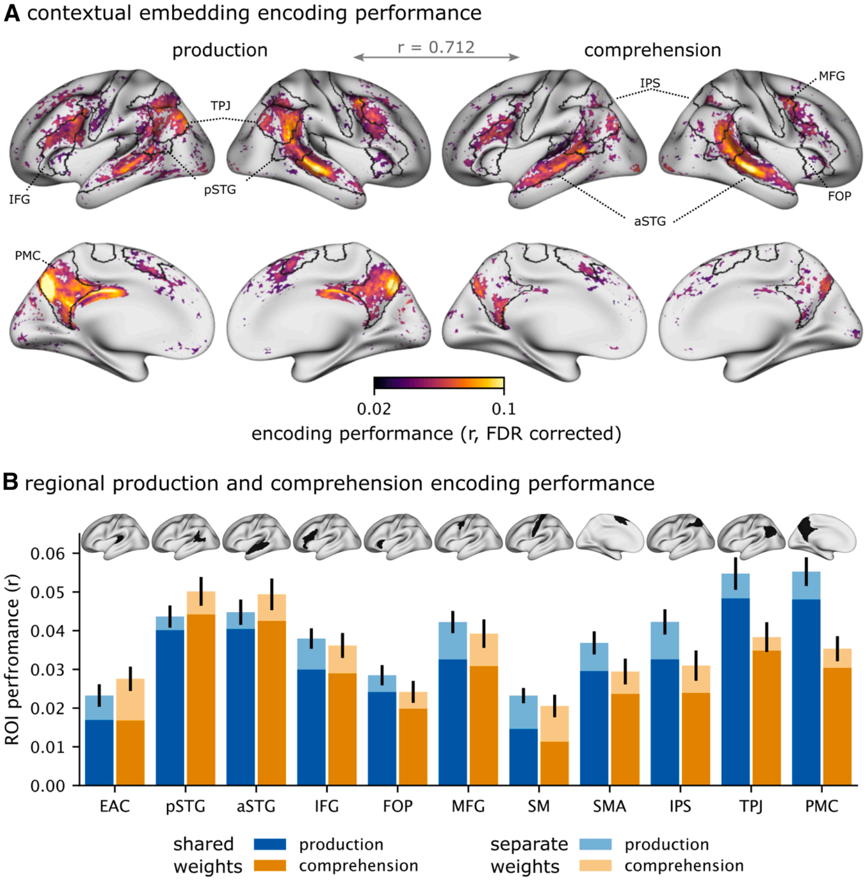

Figure 2. Within-subject speaking and listening encoding performance

(A) Encoding model performance of the separate-weights model for production and comprehension relative to the control feature spaces.

(B) We summarized the un-thresholded encoding performance of the shared- and separate-weights models in 11 ROIs spanning the extended language network, averaged across left and right hemispheres. Error bars indicate the standard error of the mean over parcels within each ROI.

See also Figures S5 and S6 and Table S3.

Contextual embeddings capture both production and comprehension

We first validated that we can successfully model brain activity during spontaneous production and comprehension in our hyperscanning paradigm. To do so, we built two models to quantify linguistic processing and to measure the cortical overlap between production and comprehension. In one, we constrained the model to learn one set of shared weights for production and comprehension for all feature spaces. In this model, a vertex must code for both processes with the same functional tuning (i.e., shared weights) to be well predicted. In the second model, we split all regressors into separate sets for production and comprehension, allowing the model to learn separate weights for each process (Figure 1C). We treat the confound, acoustic, and phonemic feature sets as nuisance variables and report only the LLM contextual embedding performance. We first inspect the performance of the second, more flexible model, which we expect to outperform the unified constrained model.

Using the more flexible model with separate weights for production and comprehension, we found significant within-subject encoding performance throughout the core language network: STG, IFG, and MFG for speech production and speech comprehension (Figure 2A). Moreover, encoding performance extended bilaterally to higher-level regions like TPJ and PMC. We found considerable spatial overlap between encoding performance for production and comprehension—i.e., vertices well predicted during production are also likely to be well predicted during comprehension ($r = 0.712$, $p < 1e{-}5$). To quantify whether production and comprehension encoding rely on shared or divergent weights, we compared the performance of the shared-weights and separate-weights models. We found that across all 11 ROIs, 80% of encoding performance can be attributed to shared functional tuning rather than idiosyncratic production- or comprehension-specific variance (Figure 2B). Peripheral regions for speech perception (EAC) and speech articulation (SM) showed the least shared tuning, maintaining modality-specific representations. These results suggest that cortical activity during both production and comprehension keys to similar features captured by the LLM embeddings throughout the language network. Higher-level language areas are more likely to show overlap because they do not need to represent acoustic or motor features present in peripheral regions, whereas lower-level areas show less overlap because speech processing and articulatory planning occur at the peripherals of the processing hierarchy.

We observed several qualitative differences across tasks, regions, and hemispheres. First, overall encoding performance appears higher in the right STG and TPJ than in the left-hemisphere homologs. Second, overall encoding performance appears stronger for production, especially in bilateral PMC and right TPJ. Third, encoding performance for comprehension appears stronger and more bilateral in STG than in production. Despite these differences, the overall encoding performance suggests that LLM embeddings provide a rich basis for modeling linguistic encoding throughout much of the cortex.

Story-listening comprehension shares a subset of linguistic features with interactive production and comprehension neural systems

In addition to the hyperscanning conversations paradigm, we recorded participants as they listened to a 13-min story in a separate scanning session. This presented a unique opportunity to compare linguistic processing during spontaneous production, (inter)active comprehension, and non-interactive

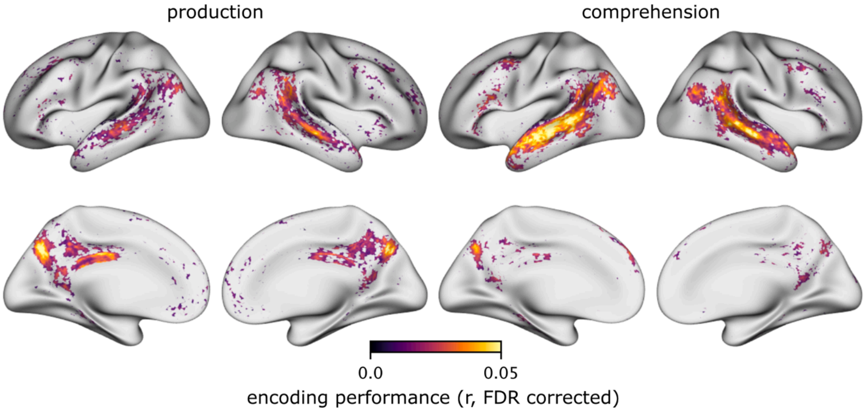

Figure 3. Encoding models trained on passive listening partially generalize to neural responses during conversations

Participants passively listened to a 13-min story in a separate scanning session before the hyperscanning procedure. We estimated encoding models using the same four feature spaces from this passive listening-only dataset. Then we evaluated how well they generalize to data acquired during conversational production and comprehension conversations. Here, we present only the performance of the contextual embedding feature space, after testing for significance and correcting for multiple comparisons. We found significant encoding performance in STG and PMC, which is significantly weaker than when training on conversations.

See also Figures S1 and S7.

comprehension. Specifically, we aimed to test the shared processing between passive listening and active comprehension and production. To do so, for each subject, we estimated a comprehension encoding model using the story data only and then evaluated the fitted model on the subject's conversation data. We extracted the same four feature spaces from the story and evaluated the model performance similarly to the conversation models.

We found significant within-subject generalization performance from the passive listening paradigm to the conversational paradigm for production and comprehension (Figure 3). Generalization to conversational comprehension was stronger than to production. However, both were lower than when training on conversational data, capturing only a portion of the variance as training on conversations—even when equating their training data (Figure S1). Training on conversation data resulted in an increase of +41% in average encoding performance for comprehension and an increase of +49% for production. A paired t-test found a significant difference (p < 0.00212) between subjects' average vertex encoding performance when training on conversation or story. Generalization performance was more bilateral than performance based on the conversational paradigm. An overlapping set of regions was well predicted, particularly STG during comprehension and PMC during production. Notably, generalization was poorer for IFG and MFG compared with temporal regions. Though incomplete, generalization from passive comprehension to both production and comprehension in a conversation context provides further evidence for a common subset of linguistic features that span both processes, while still highlighting the boost in these systems during active, naturalistic communication.

Contextual embeddings outperform other features of speech and language

Our modeling framework allows us to test different hypotheses about features of brain activity during production and comprehension by comparing the performance of different models. So far, we have only reported the performance of the contextual word embeddings from a pre-trained language model. However, we can decompose the joint model performance into the relative contribution from each feature space. Here, we report the performance of each feature space and then use a variance partitioning analysis to compute the unique variance predicted by the contextual LLM embeddings.

We found that during both production and comprehension, the contextual LLM embeddings outperformed all other feature spaces regarding correlation strength and cortical coverage (Figures 4A and S2). Among the lower-level control feature spaces, we observed that the acoustic features were the most predictive, especially in EAC and STG. By contrast, the articulation band was least predictive throughout all regions (likely due to collinearity with the better-fitting acoustic space). Moreover, the confounding variables were most predictive in SM, EAC, and aSTG. These regions are likely to exhibit large signal fluctuations between speech production or comprehension and are more susceptible to regressors such as word rate.

Next, we performed a variance partitioning analysis[15,54,55] to isolate the unique variance explained by the contextual LLM embeddings. We use hierarchical regression to compare a full model with all features and a nested model excluding the features of interest. In this analysis, the full model is composed of the LLM contextual embeddings (L), acoustic (A), and articulatory phonemic (P) features, resulting in encoding performance $R_{L,A,P}$. The nested model is the same, except that it excludes the LLM contextual embeddings from the predictors. Therefore, the unique contextual variance can be calculated as $U_L = R_{L,A,P} - R_{A,P}$ (displayed as U_{LLM} in the figure). The contextual embeddings accounted for unique variance bilaterally across all previously reported brain regions (Figure 4B). Together, these results suggest that while part of the variability in brain activity can be predicted by acoustic speech features, the contextual word embeddings of LLMs provide unique predictive power, especially in higher-order regions. By precisely quantifying the extent of the embeddings' advantage in each region of interest, we replicate previous findings that contextual word embeddings outperform simpler linguistic features[12,22,24] and that semantics are more predictive than syntax or phonology[56–58] by precisely quantifying the extent of the embeddings' advantage in each region of interest. Crucially, this also demonstrates that our encoding model is able to "decompose" the neural activity into a linguistic component that is separate from acoustic and phonetic-related activity.

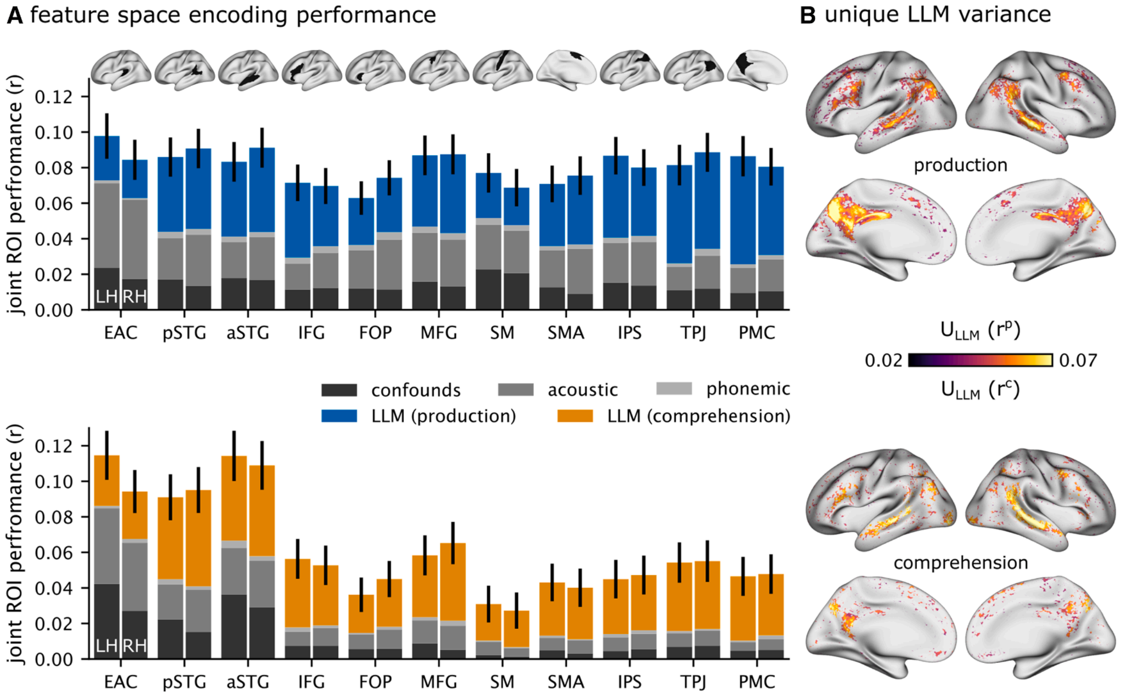

Figure 4. Model comparison and variance partitioning

We compared the variance explained by LLM embeddings with other linguistic feature spaces.

(A) The joint encoding performance of the full model was decomposed into the contribution of each space separately for production and comprehension. Error bars indicate standard error of the mean for the LLM band only, over voxels within each ROI.

(B) We performed a variance partitioning analysis within subjects to quantify the unique contribution of LLM word embeddings. We trained one full encoding model with all features and a nested model with all features, excluding the LLM word embeddings. Then, we subtracted the nested model performance from the full model to quantify the unique variance explained by the LLM embeddings.

See also Figure S2.

Model-based brain-to-brain coupling between conversational partners

When two people converse, we expect their brain activity to align along certain shared features between speech production and comprehension.[3,8,12,47] Consider a face-to-face conversation: neural activity may align on linguistic features (e.g., the meaning of words) and non-linguistic features (e.g., gestures and facial expressions). To isolate *linguistic* features of shared brain activity across brains, we estimated encoding models from LLM embeddings (jointly with control features) and evaluated how well models trained on one subject generalize to their conversational partner. Specifically, given subject A and their conversational partner, subject B, we correlated subject A's production model predictions with subject B's actual comprehension neural responses (Figure 5A). This analysis enabled us to test whether subject A's encoding models in one conversational role can generalize and predict their partner's neural responses in the other conversational role, vertex by vertex.[12,59] Our previous results showed that production and comprehension rely on similar brain regions and share similar linguistic features *within* subjects. This analysis reveals areas where production and comprehension are linguistically coupled *between* subjects.

We found significant model-based speaker-listener coupling for LLM embeddings in the right hemisphere along pSTG, extending into the TPJ, the MFG, and bilaterally in the precuneus in PMC (Figure 5B). Because the trained encoding model has to generalize to another subject's brain performing a different process (production versus comprehension), the overall magnitude of the correlation is lower. Interestingly, this model-based linguistic coupling appears right-lateralized (in right-handed subjects). While relatively few vertices in left-hemisphere language areas were significant, we observed strong coupling in right-lateralized temporal areas and in bilateral PMC. For example, brain-to-brain coupling for LLM embeddings was found in the right TPJ, a structure commonly associated with mentalizing and social cognition.[60,61] Thus, unlike within subjects, where we find broad and bilateral model-based coupling (e.g., in STG, IFG, and PMC), model-based coupling between speaker and listener relies on right-lateralized pSTG and TPJ regions and bilateral precuneus, which are regarded as higher-order cognition areas. We also correlated the functional tuning of voxels (encoding model weights) across all pairs of subjects to compare between actual and pseudo-dyads. We found weight similarity across many language areas between actual dyads and pseudo-dyads, with greater alignment within actual dyads (Figure S3). This suggests that certain features in the high-dimensional embedding space are shared across conversations, and these may afford communicators a generalized linguistic understanding. Some features particular to individual conversations may be

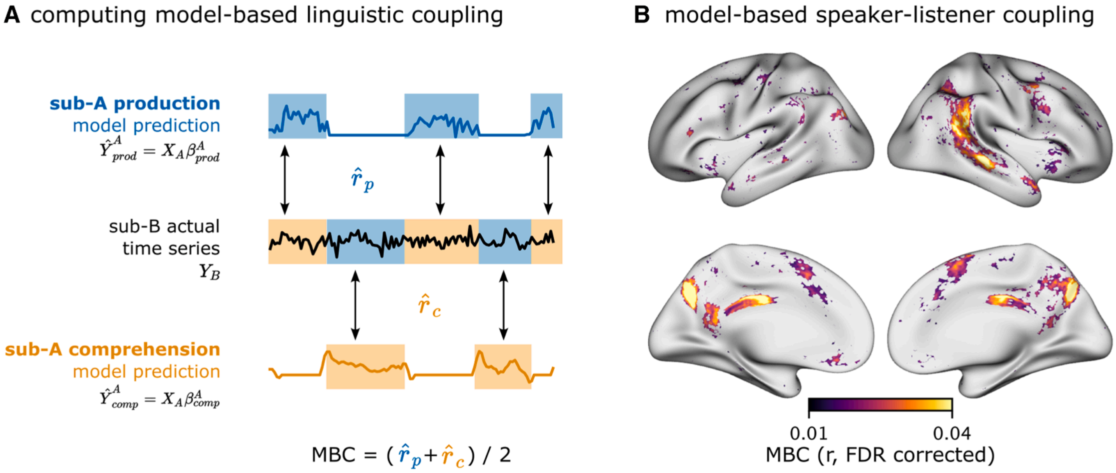
A computing model-based linguistic coupling
sub-A production
model prediction
$\hat{Y}^{A}_{prod} = X_A \beta^{A}_{prod}$
$\hat{r}_p$
sub-B actual
time series
Y_B
$\hat{r}_c$
sub-A comprehension
model prediction
$\hat{Y}^{A}_{comp} = X_A \beta^{A}_{comp}$
MBC = ($\hat{r}_p$+$\hat{r}_c$) / 2
B model-based speaker-listener coupling
0.01
0.04
MBC (r, FDR corrected)

Figure 5. Model-based speaker-listener coupling

(A) Schematic of model-based coupling (MBC). We use the already-trained encoding models from subject A's production data to predict subject B's time series during comprehension. We correlate subject A's model predictions with subject B's time series separately for production and comprehension to obtain two correlations per subject per trial.

(B) We average production and comprehension coupling correlations to obtain a group map of MBC.

See also Figure S3.

intertwined with, or sit on top of, those for language processing, and these may afford high-level situation model representations that underlie mutual understanding.

Spatial and temporal network structure in model-based conversational coupling

So far, we have restricted the scope of speaker-listener coupling in both spatial and temporal dimensions for simplicity: we have only considered coupling between one brain area in the speaker and the homologous area in the listener, and we have only considered instantaneous, or "zero-lag," coupling between partners. The reality is much more complicated. For example, activity in some areas of the speaker's brain may be coupled to activity in *different* regions of the listener's brain, and in some cases, the speaker's brain may *precede* that of the listener.[8,12] Here, we briefly explore variations in coupling along both of these axes. Since vertices are plentiful, adding spatial and temporal dimensions would exponentially increase the number of comparisons. Thus, we constrain this exploratory analysis to the 11 discussed ROIs.

We first assessed how well a model trained on the speaker's brain activity in one ROI generalizes to the listener's brain activity across all other ROIs. We did this by averaging the predicted and actual time series within each ROI across its vertices. This generated an inter-regional generalization matrix that summarizes the speaker-listener coupling across all ROI pairs at lag 0 (Figure 6A). We observed that the right hemisphere is more connected between speaker and listener than the left hemisphere. Moreover, some areas are relatively uncoupled from other regions (e.g., SM), whereas others are coupled with multiple areas (e.g., pSTG). Interestingly, this matrix has no clear diagonal, meaning that speaker-listener coupling across areas is similarly strong (or weak) to coupling between homologs.

To investigate temporal variation in linguistic coupling, we cross-correlated the predicted time series in subject A with the actual time series of subject B at varying lags (building off of Figure 5A). This resulted in a temporal profile of interregional model-based speaker-listener coupling for each ROI pair (484). A carpet plot of these profiles sorted by their peak lag suggests different clusters of temporal coupling (Figures 6B and S4). Most pairs of regions exhibit peak coupling at lag 0 ± 3 s (e.g., TPJ, PMC, and FOP). For a subset of region pairs, the speaker's brain precedes the listener's brain. For example, the speaker's left MFG and SMA precede the listener's brain activity (e.g., the speaker's MFG precedes the listener's pSTG). In another subset of region pairs, the listener's brain appears to precede the speaker. For example, the speaker's right aSTG and pSTG tend to lag behind the listener's brain activity (e.g., the speaker's aSTG lags behind the listener's pSTG). These results suggest that linguistic coupling between conversation partners is spatially and temporally extended.

DISCUSSION

The neural systems involved in speech production and comprehension may require different processes, but they must converge on similar representations. After all, a shared linguistic space is necessary to align the linguistic information across the speaker's and listener's brains. In this paper, we aimed to quantify the extent to which production and comprehension rely on shared neural machinery during natural conversations. Our findings revealed that speech production and comprehension recruit similar brain areas and shared linguistic representations when engaged in natural conversation (Figure 2). The brain's linguistic processing during passive story listening generalizes to spontaneous speech production and comprehension during

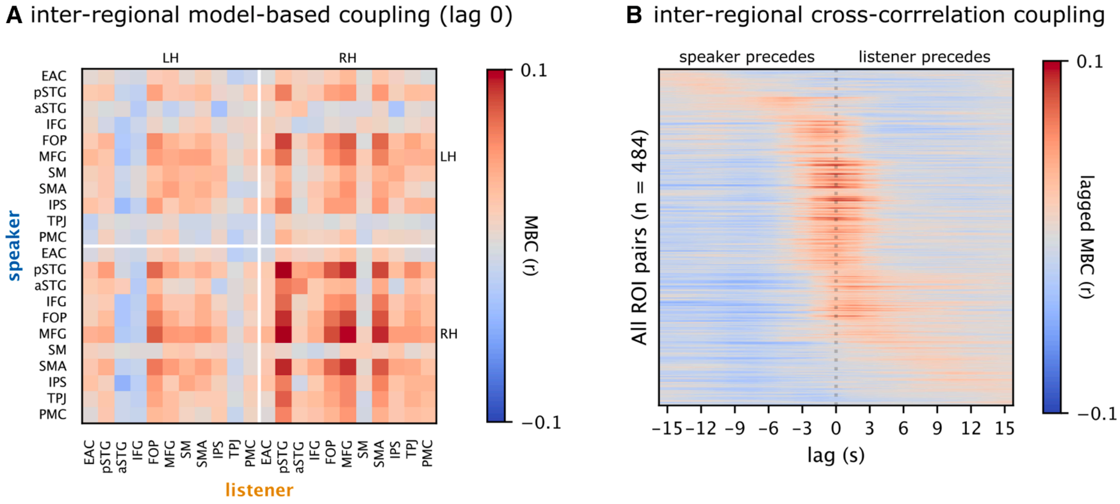

Figure 6. Inter-regional and cross-correlated model-based coupling

(A) We extend the speaker-listener model-based coupling results (Figure 5) along two dimensions. First, we correlate a speaker's model-based prediction (averaged across vertices within ROIs) to all ROIs in the listener.

(B) Second, for each pair of ROIs (a total of 484 pairs across both hemispheres), we cross-correlate the speaker's predicted time series and the listener's actual time series to extend coupling temporally. Rows are ordered by the lag at which they achieve maximum encoding performance.

See also Figure S4.

conversations (Figure 3). However, passive encoding performance was significantly weaker and missed key frontal language areas compared with when training on active conversations. The model comparison analyses demonstrated that contextual embeddings from an LLM better capture the linguistic features shared between production and comprehension than other candidate models (Figure 4). Finally, our model-based coupling analysis revealed brain-to-brain production-comprehension coupling in high-level cortical areas, particularly right-hemisphere areas associated with social cognition (Figure 5). Our results extend previous work by explicitly modeling the linguistic features encoded in brain activity, by simultaneously recording interacting participants in free-form conversations, and by quantifying the overlap in speech production and comprehension across the entire cortex. Overall, our results suggest that speech comprehension and speech production systems align on a set of shared, intermediate features, allowing the brain to translate between the two processes effectively.

We identified a unified language network with shared weights engaged during production and comprehension in real-time conversations. Encoding models have become essential for mapping linguistic features (e.g., acoustic, syntactic, and semantic features) to brain activity. Many recent studies have applied them during passive language comprehension.[14,15,22–24,62–64] However, only a handful of recent studies have begun leveraging encoding models for spontaneous language production and active comprehension,[16–18] and even fewer have simultaneously recorded two participants engaged in dialogue.[12,65] Our encoding models were able to predict neural responses during spontaneous speech production and comprehension (Figure 2A). We found that brain regions that can be predicted during speech production can also be predicted during comprehension, similar to previous findings in ECoG[12,22] and fMRI.[17] Building on this result, we directly tested the amount of overlap between these neural systems by constraining the encoding model to share weights—i.e., use the same functional tuning to predict both production and comprehension brain activity. We found that most brain regions exhibited shared linguistic representations between production and comprehension (Figure 2B). The features shared between the neural systems for production and comprehension are best captured by the contextual word embeddings, relative to control features (e.g., acoustic and phonemic features). LLMs encode a variety of linguistic features related to words, grammar, and contextual meaning in a high-dimensional embedding space. We believe these embeddings capture what is meaningful about the content—the linguistic, semantic, and social representations—that people share when they communicate. This provides evidence for an overlap between speech production and comprehension, which relies on a unified and shared language network. Part of this common network constitutes well-established language regions[66] and extends into general systems responsible for interactive social cognition. We also found a production-comprehension overlap in low-level perceptual and motor areas (e.g., EAC and SM), suggesting that modality-specific areas may be more localized than previously thought. By using natural conversations, we were able to demonstrate how participants engage these neural processes in real-world, interactive communication[67,68] that embodies the principles of ecological validity in social neuroscience.[32,33,69]

Our data allowed us to test a novel contrast of language use for each subject: in one case only using one's comprehension system while listening to a story and in the other using both production and comprehension systems to engage in conversation. We leveraged this to apply a conservative test: whether comprehension-only neural processing is related to production and comprehension systems during conversation. Using encoding models trained on the story-listening data, we found that they generalized across paradigms to both production and comprehension during conversation (Figure 3). This suggests that even for highly different linguistic contexts/tasks (i.e., passive comprehension versus active conversation), there are *some* linguistic features that are processed similarly, especially in the superior temporal cortex. On one hand, this implies that certain brain areas process a core set of linguistic features regardless of usage. On the other hand, this result implies that the *degree* of representation overlap in certain brain regions (e.g., IFG) depends on the way language is used. For example, the active comprehension required by real-time communication—like inferring the speaker's intended meaning or planning a response—may elicit deeper linguistic processing that is uniquely captured by models trained on conversational data.

Advances in simultaneous neuroimaging allowed us to move beyond asynchronous protocols of speaker-listener coupling to real-time, turn-taking conversations.[25,45] Our hyperscanning paradigm allowed us to simultaneously record brain activity during speech production and speech comprehension in two interacting subjects. Whereas landmark studies were limited to relating a single subject's production to multiple subjects' comprehension responses acquired at a later time,[3,8] our paradigm engages each subject's production and comprehension processes in an (inter)-active, real-time, turn-taking conversation. We found that conversations recruit more brain regions and different representations than non-interactive paradigms (Figure S1). Similar to studies of asynchronous communication, we observed production-comprehension coupling in PMC, pSTG, and TPJ (Figure 5B). These studies found more speaker-listener coupling in lateral left-hemisphere language areas than our findings. This difference is likely due to our real-time, dyadic conversation paradigm, compared with their asynchronous monologue design. Extra-linguistic features of conversation, such as situational or social factors, conversational dynamics (e.g., turn-taking), and other contextual structures may play an outsized role in this paradigm, and resolving the communicative challenges raised in these contexts may rely more heavily on the right hemisphere. While we did not observe strong speaker-listener coupling in the left hemisphere, this absence of an effect does not indicate that there is no coupling at all. Zada and colleagues,[12] for example, found extensive speaker-listener coupling in the left hemisphere using the fast temporal resolution with ECoG. The slow temporal resolution of fMRI and the short turns during conversations (i.e., quick production-comprehension switches) may make it difficult to capture this activity. Another possibility is that left-hemisphere language areas are more topographically idiosyncratic across individuals,[70] resulting in mismatched vertex-level coupling. Subject-specific language localizers and a more nuanced connectivity analysis may better capture linguistic coupling across subjects.

We speculate that interactive communication, where partners must actively listen and figuratively "speak to" one another's thoughts and intentions, may engage the social brain in a way that traditional language paradigms do not. Historically, language processing—both comprehension and production processes—has been associated with the left hemisphere.[71–75] On the other hand, both ISC analyses[76] and encoding models[14] tend to yield largely bilateral maps during natural language comprehension. In this study, we observed brain-to-brain linguistic coupling in the right-lateralized superior temporal cortex, TPJ, and prefrontal cortex, as well as the bilateral precuneus and posterior cingulate. This result indicates that the same features that mediate between comprehension and production processes within a brain are also partly shared across individuals. However, these areas are not simply right-hemisphere homologs of typical language regions.[66,77] In the neuropsychology literature, the right hemisphere has been associated with affective, prosodic, and paralinguistic features of speech, as well as pragmatic and discourse-level processing that unfold over longer timescales, relative to the left hemisphere.[78–84] Neuroimaging work has generally corroborated these findings.[85–88] For example, Yarkoni and colleagues[89] reported a very similar set of regions to ours, including right TPJ and bilateral posterior cingulate and precuneus, involved explicitly in tracking narrative comprehension across sentences. Interestingly, several of these areas overlap with regions often associated with mentalizing and other aspects of social cognition,[90,91] highlighting the key role that the social brain may play in real-time, naturalistic social interactions. We suspect that right-hemisphere brain regions, in particular, pick up on socially relevant linguistic features within the word embedding space[14] that are shared between speakers. Future work is needed to relate the extent of neural alignment between conversation partners and conversational content or behavioral outcomes.

Limitations of the study

Our study has some limitations worth noting. First, our hyperscanning conversation paradigm imposes certain restrictions on the "naturalness" of the data. Real-world conversations are typically face-to-face and use multiple communication channels beyond language (e.g., facial expressions and body language). These are not accessible in an MRI machine. Our paradigm allowed subjects to spontaneously share their thoughts given an open-ended prompt, speak for as long as they wanted within the trial, and explicitly exchange turns. However, necessary limitations of an fMRI paradigm (e.g., talking in an fMRI machine, not face-to-face, and with an on-screen countdown) may have induced additional neural processes (e.g., task monitoring) that differ from those occurring during truly spontaneous, naturalistic conversation. That said, we do not expect that additional processes impact our brain-to-brain coupling results because of our strict confound modeling and the unlikely possibility that they contain a predictable signal from word embeddings. Moreover, speakers were required to press a button to transfer the mic to their partner. This procedure deviates from face-to-face conversation and could minimize processes related to turn-taking management, such as prediction and preparation.[92] We suspect that without the constraint of passing the mic, production and comprehension processes may overlap to a greater degree.

Conclusion

LLMs are trained to predict the next word in large text corpora. After training, these models can generate increasingly fluent, surprisingly meaningful language, one word at a time, by sampling from a probability distribution of upcoming words. These models do not have dedicated systems for comprehension or production resembling anything like the human brain. Why do these models capture neural activity so well during language comprehension and production? The architecture, objective function, and end-to-end training regime of an LLM constrain the model to learn representations for "comprehension" that are intrinsically useful for "production" in a continuous input-output loop. This is distinct from both (1) cognitive architectures where the production system scaffolds comprehension[93] and (2) cognitive architectures where both production and comprehension systems key into an amodal conceptual system.[94] We speculate that this constraint, which forces language models to learn *shared* representations that are useful in pursuing its objective function of producing real-world linguistic outputs in response to real-world linguistic inputs, may yield embeddings that can capture brain activity during both comprehension and production. While the brain has specialized systems for perception and production, our findings suggest that much of the brain's language machinery occupies a middle ground similar to LLM embeddings: multimodal, active representations with mixed features for both comprehension and production.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Zaid Zada (zzada@princeton.edu).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- Raw fMRI data have been deposited at the NIMH Data Archive and are publicly available as of the date of publication at NIMH: https://dx.doi.org/10.15154/w7y8-yd67. The conversation audio and transcript data reported in this study cannot be deposited in a public repository because of the sensitive and personally identifiable information of the conversation content. To request access, contact Diana I. Tamir (dtamir@princeton.edu).
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We would like to extend our thanks to Ahmad Samara and Itamar Jalon for their helpful feedback on head motion correction and fMRI preprocessing details. We are also grateful to Junfeng Lu and the anonymous reviewers for their constructive feedback that strengthened the paper. This work was supported by the National Institutes of Health grants R21MH127284 (D.I.T.) and R01DC 022534 (U.H.).

AUTHOR CONTRIBUTIONS

Z.Z., S.A.N., U.H., and D.I.T.: conceptualization; Z.Z., S.A.N., L.M.-T., U.H., and D.I.T.: writing – review & editing; S.A.N. and D.I.T.: supervision; L.T., S.M.B., and S.S.: data curation; D.I.T.: funding acquisition; Z.Z.: formal analysis, investigation, methodology, software, visualization, and writing – original draft.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Participants
- METHOD DETAILS
  - ○ Design
  - ○ MRI acquisition
  - ○ Conversation audio transcription
  - ○ fMRIPrep preprocessing
  - ○ Functional data preprocessing
  - ○ Confound and head motion correction
  - ○ Defining cortical regions of interest
  - ○ Linguistic features for encoding analysis
- QUANTIFICATIONS AND STATISTICAL ANALYSIS
  - ○ Encoding model construction and evaluation
  - ○ Speaker–listener model-based coupling
  - ○ Story-listening task and analysis
  - ○ Software resources

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2025.11.004.

Received: July 7, 2025
Revised: September 24, 2025
Accepted: November 8, 2025
Published: December 17, 2025

REFERENCES

1. Papathanassiou, D., Etard, O., Mellet, E., Zago, L., Mazoyer, B., and Tzourio-Mazoyer, N. (2000). A common language network for comprehension and production: a contribution to the definition of language epicenters with PET. Neuroimage *11*, 347–357. https://doi.org/10.1006/nimg.2000.0546.
2. Menenti, L., Pickering, M.J., and Garrod, S.C. (2012). Toward a neural basis of interactive alignment in conversation. Front. Hum. Neurosci. *6*, 185. https://doi.org/10.3389/fnhum.2012.00185.
3. Silbert, L.J., Honey, C.J., Simony, E., Poeppel, D., and Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. Proc. Natl. Acad. Sci. USA. *111*, E4687–E4696. https://doi.org/10.1073/pnas.1323812111.
4. Gambi, C., and Pickering, M.J. (2017). Models linking production and comprehension. Preprint at Wiley. In The Handbook of Psycholinguistics, E.M. Fernández and H.S. Cairns, eds. (John Wiley & Sons), pp. 157–181. https://doi.org/10.1002/9781118829516.ch7.
5. Giglio, L., Ostarek, M., Weber, K., and Hagoort, P. (2022). Commonalities and asymmetries in the neurobiological infrastructure for language production and comprehension. Cereb. Cortex *32*, 1405–1418. https://doi.org/10.1093/cercor/bhab287.
6. Hu, J., Small, H., Kean, H., Takahashi, A., Zekelman, L., Kleinman, D., Ryan, E., Nieto-Castañón, A., Ferreira, V., and Fedorenko, E. (2023). Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. Cereb. Cortex *33*, 4384–4404. https://doi.org/10.1093/cercor/bhac350.
7. Awad, M., Warren, J.E., Scott, S.K., Turkheimer, F.E., and Wise, R.J.S. (2007). A common system for the comprehension and production of narrative speech. J. Neurosci. *27*, 11455–11464. https://doi.org/10.1523/JNEUROSCI.5257-06.2007.

8. Stephens, G.J., Silbert, L.J., and Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. Proc. Natl. Acad. Sci. USA. *107*, 14425–14430. https://doi.org/10.1073/pnas.1008662107.
9. Menenti, L., Gierhan, S.M.E., Segaert, K., and Hagoort, P. (2011). Shared language: overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. Psychol. Sci. *22*, 1173–1182. https://doi.org/10.1177/0956797611418347.
10. Segaert, K., Menenti, L., Weber, K., Petersson, K.M., and Hagoort, P. (2012). Shared syntax in language production and language comprehension–an FMRI study. Cereb. Cortex *22*, 1662–1670. https://doi.org/10.1093/cercor/bhr249.
11. Yarkoni, T., and Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspect. Psychol. Sci. *12*, 1100–1122. https://doi.org/10.1177/1745691617693393.
12. Zada, Z., Goldstein, A., Michelmann, S., Simony, E., Price, A., Hasenfratz, L., Barham, E., Zadbood, A., Doyle, W., Friedman, D., et al. (2024). A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. Neuron *112*, 3211–3222.e5. https://doi.org/10.1016/j.neuron.2024.06.025.
13. Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., and Mitchell, T. (2014). Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. PLoS One *9*, e112575. https://doi.org/10.1371/journal.pone.0112575.
14. Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., and Gallant, J.L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature *532*, 453–458. https://doi.org/10.1038/nature17637.
15. de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., and Theunissen, F.E. (2017). The Hierarchical Cortical Organization of Human Speech Processing. J. Neurosci. *37*, 6539–6557. https://doi.org/10.1523/JNEUROSCI.3267-16.2017.
16. Goldstein, A., Wang, H., Niekerken, L., Schain, M., Zada, Z., Aubrey, B., Sheffer, T., Nastase, S.A., Gazula, H., Singh, A., et al. (2025). A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. Nat. Hum. Behav. *9*, 1041–1055. https://doi.org/10.1038/s41562-025-02105-9.
17. Yamashita, M., Kubo, R., and Nishimoto, S. (2025). Conversational content is organized across multiple timescales in the brain. Nat. Hum. Behav. *9*, 2066–2078. https://doi.org/10.1038/s41562-025-02231-4.
18. Cai, J., Hadjinicolaou, A.E., Paulk, A.C., Soper, D.J., Xia, T., Wang, A.F., Rolston, J.D., Richardson, R.M., Williams, Z.M., and Cash, S.S. (2025). Natural language processing models reveal neural dynamics of human conversation. Nat. Commun. *16*, 3376. https://doi.org/10.1038/s41467-025-58620-w.
19. Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. Trans. Assoc. Comput. Linguist. *3*, 211–225. https://doi.org/10.1162/tacl_a_00134.
20. De Deyne, S., Navarro, D.J., Perfors, A., Brysbaert, M., and Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. Behav. Res. Methods *51*, 987–1006. https://doi.org/10.3758/s13428-018-1115-7.
21. Caucheteux, C., Gramfort, A., and King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. Nat. Hum. Behav. *7*, 430–441. https://doi.org/10.1038/s41562-022-01516-2.
22. Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S.A., Feder, A., Emanuel, D., Cohen, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. Nat. Neurosci. *25*, 369–380. https://doi.org/10.1038/s41593-022-01026-4.
23. Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., and De Lange, F.P. (2022). A hierarchy of linguistic predictions during natural language comprehension. Proc. Natl. Acad. Sci. USA. *119*, e2201968119. https://doi.org/10.1073/pnas.2201968119.
24. Schrimpf, M., Blank, I.A., Tuckute, G., Kauf, C., Hosseini, E.A., Kanwisher, N., Tenenbaum, J.B., and Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. Proc. Natl. Acad. Sci. USA. *118*, e2105646118. https://doi.org/10.1073/pnas.2105646118.
25. Redcay, E., and Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. Nat. Rev. Neurosci. *20*, 495–505. https://doi.org/10.1038/s41583-019-0179-4.
26. Grice, H.P. (1975). Logic and conversation. In Speech Acts, P. Cole and J.L. Morgan, eds. (BRILL), pp. 41–58. https://doi.org/10.1163/9789004368811_003.
27. Pickering, M.J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behav. Brain Sci. *27*, 169–190. ; discussion 190. https://doi.org/10.1017/S0140525X04000056.
28. Brennan, S.E., and Clark, H.H. (1996). Conceptual pacts and lexical choice in conversation. J. Exp. Psychol. Learn. Mem. Cogn. *22*, 1482–1493. https://doi.org/10.1037//0278-7393.22.6.1482.
29. Clark, H.H., and Brennan, S.E. (1991). Grounding in communication. In Perspectives on Socially Shared Cognition, L.B. Resnick, J.M. Levine, and S.D. Teasley, eds. (American Psychological Association), pp. 127–149. https://doi.org/10.1037/10096-006.
30. Wilkes-Gibbs, D., and Clark, H.H. (1992). Coordinating beliefs in conversation. J. Mem. Lang. *31*, 183–194. https://doi.org/10.1016/0749-596X(92)90010-U.
31. Clark, H.H. (1996). Using Language (Cambridge University Press). https://doi.org/10.1017/CBO9780511620539.
32. Nastase, S.A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. Neuroimage *222*, 117254. https://doi.org/10.1016/j.neuroimage.2020.117254.
33. Hasson, U., and Honey, C.J. (2012). Future trends in Neuroimaging: Neural processes as expressed within real-life contexts. Neuroimage *62*, 1272–1278. https://doi.org/10.1016/j.neuroimage.2012.02.004.
34. Jiang, J., Dai, B., Peng, D., Zhu, C., Liu, L., and Lu, C. (2012). Neural Synchronization during Face-to-Face Communication. J. Neurosci. *32*, 16064–16069. https://doi.org/10.1523/JNEUROSCI.2926-12.2012.
35. Kinreich, S., Djalovski, A., Kraus, L., Louzoun, Y., and Feldman, R. (2017). Brain-to-Brain Synchrony during Naturalistic Social Interactions. Sci. Rep. *7*, 17060. https://doi.org/10.1038/s41598-017-17339-5.
36. Zadbood, A., Chen, J., Leong, Y.C., Norman, K.A., and Hasson, U. (2017). How We Transmit Memories to Other Brains: Constructing Shared Neural Representations Via Communication. Cereb. Cortex *27*, 4988–5000. https://doi.org/10.1093/cercor/bhx202.
37. Liu, L., Li, H., Ren, Z., Zhou, Q., Zhang, Y., Lu, C., Qiu, J., Chen, H., and Ding, G. (2022). The "Two-Brain" Approach Reveals the Active Role of Task-Deactivated Default Mode Network in Speech Comprehension. Cereb. Cortex, bhab521. Cereb. Cortex *32*, 4869–4884. https://doi.org/10.1093/cercor/bhab521.
38. Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S.A., and Hasson, U. (2022). Teacher–student neural coupling during teaching and learning. Soc. Cogn. Affect. Neurosci. *17*, 367–376. https://doi.org/10.1093/scan/nsab103.
39. Chang, C.H.C., Nastase, S.A., Zadbood, A., and Hasson, U. (2024). How a speaker herds the audience: multibrain neural convergence over time during naturalistic storytelling. Soc. Cogn. Affect. Neurosci. *19*, nsae059. https://doi.org/10.1093/scan/nsae059.
40. Montague, P.R., Berns, G.S., Cohen, J.D., McClure, S.M., Pagnoni, G., Dhamala, M., Wiest, M.C., Karpov, I., King, R.D., Apple, N., et al. (2002). Hyperscanning: simultaneous fMRI during linked social interactions. Neuroimage *16*, 1159–1164. https://doi.org/10.1006/nimg.2002.1150.

41. Babiloni, F., and Astolfi, L. (2014). Social neuroscience and hyperscanning techniques: Past, present and future. Neurosci. Biobehav. Rev. *44*, 76–93. https://doi.org/10.1016/j.neubiorev.2012.07.006.

42. Wheatley, T., Boncz, A., Toni, I., and Stolk, A. (2019). Beyond the Isolated Brain: The Promise and Challenge of Interacting Minds. Neuron *103*, 186–188. https://doi.org/10.1016/j.neuron.2019.05.009.

43. Czeszumski, A., Eustergerling, S., Lang, A., Menrath, D., Gerstenberger, M., Schuberth, S., Schreiber, F., Rendon, Z.Z., and König, P. (2020). Hyperscanning: A Valid Method to Study Neural Inter-brain Underpinnings of Social Interaction. Front. Hum. Neurosci. *14*, 39. https://doi.org/10.3389/fnhum.2020.00039.

44. Nam, C.S., Choo, S., Huang, J., and Park, J. (2020). Brain-to-Brain Neural Synchrony During Social Interactions: A Systematic Review on Hyperscanning Studies. Appl. Sci. (Basel) *10*, 6669. https://doi.org/10.3390/app10196669.

45. Tsoi, L., Burns, S.M., Falk, E.B., and Tamir, D.I. (2022). The promises and pitfalls of functional magnetic resonance imaging hyperscanning for social interaction research. Soc. Pers. Psychol. Compass *16*, e12707. https://doi.org/10.1111/spc3.12707.

46. Speer, S.P.H., Mwilambwe-Tshilobo, L., Tsoi, L., Burns, S.M., Falk, E.B., and Tamir, D.I. (2024). Hyperscanning shows friends explore and strangers converge in conversation. Nat. Commun. *15*, 7781. https://doi.org/10.1038/s41467-024-51990-7.

47. Hasson, U., Ghazanfar, A.A., Galantucci, B., Garrod, S., and Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. Trends Cogn. Sci. *16*, 114–121. https://doi.org/10.1016/j.tics.2011.12.007.

48. Schoot, L., Hagoort, P., and Segaert, K. (2016). What can we learn from a two-brain approach to verbal interaction? Neurosci. Biobehav. Rev. *68*, 454–459. https://doi.org/10.1016/j.neubiorev.2016.06.009.

49. Brouwer, G.J., and Heeger, D.J. (2009). Decoding and reconstructing color from responses in human visual cortex. J. Neurosci. *29*, 13992–14003. https://doi.org/10.1523/JNEUROSCI.3577-09.2009.

50. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019) GPT-2 Language Models Are Unsupervised Multitask Learners. 24. OpenAI Technical Report. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

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

52. Nunez-Elizalde, A.O., Huth, A.G., and Gallant, J.L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. Neuroimage *197*, 482–492. https://doi.org/10.1016/j.neuroimage.2019.04.012.

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

54. Lescroart, M.D., Stansbury, D.E., and Gallant, J.L. (2015). Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. Front. Comput. Neurosci. *9*, 135. https://doi.org/10.3389/fncom.2015.00135.

55. Lee Masson, H., and Isik, L. (2021). Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. Neuroimage *245*, 118741. https://doi.org/10.1016/j.neuroimage.2021.118741.

56. Nieuwland, M.S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., et al. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. eLife *7*, e33468. https://doi.org/10.7554/eLife.33468.

57. Pickering, M.J., and Gambi, C. (2018). Predicting while comprehending language: A theory and review. Psychol. Bull. *144*, 1002–1044. https://doi.org/10.1037/bul0000158.

58. Pickering, M.J., and Strijkers, K. (2024). Language production and prediction in a parallel activation model. Top. Cogn. Sci. *17*, 936–947. https://doi.org/10.1111/tops.12775.

59. Toneva, M., Williams, J., Bollu, A., Dann, C., and Wehbe, L. (11–13 Apr 2022). Same Cause; Different Effects in the Brain. B. Schölkopf, C. Uhler, and K. Zhang, eds. *177*, 787–825.

60. Frith, C.D., and Frith, U. (1999). Interacting Minds–A Biological Basis. Science *286*, 1692–1695. https://doi.org/10.1126/science.286.5445.1692.

61. Frith, C.D., and Frith, U. (2021). Mapping Mentalising in the Brain. In The Neural Basis of Mentalizing, M. Gilead and K.N. Ochsner, eds. (Springer International Publishing), pp. 17–45. https://doi.org/10.1007/978-3-030-51890-5_2.

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

63. Deniz, F., Nunez-Elizalde, A.O., Huth, A.G., and Gallant, J.L. (2019). The Representation of Semantic Information Across Human Cerebral Cortex During Listening Versus Reading Is Invariant to Stimulus Modality. J. Neurosci. *39*, 7722–7736. https://doi.org/10.1523/JNEUROSCI.0675-19.2019.

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

65. Spiegelhalder, K., Ohlendorf, S., Regen, W., Feige, B., Tebartz Van Elst, L., Weiller, C., Hennig, J., Berger, M., and Tüscher, O. (2014). Interindividual synchronization of brain activity during live verbal communication. Behav. Brain Res. *258*, 75–79. https://doi.org/10.1016/j.bbr.2013.10.015.

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

67. Hagoort, P. (2019). The neurobiology of language beyond single-word processing. Science *366*, 55–58. https://doi.org/10.1126/science.aax0289.

68. Wheatley, T., Thornton, M.A., Stolk, A., and Chang, L.J. (2024). The Emerging Science of Interacting Minds. Perspect. Psychol. Sci. *19*, 355–373. https://doi.org/10.1177/17456916231200177.

69. Zaki, J., and Ochsner, K. (2009). The need for a cognitive neuroscience of naturalistic social cognition. Ann. N. Y. Acad. Sci. *1167*, 16–30. https://doi.org/10.1111/j.1749-6632.2009.04601.x.

70. Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., and Kanwisher, N. (2010). New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J. Neurophysiol. *104*, 1177–1194. https://doi.org/10.1152/jn.00032.2010.

71. Broca, P. (1865). Sur le siège de la faculté du langage articulé. Bull. Soc. Anthropol. Paris *6*, 377–393. https://doi.org/10.3406/bmsap.1865.9495.

72. Dax, M. (1865). Lesions de la motie gauche de l'encephale coincident avec l'oublie des signes de la pensee. Gaz Hbd Med Chir *2*, 259–262.

73. Wernicke C. Der Aphasische Symptomencomplex. In: Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis,Wernicke C., editor. Springer Berlin Heidelberg; 1974. p. 1–70. https://doi.org/10.1007/978-3-642-65950-8_1.

74. Knecht, S., Deppe, M., Dräger, B., Bobe, L., Lohmann, H., Ringelstein, E.-B., and Henningsen, H. (2000). Language lateralization in healthy right-handers. Brain *123*, 74–81. https://doi.org/10.1093/brain/123.1.74.

75. Corballis, M.C. (2014). Left brain, right brain: facts and fantasies. PLoS Biol. *12*, e1001767. https://doi.org/10.1371/journal.pbio.1001767.

76. Nastase, S.A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C.J., Yeshurun, Y., Regev, M., et al. (2021). The "Narratives" fMRI dataset for evaluating models of

naturalistic language comprehension. Sci. Data *8*, 250. https://doi.org/10.1038/s41597-021-01033-3.

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

78. Beeman, M. (1993). Semantic processing in the right hemisphere may contribute to drawing inferences from discourse. Brain Lang. *44*, 80–120. https://doi.org/10.1006/brln.1993.1006.

79. Beeman, M.J., and Chiarello, C. (1998). Complementary Right- and Left-Hemisphere Language Comprehension. Curr. Dir. Psychol. Sci. *7*, 2–8. https://doi.org/10.1111/1467-8721.ep11521805.

80. Kaplan, J.A., Brownell, H.H., Jacobs, J.R., and Gardner, H. (1990). The effects of right hemisphere damage on the pragmatic interpretation of conversational remarks. Brain Lang. *38*, 315–333. https://doi.org/10.1016/0093-934x(90)90117-y.

81. Heilman, K.M., Scholes, R., and Watson, R.T. (1975). Auditory affective agnosia. Disturbed comprehension of affective speech. J. Neurol. Neurosurg. Psychiatry *38*, 69–72. https://doi.org/10.1136/jnnp.38.1.69.

82. Lindell, A.K. (2006). In your right mind: right hemisphere contributions to language processing and production. Neuropsychol. Rev. *16*, 131–148. https://doi.org/10.1007/s11065-006-9011-9.

83. McNeely, H.E., and Parlow, S.E. (2001). Complementarity of linguistic and prosodic processes in the intact brain. Brain Lang. *79*, 473–481. https://doi.org/10.1006/brln.2001.2502.

84. Oderbolz, C., Poeppel, D., and Meyer, M. (2025). Asymmetric Sampling in time: Evidence and perspectives. Neurosci. Biobehav. Rev. *171*, 106082. https://doi.org/10.1016/j.neubiorev.2025.106082.

85. Bottini, G., Corcoran, R., Sterzi, R., Paulesu, E., Schenone, P., Scarpa, P., Frackowiak, R.S., and Frith, C.D. (1994). The role of the right hemisphere in the interpretation of figurative aspects of language. A positron emission tomography activation study. Brain *117*, 1241–1253. https://doi.org/10.1093/brain/117.6.1241.

86. Robertson, D.A., Gernsbacher, M.A., Guidotti, S.J., Robertson, R.R., Irwin, W., Mock, B.J., and Campana, M.E. (2000). Functional neuroanatomy of the cognitive process of mapping during discourse comprehension. Psychol. Sci. *11*, 255–260. https://doi.org/10.1111/1467-9280.00251.

87. Gernsbacher, M.A., and Kaschak, M.P. (2003). Neuroimaging studies of language production and comprehension. Annu. Rev. Psychol. *54*, 91–114. https://doi.org/10.1146/annurev.psych.54.101601.145128.

88. Vigneau, M., Beaucousin, V., Hervé, P.-Y., Jobard, G., Petit, L., Crivello, F., Mellet, E., Zago, L., Mazoyer, B., and Tzourio-Mazoyer, N. (2011). What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing? Insights from a meta-analysis. Neuroimage *54*, 577–593. https://doi.org/10.1016/j.neuroimage.2010.07.036.

89. Yarkoni, T., Speer, N.K., and Zacks, J.M. (2008). Neural substrates of narrative comprehension and memory. Neuroimage *41*, 1408–1425. https://doi.org/10.1016/j.neuroimage.2008.03.062.

90. Saxe, R. (2006). Uniquely human social cognition. Curr. Opin. Neurobiol. *16*, 235–239. https://doi.org/10.1016/j.conb.2006.03.001.

91. Frith, C.D., and Frith, U. (2012). Mechanisms of social cognition. Annu. Rev. Psychol. *63*, 287–313. https://doi.org/10.1146/annurev-psych-120710-100449.

92. Levinson, S.C., and Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. Front. Psychol. *6*, 731. https://doi.org/10.3389/fpsyg.2015.00731.

93. Pulvermüller, F., and Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. Nat. Rev. Neurosci. *11*, 351–360. https://doi.org/10.1038/nrn2811.

94. Mahon, B.Z., and Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. J. Physiol. Paris *102*, 59–70. https://doi.org/10.1016/j.jphysparis.2008.03.004.

95. Dupré La Tour, T., and Di, V., Oleggio Castello, M., and Gallant, J.L. (2024). The Voxelwise Modeling framework: a tutorial introduction to fitting encoding models to fMRI data. Preprint. https://doi.org/10.31234/osf.io/t975e,.

96. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-Art Natural Language Processing (Association for Computational Linguistics). https://doi.org/10.18653/v1/2020.emnlp-demos.6.

97. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

98. Bain, M., Huh, J., Han, T., and Zisserman, A. (2023). WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In INTERSPEECH 2023 (ISCA), pp. 4489–4493. https://doi.org/10.21437/Interspeech.2023-78.

99. 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., et al. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods *16*, 111–116. https://doi.org/10.1038/s41592-018-0235-4.

100. Gorgolewski, K., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., and Ghosh, S.S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. Front. Neuroinform. *5*, 13. https://doi.org/10.3389/fninf.2011.00013.

101. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. Front. Neuroinform. *8*, 14. https://doi.org/10.3389/fninf.2014.00014.

102. Aron, A., Melinat, E., Aron, E.N., Vallone, R.D., and Bator, R.J. (1997). The Experimental Generation of Interpersonal Closeness: A Procedure and Some Preliminary Findings. Pers. Soc. Psychol. Bull. *23*, 363–377. https://doi.org/10.1177/0146167297234003.

103. Esteban, O., Blair, R., Markiewicz, C.J., Berleant, S.L., Moodie, C., Ma, F., Isik, A.I., Erramuzpe, A., Kent, M., James, D., et al. (2018). fMRIPrep. Software. https://doi.org/10.5281/zenodo.852659.

104. Gorgolewski, K.J., Esteban, O., Markiewicz, C.J., Ziegler, E., Ellis, D.G., Notter, M.P., Jarecka, D., Johnson, H., Burns, C., Manhães-Savio, A., et al. (2018). Nipype. Software. https://doi.org/10.5281/zenodo.596855.

105. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., and Gee, J.C. (2010). N4ITK: Improved N3 Bias Correction. IEEE Trans. Med. Imaging *29*, 1310–1320. https://doi.org/10.1109/TMI.2010.2046908.

106. Avants, B.B., Epstein, C.L., Grossman, M., and Gee, J.C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. *12*, 26–41. https://doi.org/10.1016/j.media.2007.06.004.

107. Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging *20*, 45–57. https://doi.org/10.1109/42.906424.

108. Dale, A.M., Fischl, B., and Sereno, M.I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. NeuroImage *9*, 179–194. https://doi.org/10.1006/nimg.1998.0395.

109. Klein, A., Ghosh, S.S., Bao, F.S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Chaibub Neto, E.C., et al. (2017). Mindboggling morphometry of human brains. PLoS Comput. Biol. *13*, e1005350. https://doi.org/10.1371/journal.pcbi.1005350.

110. Fischl, B., Sereno, M.I., Tootell, R.B.H., and Dale, A.M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical

surface. Hum. Brain Mapp. *8*, 272–284. https://doi.org/10.1002/(SICI)1097-0193(1999)8:4<272::AID-HBM10>3.0.CO;2-4.

111. Huntenburg, J.M. (2014). Evaluating Nonlinear coregistration of BOLD EPI and T1w Images (PhD Thesis (Freie Universität Berlin)).

112. Wang, S., Peterson, D.J., Gatenby, J.C., Li, W., Grabowski, T.J., and Madhyastha, T.M. (2017). Evaluation of Field Map and Nonlinear Registration Methods for Correction of Susceptibility Artifacts in Diffusion MRI. Front. Neuroinform. *11*, 17. https://doi.org/10.3389/fninf.2017.00017.

113. Treiber, J.M., White, N.S., Steed, T.C., Bartsch, H., Holland, D., Farid, N., McDonald, C.R., Carter, B.S., Dale, A.M., and Chen, C.C. (2016). Characterization and Correction of Geometric Distortions in 814 Diffusion Weighted Images. PLoS One *11*, e0152472. https://doi.org/10.1371/journal.pone.0152472.

114. Greve, D.N., and Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. NeuroImage *48*, 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060.

115. Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. NeuroImage *17*, 825–841. https://doi.org/10.1016/s1053-8119(02)91132-8.

116. Cox, R.W., and Hyde, J.S. (1997). Software tools for analysis and visualization of fMRI data. NMR Biomed. *10*, 171–178. https://doi.org/10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L.

117. Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., and Petersen, S.E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. NeuroImage *84*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048.

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

119. Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S.J., and Turner, R. (1996). Movement-Related effects in fMRI time-series. Magn. Reson. Med. *35*, 346–355. https://doi.org/10.1002/mrm.1910350312.

120. Satterthwaite, T.D., Elliott, M.A., Gerraty, R.T., Ruparel, K., Loughead, J., Calkins, M.E., Eickhoff, S.B., Hakonarson, H., Gur, R.C., Gur, R.E., et al. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. NeuroImage *64*, 240–256. https://doi.org/10.1016/j.neuroimage.2012.08.052.

121. Ciric, R., Wolf, D.H., Power, J.D., Roalf, D.R., Baum, G.L., Ruparel, K., Shinohara, R.T., Elliott, M.A., Eickhoff, S.B., Davatzikos, C., et al. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. Neuroimage *154*, 174–187. https://doi.org/10.1016/j.neuroimage.2017.03.020.

122. Parkes, L., Fulcher, B., Yücel, M., and Fornito, A. (2018). An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. Neuroimage *171*, 415–436. https://doi.org/10.1016/j.neuroimage.2017.12.073.

123. Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., et al. (2016). A multi-modal parcellation of human cerebral cortex. Nature *536*, 171–178. https://doi.org/10.1038/nature18933.

124. Lipkin, B., Tuckute, G., Affourtit, J., Small, H., Mineroff, Z., Kean, H., Jouravlev, O., Rakocevic, L., Pritchett, B., Siegelman, M., et al. (2022). Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. Sci. Data *9*, 529. https://doi.org/10.1038/s41597-022-01645-3.

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

126. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

127. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

128. Gale, D.J., Vos de Wael, R., Benkarim, O., and Bernhardt, B. (2021) Surfplot: Publication-Ready Brain Surface Figures (Zenodo). https://doi.org/10.5281/zenodo.5942281

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
fMRI data	NIMH Data Archive	10.15154/w7y8-yd67; NIMH: 4349
Software and algorithms		
Original code	This paper	10.5281/zenodo.17468195
Himalaya	Dupré La Tour et al.[95]	10.1016/j.neuroimage.2022.119728
HuggingFace Transformers	Wolf et al.[96]	10.18653/v1/2020.emnlp-demos.6; RRID: SCR_027381
SciPy	Virtanen et al.[97]	10.1038/s41592-019-0686-2; RRID: SCR_008058
WhisperX	Bain et al.[98]	https://github.com/m-bain/whisperX
fMRIPrep 20.2.0	Esteban et al.[99]	10.1038/s41592-018-0235-4; RRID: SCR_016216
Nipype 1.5.1	Gorgolewski et al.[100]	10.3389/fninf.2011.00013; RRID: SCR_002502
Nilearn 0.6.2	Abraham et al.[101]	10.3389/fninf.2014.00014; RRID: SCR_001362

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Participants

Thirty dyads (*N* = 60 participants) engaged in real-time conversations while they were simultaneously scanned with fMRI hyperscanning. These data are a subset of a larger dataset collected with additional conditions and participants.[46] Participants were recruited from Princeton University and received monetary compensation for their participation. All participants provided informed consent in a manner approved by the Princeton University Institutional Review Board. Eligibility requirements included: must be 18 years or older, right-handed, and with normal or corrected vision. Of the 58 included participants, 41 were female, and the average age was 20.74 (minimum 18, maximum 36). Participants reported their race or ethnicity as African American = 8, Asian = 19, Caucasian = 26, Other = 5. Sex and gender were not considered in any analyses because we did not expect any differences in encoding performance or brain-to-brain coupling based on these variables. One dyad was excluded due to an unexpected scanning issue that resulted in fewer conversations than others.

METHOD DETAILS

Design

Two participants at a time arrived at two fMRI scanners in adjacent rooms. The participants did not know each other before the experiment but briefly met before entering the scanners. Participants were instructed to engage in prompted conversations across five runs. Prompts were specifically designed to increase the level of intimacy of conversations across the runs, and are based on stimuli from Aron and colleagues[102] (Table S1). Each run was 13:36 minutes long and consisted of four trials. We only used two trials of each run because the other two trials were not spontaneous conversations, and were used for a different experiment. Each trial was 03:21 minutes long and started with the topic prompt displayed on screen for 9 seconds, followed by the conversation for 180 seconds, and ended with 12 seconds of a fixation cross (Figure 1). The participant who would start speaking first was randomly assigned. Once a participant finished their utterance, they were instructed to press a button to "pass the virtual mic" to their conversational partner. When a participant had the virtual mic, the screen displayed the text "your turn to speak, when you want to pass the mic, press '1'", followed by a countdown timer displaying the number of seconds left. When listening, the screen showed "your turn to listen", followed by the same countdown timer. Participants were instructed to fill the entire three minutes. After all runs, participants filled out a survey answering questions about the level of enjoyment, similarity, and closeness they felt during their conversations.

To get a sense of the language used in the conversations, we collated the part-of-speech category across all words (Table S2). We found that pronouns are most common, followed by verbs and nouns. We also aimed to estimate repeated words within conversations, specifically between utterances. We used the ROGUE NLP metric to compute unigram and bi-grams similarity between each pair of utterances. The score is normalized between 0 and 1, with higher scores indicating higher similarity (overlap) between utterances. We averaged the ROGUE score across 1,551 pairs of utterances. For unigrams, we obtained a score of M=0.244, std=0.148; and for bigrams, we obtained a score of M=0.055, std=0.051. This suggests that while speakers sometimes use the same words, they rarely use them in the same bigrams. Given the low likelihood of participants using the same words in the same context when responding to their partner, we don't expect this to influence brain-to-brain coupling.

MRI acquisition

We recorded neuroimaging data using 3T Siemens Skyra and 3T Siemens Prisma MRI systems. Both machines were configured using the same scanning parameters. Functional scans were acquired with whole brain coverage in interleaved order: 3.0 mm slice thickness, 3.0 × 3.0 mm in-plane resolution, flip angle = 80°, TE = 28 ms, TR = 1500 ms, multiband acceleration factor = 2. A T1-weighted image was acquired for anatomical reference: 1.0 ×.0 × 1.0 mm resolution, 176 sagittal slices, flip angle = 9°, TE = 2.98 ms, TR = 2300 ms. To minimize head movement, the subjects' heads were stabilized with foam padding.

Conversation audio transcription

Each three-minute audio segment was transcribed, aligned, and diarized (assigned unique speaker labels) at the word level using WhisperX[98]—an automatic speech recognition tool. We used the *faster-whisper-large-v2* model and set the minimum and maximum speakers to two. Each resulting transcription consisted of each word spoken, its onset and duration, and the identity of the speaker.

fMRIPrep preprocessing

Results included in this manuscript come from preprocessing performed using *fMRIPrep* 20.2.0,[99,103] which is based on *Nipype* 1.5.1[100,104] and *Nilearn* 0.6.2.[101]

T1-weighted images were corrected for intensity non-uniformity (INU) with *N4BiasFieldCorrection*,[105] distributed with ANTs 2.3.3,[106] and used as a reference throughout the workflow. The T1 reference was then skull-stripped with a *Nipype* implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1 image using *fast* (FSL 5.0.9[107]). Brain surfaces were reconstructed using *recon-all* (FreeSurfer 6.0.1[108]), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle.[109] Individual cortical surface reconstructions were aligned to the *fsaverage6* surface template (40,962 vertices per hemisphere) based on sulcal curvature.[110]

Functional data preprocessing

For each of the 6 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated. A deformation field to correct for susceptibility distortions was estimated based on *fMRIPrep*'s *fieldmap-less* approach. The deformation field is constructed by co-registering the BOLD reference to the same-subject T1 reference with inverted intensity.[111,112] Registration is performed with *antsRegistration* (ANTs 2.3.3), and the process is regularized by constraining deformation to be nonzero only along the phase-encoding direction, and modulated with an average fieldmap template.[113] Based on the estimated susceptibility distortion, a corrected BOLD reference was calculated for a more accurate co-registration with the anatomical reference.

The BOLD reference was then co-registered to the T1w reference using FreeSurfer's *bbregister*, which implements boundary-based registration.[114] Co-registration was configured with six degrees of freedom. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) were estimated before any spatiotemporal filtering using *mcflirt* $^{\text{FSL 5.0.9}}$ [115]. BOLD runs were slice-time corrected using *3dTshift* from AFNI 20160207.[116] The BOLD time series were ultimately resampled onto the *fsaverage6* surface template using FreeSurfer's *mri_vol2surf*. Resampling was performed with a single interpolation step by applying a single, composite transform to correct for head motion, slice-timing, susceptibility distortions, and normalization to the surface template. All subsequent analyses were applied to the vertex-level functional data in surface space; our use of the term "vertex" is otherwise synonymous with the use of "voxel" in volumetric analyses (e.g., "voxelwise encoding models").

Several confounding time series were calculated while preprocessing the BOLD data: six head motion parameters, framewise displacement (FD), and a set of physiological components. FD was estimated for each functional run by computing the absolute sum of relative motions.[117] FD was calculated for each functional run using the implementation in *Nipype*.[117] The three global signals are extracted within the CSF, the white matter, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for anatomically constrained component-based noise correction (aCompCor[118]). Principal components are estimated after high-pass filtering the preprocessed BOLD time series using a discrete cosine filter with 128s cut-off. We retained 10 aCompCor components, five estimated from a white matter mask, and five from a CSF mask.

Confound and head motion correction

A typical fMRI signal cleaning pipeline involves regressing out nuisance variables from fMRIPrep's output from the BOLD signal across an entire run or scan.[119–122] Nuisance variables include head motion (e.g., rigid-body motion parameters), physiological noise (e.g., cardiac fluctuations), and scanner noise (e.g., signal drift). However, our hyperscanning paradigm with freely alternating speech production and comprehension between subjects requires additional task-related nuisance variables.

From fMRIPrep confounds, we chose the six head motion variables, all available cosine variables, and the top five components from aCompCor for white matter and CSF masks, separately. This resulted in 26 nuisance regressors. Next, we added five regressors based on the task structure (see the previous Design section). Three boxcar regressors were initialized with zeros across the entire run and populated with ones for (1) indicating the two different trial types, (2) indicating turn to speak, and (3) indicating turn to listen. Two indicator regressors were initialized with zeros and filled with ones when either (1) the subject pressed the button to end their turn, or (2) their conversation partner pressed the button (the instructions on the screen switched each time a button was pressed). These regressors were convolved with an HRF to account for the hemodynamic response using Nilearn's *glm.first_level.glover_hrf* implementation. Finally, all confound variables were passed to Nilearn's *signal.clean* function to detrend, regress out the variables, and z-score the time series.

Defining cortical regions of interest

In order to summarize results across the cortex, we first aggregated the 40,962 vertices in each hemisphere into 180 parcels from a widely-used Glasser multimodal parcellation.[123] Then, we defined an extended parcel-level language network from four primary sources: a collection of functionally defined language regions,[70] a probabilistic atlas based on language localizer tasks in 806 subjects,[124] an activation map corresponding to the "language" topic from NeuroSynth,[125] and an intersubject correlation map (ISC) based on 345 subjects listening to natural stories.[76] We thresholded the probabilistic atlas at p=0.10, the NeuroSynth map at t=0.10, and the intersubject map at r=0.10. We overlaid these four maps to form an extended "meta" map of language areas (Figure S5A).

We grouped the 55 parcels within this final brain map into 11 regions of interest based on their spatial proximity and previously identified groupings (Figure S5B; Table S3). Specifically, following the networks identified by Glasser and colleagues,[123] we identified the following regions: early auditory cortex (EAC), posterior and anterior superior temporal gyrus (pSTG, aSTG), inferior and middle frontal gyri (IFG, MFG), somatomotor cortex (SM), supplementary motor area (SMA), frontal operculum (FOP), intraparietal sulcus (IPS), temporoparietal junction (TPJ), and posterior medial cortex (PMC). Finally, given that the maps derived from prior studies may be biased toward comprehension tasks, we defined a somatomotor region of interest we expect to be involved in language production.[3] Note that we are deliberately defining a more inclusive "language network" than prior work[70] to explore both more peripheral perception (e.g., EAC) and production (e.g., SM) areas, as well as higher-level areas that may be involved in narrative and social cognition (e.g., TPJ, PMC).

Linguistic features for encoding analysis

In vertex-wise encoding analysis, we use ridge regression to learn a linear model mapping from a set of explicit features (i.e., design matrix) to the observed brain activity.[51] We first re-represent the language task and stimulus in one or more feature spaces. We defined several feature spaces from the conversation stimuli to build these design matrices.

Task structure and nuisance variables

We computed four low-level variables from each transcript that could affect the BOLD signal.[14] For each TR, we quantify the word rate (number of words in a TR), phoneme rate (number of phonemes in a TR), word occurrence (some TRs contained no words), and a variable indicating whether it was the subject's turn to speak or listen. The word and phoneme rates were continuous, while the word onset and indicator variables were binary.

Acoustic spectral features

For each pair of subjects, we had one audio recording of the entire conversation that was recorded from one mic at a time and switched upon button presses indicating the end of turn. We computed acoustic features from the speech audio files.[15] Specifically, we used the *WhisperFeatureExtractor* class from the *HuggingFace*[96] library with the default settings to extract a spectral representation of the audio. This function uses a short-time Fourier transform to compute a mel-filter bank of 80 features that represent the spectral power density on a Mel log scale. Note that these features likely capture more than just acoustic features because they were recorded in MRI machines with different noise characteristics, and were saved into one file from two sources. Thus, at minimum, it also encodes information about the conversation turns.

Articulatory phonemic features

Following de Heer and colleagues,[15] we quantify the articulatory features of speech based on the phonemes in the transcript. Specifically, we used the CMU pronunciation dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict) to obtain the phonemes associated with each word in the transcript. We then constructed the articulatory features for each phoneme based on the place and manner of consonants, and voicing of vowels. This resulted in a binary vector of 22 features for each phoneme.

Large language model features

We extracted word embeddings from the large language model GPT-2 XL[50] using the *HuggingFace* library.[96] For each 3-minute conversation transcript, we first converted all words to GPT-2 tokens. We then passed these tokens as input to the LLM using the full context window (1024 tokens; the context increases as more tokens are available later in the conversation), where they were converted to 1,600-dimensional token embeddings and passed through the decoder layers. We extracted the activations from the middle (24th) layer to serve as contextual word embeddings.

QUANTIFICATIONS AND STATISTICAL ANALYSIS

Encoding model construction and evaluation

Encoding models were the core analytical approach we took to estimating linguistic content in the BOLD signal.[51] For all analyses, we used kernel ridge regression to prevent overfitting, and banded ridge regression to find different regularization parameters for each feature space separately.[52] Banded-ridge regression has an in-built ‘‘feature-space selection mechanism’’ that allows it to apply different regularization penalties to each feature space, thus selecting some submodels while suppressing others[53] (e.g., redundant spaces). In the case of highly correlated feature spaces, the encoding model may split the variance between them, or assign variance to only one. We used variance decomposition to quantify the *relative importance* of each feature space. This has been shown to reduce spurious correlations between features spaces and stimulus correlations.[52] We used the *MultipleKernelRidgeCV* class from the *himalaya* library[53] to perform cross-validation within the training set to select the best regularization parameter per feature space. All results we report on encoding performance were evaluated on a held-out test sample.

Design matrix construction

Each 3-minute conversation (trial) consisted of a 120-TR BOLD time series. With two trials per run (240 TRs) and five total runs, we had a total of 1,200 TRs per subject. Thus, our design matrix had 1,200 rows. The initial number of columns was based on the selected feature spaces for each analysis. For example, for the full joint model (Figure 1), we used five feature spaces: task (8 dimensions), acoustic (80), articulatory (22), and contextual embeddings (1,600). Stimuli features that were defined on the word or token level were averaged within TRs (e.g., LLM embeddings). Then, we split each feature space into two groups, for production or comprehension, and filled the gaps between one process and the other with zeros.

Model definition

We used a Scikit-learn[126] pipeline to define the full encoding model. The pipeline consisted of three main steps before model fitting. First, the regressors were mean-centered using *StandardScaler*. Then, each feature space was duplicated and shifted by 2–5 TRs (3–7.5 s) to account for the hemodynamic lag in the BOLD signal.[14] Finally, because the design matrix was wider than it is long, we used the kernel method to solve the ridge regression in its dual form.[95] Specifically, we used a linear kernel for each feature space separately before fitting the model.

Model fitting and evaluation

We used cross-validation to evaluate each model on a held-out test sample. Specifically, we defined five folds, based on the five runs, to fit a model on four runs (960 TRs), and tested it on the held-out run (240 TRs). We repeated this procedure five times, testing each run in turn, and then averaging the encoding performance across the five runs. Each run contained unique conversations based on different prompts.

Banded ridge regression allows us to perform variance decomposition: evaluating each feature space separately relative to all the others. To do this, the joint predicted time series on the held-out run can be decomposed into one time series per feature space.[53] Similarly, the encoding model performance (i.e., the correlation between the predicted and actual time series) can be split into one correlation for each feature space. Importantly, we segmented the actual and predicted time series into production and comprehension TRs to obtain their separate correlations for each process. Moreover, because of the hemodynamic response, some TRs may be affected by both processes. Thus, we selected the exclusive set of TRs where there is no overlap. To evaluate the *unique* (as opposed to relative) variance accounted for by LLM word embeddings, we used variance partitioning. In this analysis, nested regression models are fit, one with the feature of interest (LLM embedding) and one without. By comparing the performance of both models, we calculate the unique variance explained by that feature. If a model L is fully collinear with other models, then the unique variance it explains, U_L, will be 0.

Finally, we confirmed that head motion degrades encoding performance and that there is considerably more head motion during speech production than comprehension (Figure S6).

Statistical significance

We tested whether a vertex’s encoding performance correlation is statistically significant by using a two-sided, one-sample *t*-test, as implemented in SciPy.[97] All p-values were corrected for multiple comparisons by controlling the false discovery rate (FDR[127]).

Speaker–listener model-based coupling

We used the already-trained encoding models to evaluate the model-based coupling between conversation partners. The intuition behind this evaluation is to correlate one subject’s model-predicted time series with their conversational partner’s actual time series (as opposed to correlating it with their *own* actual time series). In effect, this simultaneously tests whether the model can generalize from one subject to another and from one process to another (e.g., production to comprehension).[59] Thus, we use the same evaluation procedure as described before, except with one major change. For each voxel, we correlate a subject’s predicted time series with their partner’s actual time series for the same voxel. Critically, we use the predictions from all feature spaces and compute the relative encoding performance of the LLM contextual embedding feature space only. By applying the same evaluation procedure as within-subject, we control for variance that can be explained by the nuisance feature spaces. When testing model-based coupling across regions and time (Figure 6), we first extract speaker turns that are at least 9 seconds long in order to exclude turns that are too short.

Story-listening task and analysis

Prior to hyperscanning acquisition, participants listened to a ∼13-minute story (“I Knew You Were Black” by Carol Daniel). Three participants did not complete this task and were excluded from this particular analysis. We used the same procedures as described above for conversations for the story, including MRI acquisition parameters, BOLD preprocessing, confound regression, linguistic features, and encoding model construction, training, and evaluation. However, there were two differences. First, the confound regression did not include any design structure variables. Second, we did not split regressors because the story is only comprehension—thus this model corresponds to the shared-weights model for conversations. We conducted an additional analysis to rule out the possibility that the RH lateralization is due to an idiosyncratic factor related to our scanning protocol or processing pipeline. We computed ISC between each pair of participants using the story listening task (although this task has no inherent pairing of participants, we used the dyads in conversations to better relate to our brain-to-brain coupling analysis). We found that the ISC during story listening is not lateralized to the RH. Since this data is acquired and preprocessed using the same processes as the conversation fMRI data, we are more confident that the RH lateralization we observed is more likely due to the interactive conversation paradigm (Figure S7).

Software resources

In addition to the software mentioned throughout the Methods, we used *Surfplot*[128] for visualizing brain maps.