Neuron

A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations

Highlights

- We acquired intracranial recordings in five dyads during face-to-face conversations
- Large language models can serve as a shared linguistic space for communication
- Context-sensitive embeddings track the exchange of information from brain to brain
- Contextual embeddings outperform other models for speaker-listener coupling

Authors

Zaid Zada, Ariel Goldstein, Sebastian Michelmann, ..., Orrin Devinsky, Samuel A. Nastase, Uri Hasson

Correspondence

zzada@princeton.edu

In brief

Zada et al. use contextual embeddings from large language models to capture linguistic information transmitted from the speaker's brain to the listener's brain in real-time, dyadic conversations.

Zada et al., 2024, Neuron *112*, 3211–3222
September 25, 2024
https://doi.org/10.1016/j.neuron.2024.06.025

CellPress

Neuron

CellPress
OPEN ACCESS

Article

A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations

Zaid Zada,[1,8,*] Ariel Goldstein,[1,2] Sebastian Michelmann,[1] Erez Simony,[1,3] Amy Price,[1] Liat Hasenfratz,[1] Emily Barham,[1] Asieh Zadbood,[1,4] Werner Doyle,[5] Daniel Friedman,[5] Patricia Dugan,[5] Lucia Melloni,[5] Sasha Devore,[5] Adeen Flinker,[5,6] Orrin Devinsky,[5] Samuel A. Nastase,[1,7] and Uri Hasson[1,7]

[1]Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ 08544, USA
[2]Department of Cognitive and Brain Sciences and Business School, Hebrew University, Jerusalem 9190501, Israel
[3]Faculty of Engineering, Holon Institute of Technology, Holon 5810201, Israel
[4]Department of Psychology, Columbia University, New York, NY 10027, USA
[5]Grossman School of Medicine, New York University, New York, NY 10016, USA
[6]Tandon School of Engineering, New York University, New York, NY 10016, USA
[7]These authors contributed equally
[8]Lead contact
*Correspondence: zzada@princeton.edu
https://doi.org/10.1016/j.neuron.2024.06.025

SUMMARY

Effective communication hinges on a mutual understanding of word meaning in different contexts. We recorded brain activity using electrocorticography during spontaneous, face-to-face conversations in five pairs of epilepsy patients. We developed a model-based coupling framework that aligns brain activity in both speaker and listener to a shared embedding space from a large language model (LLM). The context-sensitive LLM embeddings allow us to track the exchange of linguistic information, word by word, from one brain to another in natural conversations. Linguistic content emerges in the speaker's brain before word articulation and rapidly re-emerges in the listener's brain after word articulation. The contextual embeddings better capture word-by-word neural alignment between speaker and listener than syntactic and articulatory models. Our findings indicate that the contextual embeddings learned by LLMs can serve as an explicit numerical model of the shared, context-rich meaning space humans use to communicate their thoughts to one another.

INTRODUCTION

Language is the bedrock of human communication, allowing us to share our ideas and feelings with one another. Successful communication, however, relies on a shared agreement on the meaning of words *in context*. For example, the word *cold* can describe the temperature, a personality trait, or a viral infection, depending on the context. This contextual meaning of language resides in a shared space between people in a community of speakers: words absorb transient, agreed-upon meanings specific to their use and context.[1,2] Without a shared agreement, it would be impossible for strangers to understand one another. For example, speakers can only understand whether the word *cold* in the sentence "you're as cold as ice" refers to a personality trait or physical temperature if they understand the conversational context.

Until recently, we lacked a precise computational framework for modeling how humans use words in context as we communicate with others. To overcome this limitation, prior work has used data-driven, *unmediated coupling* methods, such as intersubject correlation (ISC), which leverage neural activity in one brain to model neural activity in another brain.[3–5] ISC analyses revealed that during natural communication, the listener's brain activity is coupled with the speaker's brain activity, and the strength of brain-to-brain coupling is proportional to the quality of communication.[6–12] ISC analysis, however, is *content-agnostic*: ISC can be driven by *any* shared signals across subjects and cannot tell us explicitly *what* features are aligned across brains—specifically, ISC cannot tell us what context-specific linguistic information is shared between speakers during real-world conversations. For example, in a face-to-face conversation, coupled neural activity across speakers is influenced not only by the words spoken but also by factors such as intonation, prosody, gestures, facial expressions, eye gaze, and other nonverbal aspects of social communication.

A new class of large language models (LLMs) has recently emerged that, for the first time, respects the richness of context in natural communication. Remarkably, these models learn from much the same shared space as humans: from real-world language generated by humans. LLMs rely on a simple, self-supervised objective (e.g., next-word prediction) to learn to produce context-specific linguistic outputs from real-world text

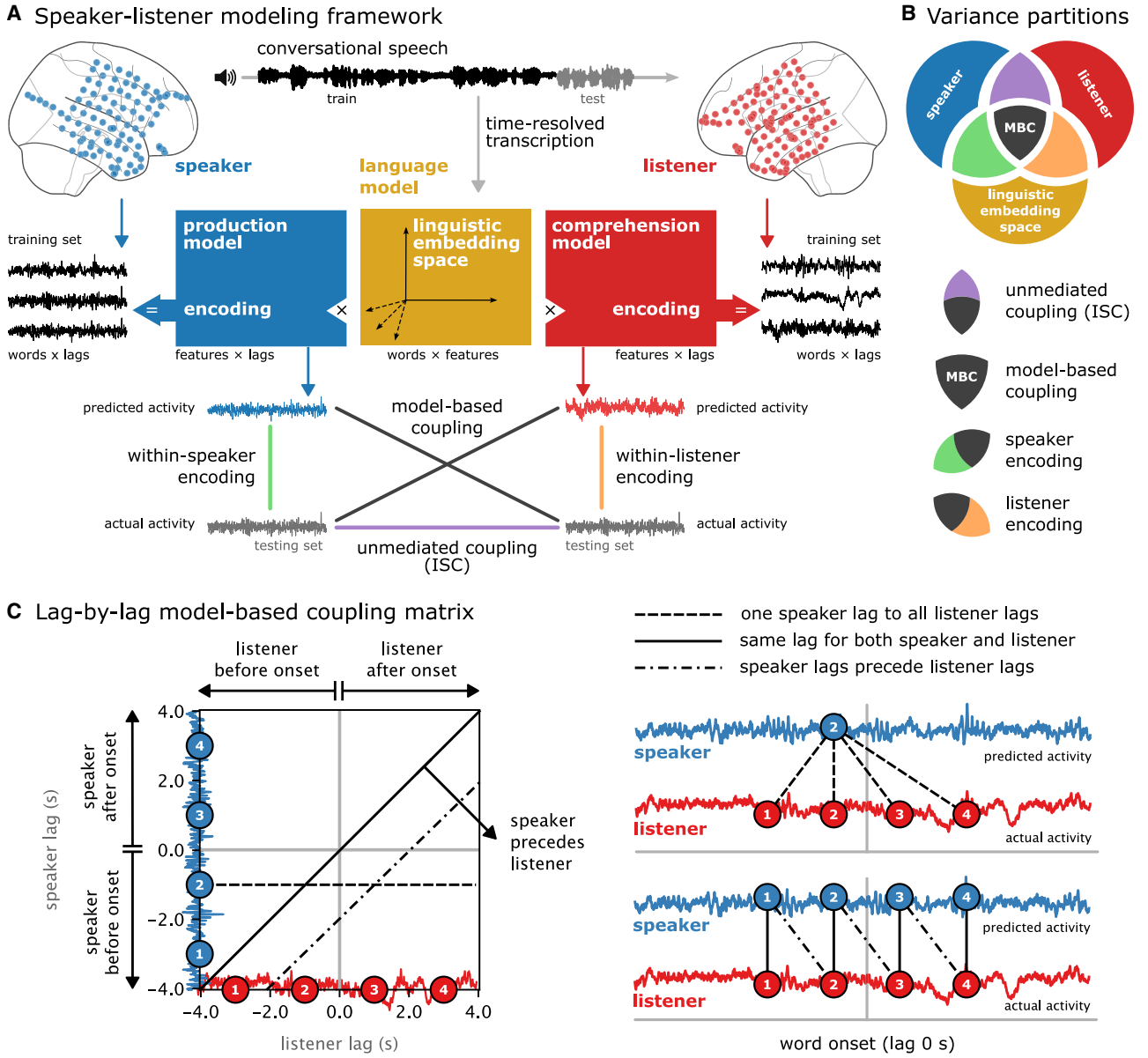

Figure 1. Encoding models for capturing speaker-listener linguistic coupling

(A) Schematic depicting how encoding models isolate the linguistic content of speaker-listener brain-to-brain coupling. Word-level neural signals for both patients are collated into speaker-listener roles and split into temporally contiguous train/test folds for cross-validation, along with their respective word embeddings. The word-by-word embeddings (yellow) populate a shared multidimensional context-dependent feature space for modeling the linguistic content encoded in both speaker and listener neural activity. Electrode-wise encoding models are estimated using ridge regression separately for the speaker (blue weight matrix) and listener (red weight matrix) to predict neural activity from the LLM embeddings at varying lags relative to word onset. Within-subject encoding performance is quantified separately for the speaker and listener by correlating the model-predicted and actual neural activity for left-out test segments of each conversation. Model-based speaker-listener coupling is quantified as the correlation between the model-predicted activity derived from the model trained on the speaker (or listener) and the actual neural activity of the listener (or speaker). Although unmediated, content-agnostic coupling methods like intersubject correlation (ISC) directly correlate brain activity between speaker and listener (purple line), our model-based coupling method effectively isolates the linguistic component of brain-to-brain coupling (black lines). That is, model-based analysis ensures that neural activities in the speaker and listener are aligned to a shared set of explicit linguistic features.

(B) Components of speaker-listener brain-to-brain coupling. Each circle represents the variance due to one of three sources: the speaker's brain activity (blue), the listener's brain activity (red), and the linguistic embedding space (yellow). Labeled intersections below represent partitions of shared variance: unmediated intersubject correlation capturing shared variance between the speaker and listener (purple) as well as shared variance between the contextual embedding space and the speaker's (green) and listener's (orange) brain activity. The central intersection of all three sources (black) represents the shared variance in speaker-listener brain-to-brain coupling captured by the contextual embeddings, i.e., model-based coupling.

(legend continued on next page)

corpora—and, in the process, implicitly encode the statistical structure of natural language into a multidimensional embedding space.[13–15] For instance, LLMs will produce different representations—i.e., contextual embeddings—for different uses of the word *cold* based on the preceding context. The capacity of these models to generate fluent, context-aware text, engage in dialogue, and meaningfully answer questions is a testament to just how much can be learned from the shared space of language and communication.[16] Interestingly, recent studies have suggested that LLMs and the brain converge on shared computational principles for natural language comprehension.[17–21]

In this study, we asked whether LLMs can provide an explicit numerical model for how context-dependent information is shared across brains during natural communication. We recorded cortical activity using electrocorticography (ECoG) in five dyadic pairs of epilepsy patients during spontaneous, interactive conversations. Modeling speech production and comprehension in this free-form setting is challenging, as each conversation has its unique trajectory, context, and dynamic. Furthermore, half the words used (50.11%) only appear once within a conversation; even words that appear more than once do not appear twice in the exact same context. LLMs, like humans, can interpret the contextual meaning of words embedded in real-world conversations. We hypothesized that this capacity can position LLM embeddings as an explicit model of the shared linguistic space by which speaker and listener communicate their thoughts—i.e., transmit their brain activity—to one another in natural conversations.

To model the transfer of linguistic information across brains, we extracted contextual embeddings for each word in the conversation from a widely used contextual language model, GPT-2.[22] Using the same set of contextual embeddings, we trained encoding models to predict brain activity during both speech production and comprehension in held-out segments of the conversations. We first demonstrate that contextual embeddings can predict the neural activity for each word as it is articulated in each unique conversation across the cortical language network during speech comprehension and production. Consistent with the flow of information during communication, our model-based coupling analyses demonstrate that the same "linguistic content" in the speaker's brain before word articulation re-emerges, word by word, in the listener's brain after each word is spoken. Our model-based coupling framework ensures that both the speaker's and listener's neural activity are aligned to the same set of context-dependent linguistic embeddings.

RESULTS

We recorded cortical activity using ECoG in five dyadic pairs of epilepsy patients during free-form, face-to-face conversations. We developed a *model-based encoding framework* to model the context-dependent linguistic content shared between brains during natural conversations.[19,23] We first filtered the neural data to high-gamma broadband power to approximate local field potentials. For each dyad, we segmented each conversation into ten non-overlapping, consecutive segments (folds). Next, we spliced the neural data into word-level epochs and collated these epochs according to speaker and listener roles. We used time-resolved text transcriptions of each conversation to extract embeddings for each word from the autoregressive LLM GPT-2.[22] LLMs encode a multitude of morphological, syntactic, semantic, and pragmatic dimensions into a unified embedding space, which we collectively refer to as linguistic content throughout this paper.

We used ridge regression to estimate separate encoding models for the speaker and listener to predict the high-gamma band neural activity for each word using the embeddings from GPT-2 (Figure 1A). We trained each encoding model on nine folds and then evaluated its performance by correlating model-predicted neural activity with the actual neural activity for each word in the remaining held-out test fold. This process was repeated for each of the ten folds, and the correlations for each fold were averaged to obtain the final encoding performance. We fit separate, independent encoding models for each electrode and each of the 129 250-ms bins spanning −4 to +4 s before and after word onset (marked 0 in the plots) during speech production and comprehension. Ridge regression was used with hyperparameter search in the training set to regularize the norm of the regression weights and minimize the risks of overfitting.[24–26] To focus on electrodes that encode linguistic content, we performed a permutation test by re-estimating within-subject production and comprehension encoding models on phase-randomized neural data ($p < 0.01$, false discovery rate [FDR] corrected); these models were estimated from non-contextual, static GPT-2 embeddings to minimize selection bias for the contextual embeddings.

Contextual embeddings predict brain activity in both the speaker and listener

We assessed whether the linguistic embedding space could capture time-resolved, word-related neural activity in both the speaker and the listener. We trained electrode-wise encoding models with 10-fold consecutive cross-validation at lags ranging from −4 to +4 s relative to word onset and measured the correlation between actual and model-predicted word-related activity in each test fold separately for speech production and speech comprehension (Figure 2). During speech production, we found maximal encoding performance in speech articulation areas along the precentral motor cortex, in the superior temporal cortex, and in higher-order language areas in the temporal pole, inferior frontal gyrus (IFG), and supramarginal gyrus (Figure 2A, blue). Inspecting the dynamics of encoding performance within the speaker's brain across lags revealed that the maximal

(C) In our model-based coupling analysis, we evaluated intersubject encoding performance at each pair of lags between the speaker's and listener's brain activity, resulting in a lag-by-lag matrix of correlation values (left). Word onset is depicted with gray lines at the center, dividing the matrix into quadrants: e.g., the bottom right quadrant corresponds to the speaker's brain activity before word onset and the listener's brain activity after word onset. The horizontal dashed line connects the speaker's lag, labeled "2," before word onset to all lags in the listener—depicted graphically on the right using dashed lines. The solid diagonal of the matrix corresponds to the simultaneous speaker and listener lags, schematically shown at the bottom right as vertical solid lines between speaker-listener lags 1–4. Lag pairs under the diagonal indicate that the speaker precedes the listener (dash-dot lines), connecting speaker lags 1–3 to listener lags 2–4.

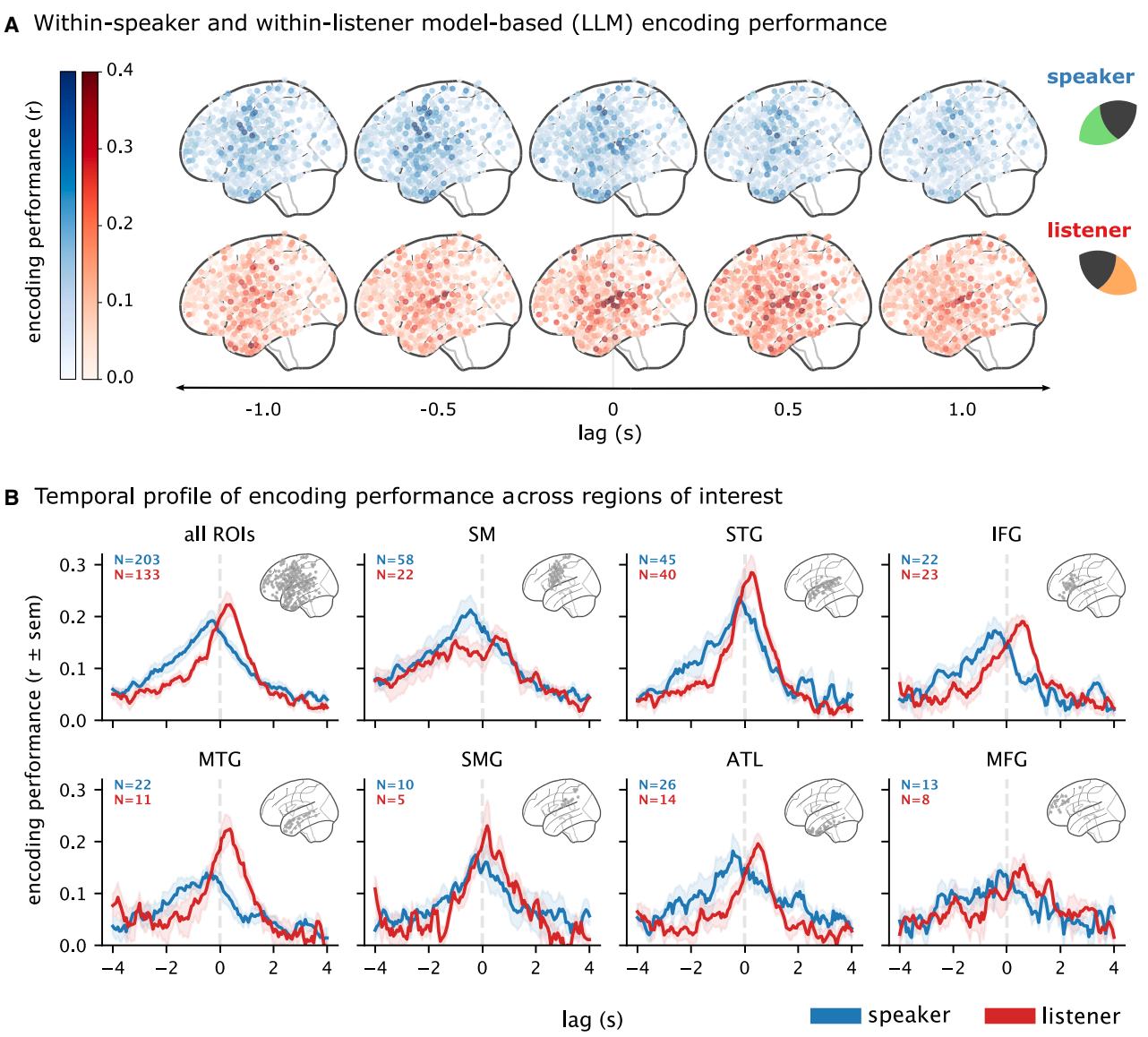

Figure 2. Within-speaker and within-listener linguistic encoding performance

Encoding models were trained to predict the neural activity from linguistic embeddings separately per lag and per electrode and evaluated using 10-fold cross-validation. Encoding performance is quantified as the correlation between word-by-word model-predicted and actual electrode activity.

(A) Encoding performance for all electrodes from all subjects at five different lags relative to word onset (lag 0). Separate models are trained for spoken words (production, blue) and heard words (comprehension, red).

(B) Encoding performance for all electrodes selected for significance ($p < 0.01$, permutation test, FDR corrected; see electrode selection in STAR Methods) across lags for all regions (B, top left) and in different regions of interest (ROIs) in the cortical language network. Error bands indicate the standard error of the mean correlation across electrodes and subjects. The number of significant electrodes in the speaker and listener is displayed in the upper left corner of each panel. SM, somatomotor cortex; STG, superior temporal gyrus; IFG, inferior frontal gyrus; MTG, middle temporal gyrus; SMG, supramarginal gyrus; ATL, anterior temporal lobe; and MFG, middle frontal gyrus.

prediction peaked ~250 ms before word onset (Figure 2B, blue; see Table S1 for peaks in individual regions of interest [ROIs]). During speech comprehension, the encoding model predicted neural responses in similar brain areas, particularly the superior temporal cortex (Figure 2A, red). Comprehension encoding performance increased gradually in superior and anterior temporal electrodes, peaked ~250 ms after word onset, and decreased over 1 s after the peak (Figure 2B, red; Table S1).

The linguistic embedding space predicts neural activity in multiple regions with different temporal dynamics and selectivity (Figure 2B). Somatomotor (SM) electrodes encode linguistic content more during speech production than comprehension, particularly before word articulation. Electrodes in the superior temporal gyrus (STG), on the other hand, encode linguistic content during both processes, although encoding performance is stronger for comprehension. In the IFG and anterior temporal

lobe (ATL), linguistic encoding during speech production peaks prior to word onset, with sustained encoding after word articulation during comprehension. Even during comprehension, encoding performance begins to ramp up before word onset. There are two possible reasons for this: (1) the embedding for a given word encodes contextual information spanning prior words, and (2) the brain may facilitate real-time processing by actively predicting the content of forthcoming words (see Goldstein et al.[19]). During speech production, encoding performance in most regions decreases rapidly after word onset; this decrease accompanies the increase in post-word-onset encoding performance in the listener (Table S1). These results demonstrate that the linguistic embedding space learned by GPT-2 captures relevant features for predicting neural activity during language production and comprehension across the cortical language network.

Contextual embeddings capture linguistic coupling between the speaker and the listener

How are the speaker's and listener's brains aligned during the conversation? The previous analysis used encoding models to predict the neural signal from word embeddings separately in the speaker and listener. To assess model-based linguistic coupling across brains, we used the encoding model already trained on the speaker's brain activity to predict the listener's brain activity (and vice versa) using the same cross-validation scheme; that is, we correlated the model-based predictions from one brain with the actual neural activity in the other brain[27] (Figure 1A). This novel model-based coupling analysis quantifies how well the model fit for speech production (or comprehension) generalizes to speech comprehension (or production) in left-out segments of each conversation. By virtue of using word embeddings from a language model, the encoding model filters out non-linguistic features that may be common between the conversants but that are not present in the conversation transcript. This model-based framework is a qualitative advance over content-agnostic, unmediated coupling methods (e.g., ISC) in that it constrains any observed speaker-listener coupling to the same set of context-dependent linguistic embeddings. To capture the temporal dynamics of speaker-listener coupling, we applied this procedure for each pair of lags in the speaker's and listener's brain activity, resulting in a lag-by-lag encoding matrix of correlation values where the y axis indexes lag in the speaker's brain and the x axis indexes lags in the listener's brain, relative to word onset (Figure 1C). In this matrix, the central axis lines signify the onset of word articulation (lag 0 s). By contrast, the matrix diagonal corresponds to simultaneous lags between the speaker and listener. Intersubject encoding below the diagonal indicates that linguistic content encoded in the speaker's brain precedes the same linguistic content encoded in the listener's brain.

We first applied this procedure across all electrodes selected for significance to estimate overall linguistic coupling across the cortical language network (n = 203 production; 133 comprehension): we averaged the *model-predicted* neural activity across electrodes (e.g., in the speaker) and correlated this with the averaged *actual* neural activity across electrodes (e.g., in the listener); different averaging schemes yielded qualitatively similar results (Figure S1). We observed time-locked, speaker-listener linguistic coupling centered around the moment of articulation of each word in the conversation (Figure 3A; training and testing on speaker or listener yielded very similar results; Figure S2). Consistent with the flow of information during communication, the linguistic coupling falls under the diagonal, indicating that the speaker's brain is "leading" the listener's brain (see Figure S3 for significance test). We found that linguistic content emerges in the speaker's brain prior to the spontaneous articulation of each word and then re-emerges in the listener's brain after each word is heard. This temporal dynamic proceeds word by word and is specific to the current word.

Model comparison of syntactic, phonemic, and LLM features

Contextual embeddings extracted from GPT-2 encode linguistic content across many dimensions of real-world text spanning morphological, syntactic, semantic, and contextual dependencies.[13–15] Unlike unmediated coupling methods (e.g., ISC analysis), our model-based framework supports rigorous model comparison. We compared encoding performance between the contextual LLM embeddings and two additional feature sets inspired by classical psycholinguistics. First, we constructed an articulatory phonemic model capturing articulatory speech features for each phoneme[28] (Table S2). Second, we constructed a syntactic model: we used spaCy to extract part-of-speech tags for each word, as well as the syntactic relations between words (based on a dependency parse tree) in each utterance (Tables S3 and S4). We evaluate these alternative feature spaces using the same encoding analysis we used for contextual embeddings. We found that contextual embeddings from GPT-2 predicted neural activity better within subjects (Figure S4) and between speaker-listener pairs than both the syntactic (Figure 3E) and articulatory (Figure 3F) features (see Figure S5 for significance tests). This was even the case in the ostensibly low-level perceptual (STG) cortex and articulatory (SM) cortex (Figures S6 and S7). This may seem surprising, but SM areas may encode ethological behavioral programs[29] that extend across longer timescales and are better captured by contextual embeddings than articulatory features.

We hypothesized that the linguistic structures captured by the classical articulatory and syntactic features may be implicitly embedded in the multidimensional contextual embeddings. To rigorously test this hypothesis, we used banded ridge regression to estimate a joint encoding model containing all three feature sets (symbolic articulatory features, symbolic syntactic features, and contextual embeddings).[26] This allows all three submodels to fairly vie for variance in predicting held-out brain activity. We found that the prediction performance of the articulatory and syntactic models is reduced nearly to zero when forced to compete for variance against the contextual embeddings (Figure S8), indicating that the symbolic articulatory and syntactic models capture very little unique variance in brain activity above and beyond what is already captured by the contextual embeddings. A variance partitioning analysis corroborated this result, revealing that the unique contribution of the contextual embeddings markedly exceeds the variance explained by articulatory and syntactic features (Figure S9).

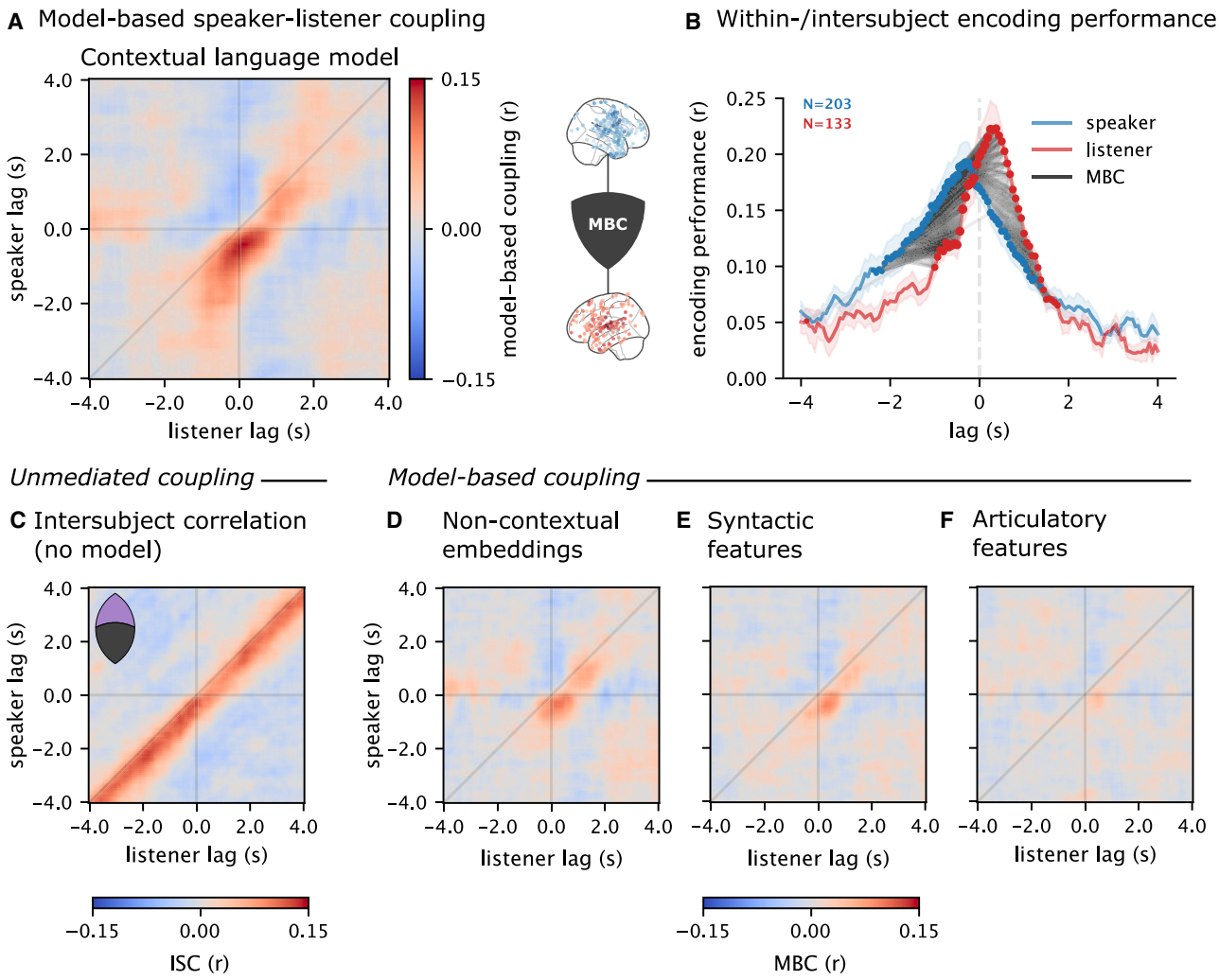

Figure 3. Speaker-listener brain-to-brain linguistic coupling

(A) Model-based speaker-listener coupling across all electrodes and regions for each pair of lags (see Figure S3 for statistical thresholding). This plot corresponds to the shared variance between the speaker's and listener's brain activity captured by the linguistic encoding model (black intersection in Figure 1B). Speaker-listener encoding peaks in the bottom right quadrant, indicating that linguistic content emerging in the speaker's brain prior to word articulation re-emerges in the listener's brain after word articulation.

(B) Within-subject encoding performance for the speaker (blue) and listener (red). Gray lines connect significant pairs of lags in the speaker-listener encoding performance matrix (A). Line transparency indicates correlation strength, while the size of each circle denotes the overall magnitude of model-based speaker-listener coupling for that lag (L2-norm of row or column). Error bands indicate the standard error of the mean correlation across electrodes and subjects.

(C) Unmediated speaker-listener intersubject correlation (ISC) without an explicit language model. This analysis directly measures the correlation between word-level neural activity but cannot isolate the word-specific linguistic content driving brain-to-brain coupling.

(D) Intersubject encoding performance using non-contextual, lexical-semantic embeddings from the trained LLM.

(E and F) (E) Intersubject encoding performance using symbolic word embeddings defined with binary vectors coding for parts of speech, syntactic dependencies, and (F) articulatory phonemic features.

Speaker-listener linguistic coupling is word, context, and conversation specific

During face-to-face communication, the speaker-listener brain responses can be coupled due to other variables, such as facial expressions, gestures, and background sounds that are not strictly linguistic in nature. We next evaluated unmediated coupling by computing the direct ISC between the speaker's and listener's brain activity (i.e., not mediated through a language model) in a way that matched the folding scheme used for cross-validation in the encoding analysis. Replicating prior results,[6–12] we found strong coupling between speaker and listener neural activity during natural conversations. Direct coupling was consistently below the diagonal, indicating that the speaker's brain activity preceded the listener's brain activity (Figure 3C). Unmediated coupling methods like ISC, however, cannot isolate the word-by-word linguistic content of the conversation. Therefore, the observed coupling is not time-locked to the articulation of each word. Words occurring before and

after the current word, regardless of their content, will contribute to the observed correlation, yielding high correlations all along the diagonal. On the other hand, the model-based speaker-listener coupling results (Figure 3A) are temporally specific, suggesting that the embeddings capture word-specific linguistic coupling in a way that cannot be observed using ISC analysis.

We next asked whether model-based speaker-listener linguistic coupling was sensitive to the specific meaning of words in context. We extracted non-contextual, lexical-semantic embeddings from GPT-2 with the same dimensionality as the contextual embeddings. In this setting, each occurrence of a given word receives the same embedding, capturing the "average" meaning of that word across all contexts. Similar to other types of word embeddings—such as word2vec and GloVe[30,31]—these representations cannot capture the unique meaning of words in context. For example, the word *cold* will receive the same embedding regardless of whether the context refers to a personality trait or the temperature. We found that speaker-listener linguistic coupling was significantly stronger for contextual embeddings over non-contextual word embeddings (Figure 3D and S5). Furthermore, to demonstrate that our results are not limited to our choice of an autoregressive language model (GPT-2), we replicated our core results using BERT,[32] a well-studied masked language model (Figure S10).

Finally, we asked whether a given speaker and listener tend to explore a conversation-specific region of the linguistic space. To examine the uniqueness of linguistic coupling in each conversation, we compared the speaker and listener weight matrices estimated by the encoding models within and across conversations. We found that the relationship between speaker and listener model weights is partly specific to each dyadic conversation with reduced generalization across conversations; that is, each conversation was biased toward a particular subset of features in the contextual embedding space (Figures S11 and S12).

Linguistic coupling across language areas within speakers and listeners

We next used LLM contextual embeddings to assess linguistic coupling across regions of the cortical language network within speakers and listeners (i.e., model-based "connectivity"). We correlated model-based predictions from one region with the actual neural activity in other regions.[27] For example, we used encoding models trained on neural activity in the speaker's ATL to predict neural activity in the speaker's STG. Similarly, we used encoding models trained on neural activity in the listener's STG to predict neural activity in the listener's ATL. This analysis yielded lag-by-lag encoding matrices across pairs of language regions within the speaker and the listener (Figures S13 and S14).

During language production, we found a dense network of inter-regional linguistic encodings with many regions encoding similar features of the linguistic embedding space (Figure 4A, top). These inter-regional connections include both higher-level (ATL and SMG) and lower-level (SM, STG, and IFG) language areas. For example, the speaker's STG and IFG are both coupled to the speaker's SM but with different temporal structures (Figure 4A, bottom; Figure S13); SM precedes STG, and coupling is closely tied to word articulation, while linguistic coupling between SM and IFG is largely synchronous and more temporally diffuse. Interestingly, SM tends to precede most regions, except for IFG. By contrast, during language comprehension, linguistic coupling reveals a sparser network comprising typical language areas (Figure 4C): STG is coupled with and generally precedes IFG, middle temporal gyrus [MTG], and ATL (Figure 4C, bottom). Unexpectedly, although SM did have strong encoding performance during comprehension (Figure 2B), it was not strongly coupled to any other region, suggesting that it encodes a particular set of linguistic features not shared with other areas (Figure S14).

Speaker-listener linguistic coupling across language areas

Model-based speaker-listener coupling was widespread across the cortical language network and partially asymmetric between different language areas (Figure 4B). In the speaker, SM, STG, and ATL were the primary drivers of the speaker-listener linguistic coupling; in the listener, STG, IFG, and ATL were the primary receivers (Figure S15). The speaker's SM was strongly coupled with the listener's STG, IFG, and ATL. Interestingly, the speaker's SM activity before word onset was coupled to the listener's STG before word onset (lower left quadrant), then differently coupled after word onset (upper right quadrant, with minimal coupling in the lower right quadrant); this suggests that both regions encode shared features that change rapidly at word articulation. By contrast, the speaker's SM before word onset was most strongly coupled to the listener's IFG after word onset (lower right quadrant). Although most speaker-listener couplings were asymmetric between language areas, the speaker's STG was coupled to the listener's STG along the diagonal around word articulation, suggesting a short delay in these lower-level regions (Figure 1C). The speaker's ATL prior to word onset was coupled with the listener's ATL around word onset with a larger speaker-listener latency (farther off the diagonal). These results suggest a network configuration where linguistic coupling between the high-level language and articulatory areas drives speech production, whereas coupling primarily between STG and IFG, followed by higher-level regions, underlies speech comprehension.

DISCUSSION

In this study, we show that contextual embeddings from an LLM encode the shared linguistic space between speakers. We develop a model-based coupling framework that quantifies how linguistic information is communicated from brain to brain in spontaneous, face-to-face conversations via a shared set of context-sensitive linguistic embeddings extracted from an LLM. Specifically, we extracted a contextual embedding for each word in each conversation and used it to recover word-specific brain activity shared between the speaker and listener. We find that context-dependent linguistic content emerges in the speaker's brain activity before word articulation, and the same linguistic content later re-emerges in the listener's brain after word articulation. This temporal dynamic matches the overall flow of information in a conversation, where speakers dynamically alternate roles in transmitting their ideas to one another.

The contextual embeddings learned by an LLM approximate a multidimensional linguistic "code"—shared across brains—that

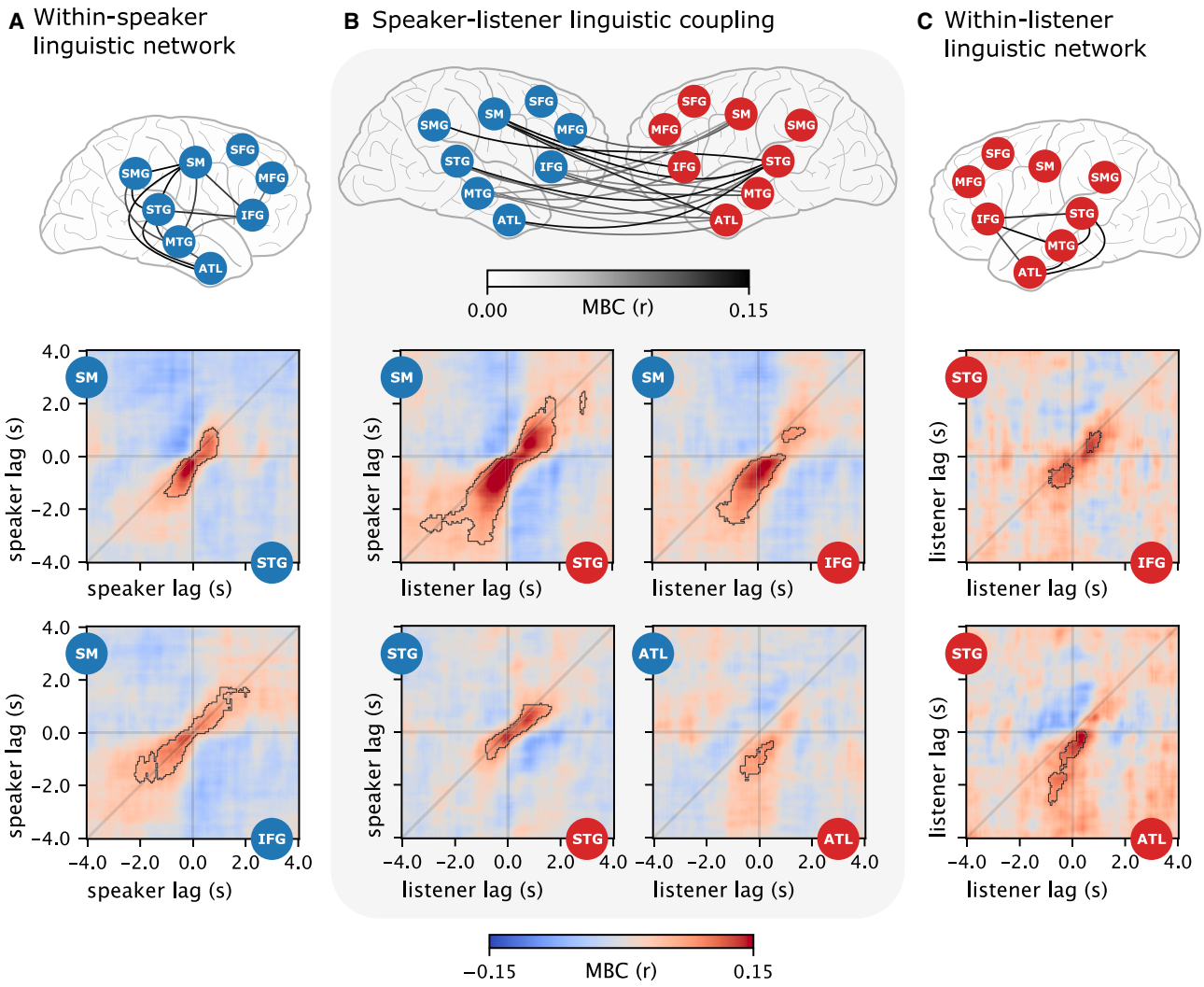

Figure 4. Inter-regional model-based coupling within and across subjects

(A) Linguistic encoding across regions of the cortical language network within the speaker's brain. The connection between the anterior temporal lobe (ATL) and somatosensory (SM), for example, indicates the maximum encoding performance across lags for a model trained on ATL and tested on SM (and vice versa). The network diagram (top) depicts significant linguistic coupling between two regions as bidirectional gray lines ($p < 0.05$, Bonferroni corrected), where the darkness of the connection indexes the strength of linguistic encoding. Below are lag-by-lag within-speaker encoding matrices for two example pairs of regions with contour lines denoting significant lags (see Figure S13 for all pairs of ROIs and the number of electrodes per area).

(B) Inter-regional model-based coupling between the speaker (blue) and listener (red) across multiple language areas. Network diagrams (top) show significant speaker-listener coupling for pairs of regions; the darkness of the connection indexes the strength of linguistic encoding. Below are lag-by-lag model-based coupling matrices for four example pairs of regions.

(C) Linguistic encoding across regions of the cortical language network within the listener's brain (top), with lag-by-lag encoding matrices for two example pairs of regions below.

both speaker and listener can use to communicate their thoughts during natural conversations. Our findings align with prior work demonstrating that LLM contextual embeddings capture linguistic features of brain activity during language comprehension (Figure 2, red).[17–21,33,34] Our findings expand over prior work in two significant ways. Unique to this study is the dyadic conversational structure, with neural activity simultaneously recorded in two patients while they freely talk to each other. Our findings expand on prior work by showing that the same contextual embeddings are shared across the brains of both speaker and listener. Our findings demonstrate that the same embeddings also capture the emergence of word-by-word linguistic features before word articulation during speech production (Figure 2, blue). LLMs can both comprehend and generate conversational language, showcasing how understanding and producing language are closely intertwined. We found that production and comprehension are closely aligned to the same abstract, context-specific embedding space, providing a common ground for speakers to transmit their thoughts to one another.[35]

Our results demonstrate that the contextual embeddings learned by an LLM better predict the *shared* activity between speaker and listener than non-contextual word embeddings (Figure 3D), as well as symbolic articulatory (Figure 3F) and syntactic (Figure 3E) features (Figures S4 and S5). LLMs are a powerful new family of models for investigating the underlying neural code supporting natural language processing in the human brain.[19,36–39] LLMs learn the context-rich meaning of language through exposure to other people's linguistic acts, not entirely dissimilar from how children learn meaningful language from their community of speakers. Furthermore, LLMs rely on simple objectives like self-supervised next-word prediction (and also from human feedback in recent models[39]) to effectively learn language. These objectives are cognitively plausible and accessible to human language learners. For example, a recent study indicates that the brain, like LLMs, actively predicts the meaning of upcoming words when processing natural spoken language.[19] Finally, unlike the symbolic models of classical psycholinguistics, LLMs are expressive enough to leverage the complex statistical dependencies of real-world language. Combined, these parallels may explain why LLM contextual embeddings prove to be powerful models of the shared, context-dependent linguistic code necessary for transmitting our thoughts in open-ended conversations (Figure 3A).

Taking advantage of the high temporal resolution of ECoG, we were able to map out how linguistic content is shared across language areas both within and across brains. We first measured model-based connectivity among language areas within the speaker's and listener's brains separately. During spontaneous, natural speech production, we observed a dense network of model-based connectivity, with many regions encoding similar features of the linguistic embedding space (Figure 4A). These inter-regional connections include higher-level (IFG, ATL, and SMG) and lower-level (STG and SM) language areas. Surprisingly, the SM cortex appears to "lead" most other regions in the speaker's brain.[40] SM and IFG were only moderately coupled,[41] while ATL plays an unexpectedly large role in speech production. SM precedes STG during speech production, potentially reflecting efferent copy or feedback control.[42–44] In some cases, network dynamics were reversed in speaking and listening. For example, in the speaker's brain, we see that IFG precedes STG; in the listener's brain, on the other hand, we observe the opposite relationship, where STG precedes IFG (Figure S16). During language comprehension (Figure 4C), model-based connectivity revealed a more sparse network comprising typical language areas proceeding from putatively lower- to higher-level areas[45] (STG, MTG, IFG, and ATL). Finally, we mapped out how linguistic content is transmitted from the speaker's brain to the listener's brain, revealing asymmetric inter-regional coupling (Figure 4B). The speaker's SM and the listener's STG and IFG are strongly coupled before and after word articulation. The speaker's ATL before word onset is aligned to the listener's ATL as each word is articulated, perhaps reflecting the ongoing context that primes each upcoming word. Interestingly, the listener's SM cortex was not tightly coupled to other regions in the listener's brain[46] but was nonetheless coupled to several regions in the speaker's brain (Figures 4, S13, and S14). Future work may combine, for example, content-agnostic partial correlation[47] or causal[48] methods with this model-based coupling method to more precisely track the flow of information within cortical circuits and from speaker to listener.

Prior work on brain-to-brain coupling has relied on content-agnostic, unmediated methods with no explicit intermediary model of shared features, such as ISC analysis.[4–12,49–51] These unmediated coupling methods can only quantify the *magnitude* of speaker-listener neural coupling but cannot capture the word- and context-specific linguistic information shared between speaker and listener. In the current manuscript, we developed a model-based coupling framework that allows us to assess the transfer of shared linguistic information across brains. Model-based coupling effectively filters out non-linguistic features and ensures that shared neural activity between speaker and listener is aligned to a common set of explicit linguistic features. Using this modeling framework, we tested several competing language models and found that the contextual embeddings learned by an LLM most robustly capture the context-specific, word-by-word linguistic information transmitted from brain to brain in conversation. This computational framework, combined with models that can reproduce real-world language, marks a paradigm shift from unmediated brain-to-brain coupling[4,52] toward a more precise, model-driven neuroscience of social interaction. Taken together, our research indicates that contextual embeddings offer an explicit, numerical model of the same linguistic code that humans use to share their thoughts with one another.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Electrocorticography acquisition
  - Electrode localization
- METHOD DETAILS
  - Signal preprocessing
  - Transcription and alignment
  - Contextual embedding extraction
  - Encoding analysis
  - Electrode selection
  - Model-based brain-to-brain coupling
  - Model comparison framework
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2024.06.025.

ACKNOWLEDGMENTS

We thank Bobbi Aubrey, Mariya Toneva, Robert Hawkins, and Diana Tamir for helpful feedback on the data and analyses. We thank Christian Keysers for helpful comments during the review process. Funding was provided by

National Institutes of Health grant DP1HD091948 (Z.Z., A.G., and U.H.) and National Institutes of Health grant R01MH112566 (S.A.N.).

AUTHOR CONTRIBUTIONS

Conceptualization, Z.Z., S.A.N., A.G., and U.H.; data curation, Z.Z., E.B., W.D., D.F., P.D., L.M., S.D., A.F., and O.D.; formal analysis, Z.Z.; funding acquisition, U.H.; investigation, E.S., A.P., A.Z., and L.H.; methodology, Z.Z., S.A.N., A.G., S.M., and U.H.; projectadministration, Z.Z., S.A.N., and U.H.; software, Z.Z.; supervision, S.A.N. and U.H.; visualization, Z.Z., S.A.N., and U.H.; writing – original draft, Z.Z., S.A.N., and U.H.; writing – review & editing, Z.Z., A.F., S.A.N., and U.H.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 3, 2023
Revised: March 26, 2024
Accepted: June 25, 2024
Published: August 2, 2024

SUPPORTING CITATION

The following references appear in the supplemental information: [53].

REFERENCES

1. Wittgenstein, L. (1953). Philosophical Investigations (Wiley-Blackwell).
2. Dor, D. (2015). The Instruction of Imagination: Language as a Social Communication Technology (Foundations of Human Interaction).
3. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. (2004). Intersubject Synchronization of Cortical Activity During Natural Vision. Science *303*, 1634–1640. https://doi.org/10.1126/science.1089506.
4. 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.
5. Nastase, S.A., Gazzola, V., Hasson, U., and Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. Soc. Cogn. Affect. Neurosci. *14*, 667–685. https://doi.org/10.1093/scan/nsz037.
6. 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.
7. Dikker, S., Silbert, L.J., Hasson, U., and Zevin, J.D. (2014). On the Same Wavelength: Predictable Language Enhances Speaker–Listener Brain-to-Brain Synchrony in Posterior Superior Temporal Gyrus. J. Neurosci. *34*, 6267–6272. https://doi.org/10.1523/JNEUROSCI.3796-13.2014.
8. 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.
9. Liu, Y., Piazza, E.A., Simony, E., Shewokis, P.A., Onaral, B., Hasson, U., and Ayaz, H. (2017). Measuring speaker–listener neural coupling with functional near infrared spectroscopy. Sci. Rep. *7*, 43293. https://doi.org/10.1038/srep43293.
10. Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y.-F., Nguyen, M., Norman, K.A., and Hasson, U. (2021). Neural alignment predicts learning outcomes in students taking an introduction to computer science course. Nat. Commun. *12*, 1922. https://doi.org/10.1038/s41467-021-22202-3.
11. 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.
12. Davidesco, I., Laurent, E., Valk, H., West, T., Milne, C., Poeppel, D., and Dikker, S. (2023). The Temporal Dynamics of Brain-to-Brain Synchrony Between Students and Teachers Predict Learning Outcomes. Psychol. Sci. *34*, 633–643. https://doi.org/10.1177/09567976231163872.
13. Manning, C.D., Clark, K., Hewitt, J., Khandelwal, U., and Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. Proc. Natl. Acad. Sci. USA *117*, 30046–30054. https://doi.org/10.1073/pnas.1907367117.
14. Linzen, T., and Baroni, M. (2021). Syntactic Structure from Deep Learning. Annu. Rev. Linguist. *7*, 195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035.
15. Pavlick, E. (2022). Semantic Structure in Deep Learning. Annu. Rev. Linguist. *8*, 447–471. https://doi.org/10.1146/annurev-linguistics-031120-122924.
16. Piantadosi, S.T. (2023). Modern language models refute Chomsky's approach to language. https://doi.org/lingbuzz/007180.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
22. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). GPT-2 Language Models are Unsupervised Multitask Learners. OpenAI.
23. Holdgraf, C.R., Rieger, J.W., Micheli, C., Martin, S., Knight, R.T., and Theunissen, F.E. (2017). Encoding and Decoding Models in Cognitive Electrophysiology. Front. Syst. Neurosci. *11*, 61. https://doi.org/10.3389/fnsys.2017.00061.
24. 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.
25. 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.
26. 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.
27. Toneva, M., Williams, J., Bollu, A., Dann, C., and Wehbe, L. (2022). Same Cause; Different Effects in the Brain. In First Conference on Causal Learning and Reasoning (MLR Press), pp. 1–35.
28. 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.
29. Graziano, M.S.A. (2016). Ethological Action Maps: A Paradigm Shift for the Motor Cortex. Trends Cogn. Sci. *20*, 121–132. https://doi.org/10.1016/j.tics.2015.10.008.

30. Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Preprint at ArXiv.

31. Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics), pp. 1532–1543. https://doi.org/10.3115/v1/D14-1162.

32. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Preprint at. In Proceedings of the 2019 Conference of the North American Chapter of the Assocation for Computational Linguistics: Human Language Technologies, pp. 4171–4186. https://doi.org/10.18653/V1/N19-1423.

33. Toneva, M., and Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In Advances in Neural Information Processing Systems, pp. 14954–14964.

34. Cai, J., Hadjinicolaou, A.E., Paulk, A.C., Williams, Z.M., and Cash, S.S. (2023). Natural language processing models reveal neural dynamics of human conversation. Preprint at bioRxiv. https://doi.org/10.1101/2023.03.10.531095.

35. Pickering, M.J., and Garrod, S. (2013). An integrated theory of language production and comprehension. Behav. Brain Sci. *36*, 329–347. https://doi.org/10.1017/S0140525X12001495.

36. Goldstein, A., Grinstein-Dabush, A., Schain, M., Wang, H., Hong, Z., Aubrey, B., Nastase, S.A., Zada, Z., Ham, E., et al. (2024). Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. Nat. Commun. *15*, 2768. https://doi.org/10.1038/s41467-024-46631-y.

37. Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. Nat. Neurosci. *22*, 1761–1770. https://doi.org/10.1038/s41593-019-0520-2.

38. Hasson, U., Nastase, S.A., and Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. Neuron *105*, 416–434. https://doi.org/10.1016/j.neuron.2019.12.002.

39. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. Adv. Neural Inf. Process. Syst. *35*, 27730–27744.

40. Glanz Iljina, O., Derix, J., Kaur, R., Schulze-Bonhage, A., Auer, P., Aertsen, A., and Ball, T. (2018). Real-life speech production and perception have a shared premotor-cortical substrate. Sci. Rep. *8*, 8898. https://doi.org/10.1038/s41598-018-26801-x.

41. Flinker, A., Korzeniewska, A., Shestyuk, A.Y., Franaszczuk, P.J., Dronkers, N.F., Knight, R.T., and Crone, N.E. (2015). Redefining the role of Broca's area in speech. Proc. Natl. Acad. Sci. USA *112*, 2871–2875. https://doi.org/10.1073/pnas.1414491112.

42. Hickok, G., Houde, J., and Rong, F. (2011). Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization. Neuron *69*, 407–422. https://doi.org/10.1016/j.neuron.2011.01.019.

43. Khalilian-Gourtani, A., Wang, R., Chen, X., Yu, L., Dugan, P., Friedman, D., Doyle, W., Devinsky, O., Wang, Y., and Flinker, A. (2022). A Corollary Discharge Circuit in Human Speech. Preprint at bioRxiv. https://doi.org/10.1101/2022.09.12.507590.

44. Ozker, M., Doyle, W., Devinsky, O., and Flinker, A. (2022). A cortical network processes auditory error signals during human speech production to maintain fluency. PLOS Biol. *20*, e3001493. https://doi.org/10.1371/journal.pbio.3001493.

45. Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nat. Rev. Neurosci. *8*, 393–402. https://doi.org/10.1038/nrn2113.

46. Cheung, C., Hamiton, L.S., Johnson, K., and Chang, E.F. (2016). The auditory representation of speech sounds in human motor cortex. eLife *5*, e12577. https://doi.org/10.7554/eLife.12577.

47. Marrelec, G., Krainik, A., Duffau, H., Pélégrini-Issac, M., Lehéricy, S., Doyon, J., and Benali, H. (2006). Partial correlation for functional brain interactivity investigation in functional MRI. NeuroImage *32*, 228–237. https://doi.org/10.1016/j.neuroimage.2005.12.057.

48. Seth, A.K., Barrett, A.B., and Barnett, L. (2015). Granger Causality Analysis in Neuroscience and Neuroimaging. J. Neurosci. *35*, 3293–3297. https://doi.org/10.1523/JNEUROSCI.4399-14.2015.

49. Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J., Michalareas, G., Van Bavel, J.J., Ding, M., and Poeppel, D. (2017). Brain-to-Brain Synchrony Tracks Real-World Dynamic Group Interactions in the Classroom. Curr. Biol. *27*, 1375–1380. https://doi.org/10.1016/j.cub.2017.04.002.

50. 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.

51. Piazza, E.A., Hasenfratz, L., Hasson, U., and Lew-Williams, C. (2020). Infant and Adult Brains Are Coupled to the Dynamics of Natural Communication. Psychol. Sci. *31*, 6–17. https://doi.org/10.1177/0956797619878698.

52. 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.

53. Yamashita, M., Kubo, R., and Nishimoto, S. (2023). Cortical representations of languages during natural dialogue. Preprint at bioRxiv. https://doi.org/10.1101/2023.08.21.553821.

54. Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. Front. Neurosci. *7*, 267. https://doi.org/10.3389/fnins.2013.00267.

55. 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.

56. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In Interspeech, pp. 498–502. https://doi.org/10.21437/Interspeech.2017-1386.

57. 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. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (Association for Computational Linguistics), pp. 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

58. Yang, A.I., Wang, X., Doyle, W.K., Halgren, E., Carlson, C., Belcher, T.L., Cash, S.S., Devinsky, O., and Thesen, T. (2012). Localization of dense intracranial electrode arrays using magnetic resonance imaging. Neuroimage *63*, 157–165. https://doi.org/10.1016/j.neuroimage.2012.06.039.

59. Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage *31*, 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021.

60. Holdgraf, C., Appelhoff, S., Bickel, S., Bouchard, K., D'Ambrosio, S., David, O., Devinsky, O., Dichter, B., Flinker, A., Foster, B.L., et al. (2019). iEEG-BIDS, extending the Brain Imaging Data Structure specification to human intracranial electrophysiology. Sci. Data *6*, 102. https://doi.org/10.1038/s41597-019-0105-7.

61. Michelmann, S., Treder, M.S., Griffiths, B., Kerrén, C., Roux, F., Wimber, M., Rollings, D., Sawlani, V., Chelvarajah, R., Gollwitzer, S., et al. (2018). Data-driven re-referencing of intracranial EEG based on independent component analysis (ICA). J. Neurosci. Methods *307*, 125–137. https://doi.org/10.1016/j.jneumeth.2018.06.021.

62. Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., and Malach, R. (2005). Coupling Between Neuronal Firing, Field Potentials, and fMRI in Human Auditory Cortex. Science *309*, 951–954. https://doi.org/10.1126/science.1110913.

63. Steinschneider, M., Fishman, Y.I., and Arezzo, J.C. (2008). Spectrotemporal Analysis of Evoked and Induced Electroencephalographic Responses in Primary Auditory Cortex (A1) of the Awake Monkey. Cereb. Cortex *18*, 610–625. https://doi.org/10.1093/cercor/bhm094.

64. Manning, J.R., Jacobs, J., Fried, I., and Kahana, M.J. (2009). Broadband Shifts in Local Field Potential Power Spectra Are Correlated with Single-Neuron Spiking in Humans. J. Neurosci. *29*, 13613–13620. https://doi.org/10.1523/JNEUROSCI.2041-09.2009.

65. Edwards, E., Nagarajan, S.S., Dalal, S.S., Canolty, R.T., Kirsch, H.E., Barbaro, N.M., and Knight, R.T. (2010). Spatiotemporal imaging of cortical activation during verb generation and picture naming. NeuroImage *50*, 291–301. https://doi.org/10.1016/j.neuroimage.2009.12.035.

66. Cohen, L., Loughlin, P., and Vakman, D. (1999). On an ambiguity in the definition of the amplitude and phase of a signal. Signal Process. *79*, 301–307. https://doi.org/10.1016/S0165-1684(99)00103-6.

67. Goldstein, A., Ham, E., Nastase, S.A., Zada, Z., Grinstein-Dabus, A., Aubrey, B., Schain, M., Gazula, H., Feder, A., Doyle, W., et al. (2022). Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. Preprint at bioRxiv. https://doi.org/10.1101/2022.07.11.499562.

68. Antonello, R., and Huth, A. (2024). Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data. Neurobiol. Lang. (Camb) *5*, 64–79. https://doi.org/10.1162/nol_a_00087.

69. Guest, O., and Martin, A.E. (2023). On Logical Inference over Brains, Behaviour, and Artificial Neural Networks. Comput. Brain Behav. *6*, 213–227. https://doi.org/10.1007/s42113-022-00166-x.

70. la Tour, T.D., Eickenberg, M., and Gallant, J. (2022). Feature-space selection with banded ridge regression. Preprint at bioRxiv. https://doi.org/10.1101/2022.05.05.490831.

71. Nichols, T.E., and Holmes, A.P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. Hum. Brain Mapp. *15*, 1–25. https://doi.org/10.1002/hbm.1058.

72. 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.

73. Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J.L. (2011). Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. Curr. Biol. *21*, 1641–1646. https://doi.org/10.1016/j.cub.2011.08.031.

74. 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.

75. Van Uden, C.E., Nastase, S.A., Connolly, A.C., Feilong, M., Hansen, I., Gobbini, M.I., and Haxby, J.V. (2018). Modeling Semantic Encoding in a Common Neural Representational Space. Front. Neurosci. *12*, 437. https://doi.org/10.3389/fnins.2018.00437.

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

77. Toneva, M., Mitchell, T.M., and Wehbe, L. (2022). Combining computational controls with natural text reveals aspects of meaning composition. Nat. Comput. Sci. *2*, 745–757. https://doi.org/10.1038/s43588-022-00354-6.

78. 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.

79. Stokes, M.G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., and Duncan, J. (2013). Dynamic Coding for Cognitive Control in Prefrontal Cortex. Neuron *78*, 364–375. https://doi.org/10.1016/j.neuron.2013.01.039.

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

81. Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4. https://doi.org/10.3389/neuro.06.004.2008.

82. Wu, M.C.-K., David, S.V., and Gallant, J.L. (2006). Complete functional characterization of sensory neurons by system identification. Annu. Rev. Neurosci. *29*, 477–505. https://doi.org/10.1146/annurev.neuro.29.051605.113024.

83. 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.

84. Levelt, W.J.M. (1993). Speaking: from Intention to Articulation (The MIT Press). https://doi.org/10.7551/mitpress/6393.001.0001.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
MNE Python Library	Gramfort et al.[54]	DOI: https://doi.org/10.3389/fnins.2013.00267
SciPy	Virtanen et al.[55]	DOI: https://doi.org/10.1038/s41592-019-0686-2
Montreal Forced Aligner	McAuliffe et al.[56]	DOI: https://doi.org/10.21437/Interspeech.2017-1386
HuggingFace Transformers	Wolf et al.[57]	DOI: https://doi.org/10.18653/v1/2020.emnlp-demos.6
Himalaya	la Tour et al.[26]	DOI: https://doi.org/10.1101/2022.05.05.490831
Original code	This paper	DOI: https://doi.org/10.5281/zenodo.12359298

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Zaid Zada (zzada@princeton.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

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. Due to the sensitive nature of the unconstrained speech, the data cannot be shared publicly; we will make the data available to reviewers upon request. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Electrocorticography acquisition

Twelve participants (6 dyads) engaged in free-form, first-time conversations (Table S5). One dyad was excluded from the analysis due to a short conversation length (3.56 minutes) and an insufficient number of words spoken (185 and 156 words for each participant, respectively). Participants were instructed to discuss any topic, including hobbies, vacation stories, movies, etc. Participants were recruited from the New York University School of Medicine Comprehensive Epilepsy Center. They provided oral and written informed consent before study participation, according to the New York University Langone Medical Center Institutional Review Board. All participants elected to undergo intracranial monitoring for clinical purposes and were informed that their clinical care was unrelated to participation in this study and that withdrawing from the study at any point would not affect their medical treatment.

Electrode placement was determined by clinicians for each participant based on clinical criteria. Brain activity was recorded from intracranially implanted subdural platinum–iridium electrodes embedded in silastic sheets (2.3 mm-diameter contacts; Ad-Tech Medical Instrument). Decisions related to electrode placement and invasive monitoring duration were determined solely on clinical grounds without reference to this or any other research study. Electrodes were arranged as grid arrays (8 × 8 contacts, 10 mm center-to-center spacing) or linear strips.

Recordings from grid, strip, and depth electrode arrays were acquired using the NicoletOne C64 clinical amplifier (Natus Neurologics), band-pass filtered from 0.16–250 Hz, and digitized at 512 Hz. Intracranial electroencephalography signals were referenced to a two-contact subdural strip facing toward the skull near the craniotomy site. All electrodes were visually inspected, and those with excessive noise artifacts, epileptiform activity, or no signal were removed from subsequent analyses.

Presurgical and postsurgical T1-weighted magnetic resonance imaging (MRI) scans were acquired for each participant, and the location of the electrodes relative to the cortical surface was determined from co-registered magnetic resonance imaging or computed tomography scans.[58] Co-registered, skull-stripped T1 images were nonlinearly registered to an MNI152 template, and electrode locations were extracted in Montreal Neurological Institute space (projected to the cortical surface) using the co-registered image.

Electrode localization

Electrodes were localized to anatomically defined cortical regions based on the Desikan-Killiany atlas in FreeSurfer.[59] Regions of interest (ROIs) consisted of one or more of the following atlas labels: somatomotor (SM) cortex, including precentral and postcentral gyri; superior temporal gyrus (STG), including the posterior superior temporal sulcus; middle temporal gyrus (MTG); middle frontal gyrus (MFG), comprising rostral and caudal middle frontal gyrus; supramarginal gyrus (SMG); inferior frontal gyrus (IFG), comprising pars opercularis, pars orbitalis, and pars triangularis; anterior temporal lobe (ATL), comprising anterior inferior temporal cortex and the temporal pole; superior frontal gyrus (SFG). Electrodes overlapping multiple regions were assigned based on their percent overlap with the largest area. Only electrodes in the left hemisphere were considered.

METHOD DETAILS

Signal preprocessing

ECoG data were standardized according to the Brain Imaging Data Structure (iEEG-BIDS)[60] and preprocessed using the MNE Python library,[54] with accompanying customized Python scripts. The pipeline consisted of the following steps: removing spikes, re-referencing, notch filtering, then high-gamma band extraction. First, large spikes exceeding four quartiles above and below the median of each channel were removed, and replacement samples were imputed using the SciPy *pchip_interpolate* function.[55] We then re-referenced all electrodes in each subject to account for shared signals across channels using an independent component analysis method.[61] We removed line noise at 60, 120, and 180 Hz frequencies using MNE's *notch_filter* function with a width of 2 Hz. Finally, we estimated broadband power at 70-200 Hz (high gamma) using an IIR filter followed by a Hilbert envelope computation to estimate the instantaneous magnitude of the signal (using *filter* and *apply_hilbert(envelope=True)* from MNE). High-frequency broadband power has been shown to correlate with local neural firing rates.[62–65] In this study, we estimated the amplitude of the Hilbert-transformed signal. It has been argued that this transformation inaccurately estimates phase angles.[66] Since we do not use the phase estimates to assess inter-brain synchrony in this manuscript, our results are not subject to this concern.

Transcription and alignment

We recorded each conversation's audio with a microphone at the full 44,100 Hz rate. In addition, the clinical amplifier also received the microphone output and was saved at 512 Hz. Each conversation was manually transcribed, and each utterance was manually identified to a speaker and aligned to the audio. Sounds such as laughter, breathing, or inaudible speech were marked to improve the alignment's accuracy. Punctuation and capitalization were included to the best ability of the transcriber. The Montreal Forced Aligner[56] was used to align the audio to the transcript automatically and to compute onsets and offsets for each word. The number of words per conversation is reported in Table S6. To finely align the audio with the brain activity, we cross-correlated the microphone audio (44,100 Hz) envelope with the clinical amplifier audio (512 Hz) envelope. The lag of the peak correlation was used to translate the word onsets to match the brain data. We further validated this alignment in two ways: first, by cross-correlating the ECoG signal with a downsampled version of the audio envelope. Most electrodes in STG exhibited clear peaks with small latency in the correlogram. Second, we extracted epochs around the onset for each word and averaged them into one "evoked" response. This showed a clear peak after onset in the audio envelope and in several electrodes.

Contextual embedding extraction

We used the pre-trained extra-large version of GPT-2[22] with 48 layers in the HuggingFace Transformers library.[57] We first converted all words from the raw transcript to GPT-2-specific tokens (full words and subwords) and then to integer identifiers. We supplied the model with each token up to the maximum allowed by the 1024-token context window to extract the embedding from the activations (hidden states) of the final word in the sequence. Embeddings at each layer are 1,600-element numerical vectors. Only the middle 24th layer was considered in subsequent analyses, as middle layers have been shown to be better predictors of neural activity.[20,67] Finally, sub-word token embeddings were averaged for each whole word to harmonize with the original transcript. To compare embedding spaces, this same procedure was used to extract embeddings from an untrained GPT-2 model—with random initial weights but the same architecture, context window, and inputs. We also extracted static, non-contextual GPT-2 embeddings corresponding to the token embedding weights learned by the model. As a final control, we generated random normal embeddings with the same dimensionality as GPT-2 embeddings (1,600 features) for each word instance: thus, two instances of the same word would receive two separate random embeddings (to mimic the fact that actual GPT-2 embeddings differ for different instances of the same word across contexts).

In order to demonstrate that our results generalize to other large language models that also capture a similar shared linguistic space, we replicated our core analyses using the widely studied masked language model BERT.[32] We used the large, cased, whole-word-masking version of BERT from the HuggingFace library. We divided each transcript into utterances composed of one or more sentences. One speaker produced each utterance at a time. Then, each utterance was prepended with the words that appeared before it, if any, to fill up the maximum context length of the model (512 tokens). Once fed into the model, we extracted the activations corresponding only to the utterance in consideration of that input and not the context. These activations were taken from the middle layer to match our procedure with GPT-2.

Contextual embeddings extracted from large language models (like GPT-2) concurrently encode multiple linguistic dimensions embedded in natural language. We use the term "linguistic content" to refer to all the different linguistic features encoded in the LLM embedding space—this comprises many correlated features of natural text, including syntactic, lexical-semantic, and contextual features, which may also be correlated with non-textual articulatory and phonemic features. These linguistic dimensions are entangled in natural language, and simply observing that a model predicts brain activity to some extent does not allow us to attribute this prediction performance to some particular linguistic dimension or feature.[68,69]

Encoding analysis

A linear model was estimated using ridge regression to predict the neural signal separately for each electrode and every lag relative to word onset. We used the full 1,600 contextual embedding from the language model as the predictors. Two encoding models were estimated for each subject: one for words they produced as a speaker and a separate model for words they heard as a listener. The neural signal was divided into epochs for every word (separately for spoken or heard words). Each epoch ranges from –4 s to +4 s in 129 bins of 250 ms overlapping frames with jumps of 62.5 ms (see Figure S17 for the relation of the 250 ms window size to results). We fit a different encoding model for each electrode and lag using ridge regression to resolve the temporal dynamics of linguistic encoding. Thus, the predictor matrix contains a row for each word comprising its model-based representation (e.g., GPT-2 embeddings), and the multiple ridge regression is used to predict a target matrix shaped number-of-words by number-of-lags for each electrode. Each target variable comprises the amplitude of high-gamma neural activity for that word at a specific lag relative to word onset and for a specific electrode. Each encoding model was trained separately, independently of encoding models trained for other lags and electrodes. The encoding model was evaluated by computing the Pearson correlation coefficient between actual and predicted neural signals for left-out test sets using 10-fold cross-validation. The data was split consecutively (i.e., into temporally contiguous segments) for cross-validation so as to minimize the autocorrelation between training and test folds. The final correlation values reported in the text are the average correlation across all test folds. We used the *RidgeCV* implementation from the *Himalaya* Python library[70] to fit the encoding models. Regularization coefficients (i.e., L2 penalty terms) were selected from 20 log-spaced values ranging from 1 to 1,000,000 using random search with nested cross-validation (5-fold inner cross-validation for hyperparameter selection within each training fold of the outer 10-fold cross-validation loop; as implemented in *Himalaya*). Note that for both the predictor matrix of embeddings and the target vector for each electrode, the number of samples corresponds to the number of words (not the number of time points).

Electrode selection

We first performed a permutation test to generate a null distribution based on phase-randomized neural signals to select electrodes involved in language production and comprehension for further analysis. Phase randomization effectively decouples the time series of neural activity from the onsets/offsets of each word and utterance. For each of the 1,000 permutations, we fit the same ridge-regression encoding models described above based on static, non-contextual GPT-2 embeddings across all electrodes and lags (Figure S18). The null distribution was constructed from the maximum encoding performance across lags; one p-value per electrode was calculated according to its own null distribution. Constructing the null distribution from maximum encoding performance values across lags effectively controls for multiple tests across lags.[71] We then controlled the false discovery rate (FDR)[72] at an alpha value of 1% to control for multiple tests across the p-values at each electrode for production and comprehension in each subject. The final number of selected electrodes per subject per region is detailed in Table S7. Of all selected electrodes, 231 were from grids, 87 were from strips, and 18 were depth electrodes.

To focus on electrodes that encode linguistic content, we performed a randomization test by re-estimating within-subject production and comprehension encoding models on phase-randomized neural data ($p < .01$, FDR corrected); these models were estimated from non-contextual, static GPT-2 embeddings. We opted to use non-contextual embeddings for electrode selection to minimize selection bias for the contextual embeddings. However, alternative electrode selection criteria replicate the same qualitative result (Figure S19).

Model-based brain-to-brain coupling

The contextual embeddings learned by LLMs are a precise, mathematically explicit model of the geometric structure of real-world language. In our study, we position these contextual embeddings as a shared feature space mediating brain-to-brain coupling, a common basis onto which both speaker and listener converge. Our model-based brain-to-brain coupling framework quantifies how closely speaker and listener are aligned to this explicit model of linguistic structure—word-by-word in real-time, interactive conversations. This modeling framework allows us to effectively track the flow of linguistic information from one brain to another in unconstrained dialogues. Critically, unlike previous applications of encoding models to language comprehension, our approach addresses the specific challenge of real-time, dyadic interactions: each participant's brain is affecting the other participant's brain, resulting in a unique, irreplicable experience in each conversation. We use the term "model-based coupling" to express this effort to explicitly model real-time, dyadic processing between two interacting brains.

We used the trained encoding models (i.e., the learned weight matrices) from the within-speaker and within-listener analyses and assessed how well the models generalize across brains, from speaker to listener (and vice versa) at varying lags. Typically, separate encoding models are trained and tested within each subject, within each electrode/voxel, and evaluated in terms of how well they

generalize to novel stimuli.[73,74] Previous research has developed methods for evaluating how encoding models generalize across subjects,[75,76] across different brain areas,[77] and across different tasks/processes, like reading and listening.[78] In the current manuscript paper, our scientific question demanded that we develop a framework for assessing five types of generalization simultaneously: testing encoding model generalization (1) across segments of the stimulus (using 10-fold cross-validation), (2) across subjects (within speaker–listener dyads), (3) across different brain regions (e.g., from SM to STG electrodes), (4) across tasks/processes (speaking/production and listening/comprehension), and (5) across lags[79] (e.g., speaker pre-word onset to listener post-word onset).

The production models are estimated separately at each lag relative to word onset (and therefore yield different predictions for each lag). For this reason, we computed correlations between the model-predicted activity and actual neural activity at each pair of lags ranging from –4 s to +4 s between speaker and listener. This analysis is not strictly symmetric if performed in the opposite direction: model-based predictions from the encoding model estimated in the listener (the comprehension model) evaluated against the speaker's actual neural activity. For this reason, we performed the same intersubject encoding analysis for the listener: we computed comprehension model predictions derived from the listener, correlated these with the speaker's actual neural activity, then averaged these correlations with those computed from the production model and evaluated against the listener's neural activity. In practice, the results for both directions are very similar (Figure S2).

This analysis yields model-based predictions (and the corresponding actual neural activity) for every word at each lag and each electrode. To construct a lag-by-lag matrix that summarizes across electrodes (e.g., across all electrodes as in Figure 3, or across electrodes within a given language area as in Figure 4), we averaged the model-predicted activity and actual activity (retaining lags) prior to computing the correlation between predicted and actual neural activity. This correlation value is visualized at each pair of lags. We perform this process for each speaker–listener pair, and the final correlations are summarized across dyads using a weighted average, where the weights correspond to the relative number of words for each dyad; this weighted averaging procedure mitigates variance in correlations derived from small training/test folds in dyads with fewer words. Each row of the resulting matrix indicates how well the model-based predictions for a given speaker lag (the row index) match the listener's brain activity at each lag. The diagonal represents the same (i.e., matching) lag between speaker and listener; anything below the diagonal indicates that the speaker precedes the listener. The bottom half of the matrix corresponds to speaker–listener linguistic coupling based on the speaker's brain activity prior to word onset. The right half of the matrix corresponds to the speaker–listener linguistic coupling based on the listener's brain activity after word onset. The bottom right quadrant of the matrix corresponds to pairs of lags where the linguistic content of the speaker's brain prior to word onset is coupled to the linguistic content of the listener's brain after word onset.

This same procedure was applied across regions within-speaker and within-listener. Usually, encoding models within subjects are trained on one electrode (and one lag), and the correlation between the model's predictions and the actual activity of the same electrode (and lag) is evaluated on a held-out test set. Following the same procedure as above, we can evaluate the encoding model predictions for one region against another region. This tests whether two brain regions in the speaker (or listener) use the same linguistic features from the embedding space. We ran this procedure for all pairs of regions within-speaker (Figure 4A and S13) and within-listener (Figures 4C and S14).

In all cases, performance is evaluated by computing correlations across the same sets of test words. In the case of the core model-based coupling analyses, we use the same predictions generated from the models trained for within-subject encoding analysis. We then compute correlations between the model-predicted activity derived from one subject (e.g., the speaker) and actual activity in the other subject (the listener) across a left-out set of test words unseen by the model. In the within-subject model-based connectivity analyses, correlations are computed between model-predicted activity derived from one region and actual activity in another region (in the same subject) across the test words. In the ISC analyses, correlations are computed between actual brain activity in one subject (e.g., the speaker) and another (e.g., the listener) for the matching set of test words.

For comparison, we computed intersubject correlation (ISC)[5,80] by applying the same procedure as above, but instead of using predicted neural activity from encoding models, we computed the correlations between the actual speaker neural activity and actual listener neural responses for the corresponding test sets.

Note that in all cases, we first fit encoding models separately for each electrode (and each lag); this is the highest level of precision for model fitting. There are several possible ways to summarize model performance across electrodes. We assessed three different summarization methods to ensure our core results were not an artifact of a particular summarization procedure. First, we averaged the model-predicted activity for the test words across all electrodes in the brain (Figure 3 and S1A) or within ROIs (Figure 4) and correlated this with the actual activity for the test words averaged across all electrodes. Second, we computed the correlation between model-predicted and actual test activity for every possible pair of electrodes from speaker to listener (Figure S1B). Although averaging improved our predictive power, the results were similar in both versions. Third, we averaged the model-predicted (and actual) activity within ROIs and then correlated the ROI-averaged model-predicted and ROI-averaged actual activity across brains. We then averaged the 64 ROI-pair correlation matrices into one whole-brain correlation matrix to summarize across the entire brain (Figure S1C). This approach yielded qualitatively similar results to the summarization methods used in panels A and B.

Model comparison framework

We use a formal model comparison framework to make more specific claims about what kinds of information are encoded in brain activity. Our approach follows the logic of the "late commitment"[81] and "system identification"[82] frameworks, where hypotheses are formulated as explicit computational model comparisons evaluated against naturalistic brain data using out-of-sample prediction.

For example, we evaluate the hypothesis that contextual information is encoded in neural activity during natural conversations by comparing encoding model performance for contextual embeddings from GPT-2 to encoding model performance for non-contextual embeddings extracted from GPT-2 (Figures S4 and S5). The model's architecture, objective function, and dimensionality are held constant in this comparison. The static ''control'' embeddings—the pre-contextual embeddings from the input layer of the model—capture only word-level lexical-semantic information and do not incorporate contextual information from preceding words; on the other hand, the contextual embeddings we evaluate have passed through multiple layers of the transformer where the self-attention mechanism incorporates contextual information from previous words into the current embedding. In this study, we do not aim to explicitly disentangle the overarching context from the meaning of individual words,[77] or explore varying scales of prediction.[83]

We compare encoding performance for embeddings with two additional feature sets inspired by classical psycholinguistics. First, we include a phonemic model capturing articulatory speech features.[28] Specifically, for each phoneme, we constructed a unique binary vector with 22 features[84]: consonant features include the manner of articulation, place of articulation, and whether the consonant was voiced or unvoiced; vowel features included height (e.g., high, mid, low) and front-to-back position (Table S2). Second, we include a syntactic model based on two complementary feature sets: we used SpaCy to extract part-of-speech tags for each word, as well as syntactic relations between words (based on a dependency parse tree) in an utterance (e.g., nominal subject, modifier, root, etc), resulting in a 75-dimensional binary vector (Tables S3 and S4). Unlike the continuous, vector-space embeddings capturing contextual (or non-contextual lexical-semantic) meaning, these indicator variables capture the occurrence of discrete, symbolic labels for articulatory or syntactic features. We submit these features to encoding analysis using the same procedure we used for contextual embeddings (Figures 3 and S4).

QUANTIFICATION AND STATISTICAL ANALYSIS

To evaluate the statistical significance of the lag-by-lag intersubject encoding matrix, we generated a null distribution of intersubject encoding for each pair of lags based on phase-randomized neural signals. We re-estimated encoding models for each of the 1,000 phase permutations. For the whole-brain analysis (Figure 3A), we computed 1,000 iterations of the lag-by-lag correlation matrix using the phase-randomized models for both speaker and the listener. Specifically, we correlated the speaker's phase-randomized model's predicted neural activity with the listener's *actual* neural activity and vice-versa (i.e., the same as intersubject encoding, except with the perturbed model). This procedure used the same selected electrodes as the non-randomized analyses. The maximum value across all pairs of lags in each matrix was submitted to the null distribution to control for multiple tests across pairs of lags. Lag pairs were considered significant at an alpha value of 1% based on this distribution, corresponding to an intersubject encoding performance (correlation) value of .052 ($p < .01$, FDR corrected) (Figures 3B and S3).

To evaluate the significance of inter-regional intersubject encoding analysis (Figure 4), we generated a null distribution using the encoding models estimated from phase-randomized neural data. The lag-by-lag intersubject encoding correlation matrix was computed once for each pair of speaker–listener regions (8 x 8), resulting in 1,000 permutation correlations for each pair of regions and each pair of lags (129 x 129). Because of unique electrode coverage per participant and the electrode selection process, the number of subjects and electrodes in each pair of regions differed. To ensure the reliability of these correlations across subjects, we only considered pairs of regions with at least four subjects (Figures S13–S15). We computed p-values for each pair of regions and each pair of lags based on the corresponding null distribution, then used Bonferroni correction to control the family-wise error rate at 5% across pairs of ROIs and pairs of lags. Any resulting pair of ROIs with at least 20 adjacent pairs of significant lags were deemed significant overall (gray lines in Figure 4). The same statistical analysis was applied to the within-speaker (Figure S13) and within-listener (Figure S14) region pairs.

We also tested for a significant difference between intersubject encoding using contextual, model-trained embeddings against static and untrained model embeddings (Figures 3 and S5). To do this, we compared the observed difference against the 10,000 permutation differences, where we randomly assigned the observation per speaker (10) and per fold (10) to either sample. The resulting p-values were correlated for multiple comparisons with FDR at $p < 0.01$.