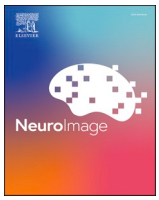
NeuroImage 311 (2025) 121180

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



The cortical architecture representing the linguistic hierarchy of the conversational speech

Ruhuiya Aili [a,1], Siyuan Zhou [b,1], Xinran Xu [a], Xiangyu He [a], Chunming Lu [a,*]

[a] State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, 100875, China
[b] Institute of Brain and Psychological Sciences, Sichuan Normal University, Chengdu, 610066, China

ARTICLE INFO

Keywords:
Conversational speech
Linguistic hierarchy
Cortical architecture
Turn
Topic
fNIRS hyperscanning

ABSTRACT

Recent studies demonstrate that the brain parses natural language into smaller units represented in lower-order regions and larger units in higher-order regions. Most of these studies, however, have been conducted on unidirectional narrative speech, leaving the linguistic hierarchy and its cortical representation in bidirectional conversational speech unexplored. To address this gap, we simultaneously measured brain activity from two individuals using functional near-infrared spectroscopy (fNIRS) hyperscanning while they engaged in a naturalistic conversation. Using a Pre-trained Language Model (PLM) and Representational Similarity Analysis (RSA), we demonstrated that conversational speech, jointly produced by two interlocutors in a turn-taking manner, exhibits a linguistic hierarchy, characterized by a boundary effect between linguistic units and an incremental context effect. Furthermore, a gradient pattern of shared cortical representation of the linguistic hierarchy was identified at the dyadic rather than the individual level. Interpersonal neural synchronization (INS) in the left superior temporal cortex was associated with turn representation, whereas INS in the medial prefrontal cortex was linked to topic representation. These findings further validated the distinctiveness of linguistic units of different sizes. Together, our results provide original evidence for the linguistic hierarchy and the underlying cortical architecture during a naturalistic conversation, extending the hierarchical nature of natural language from unidirectional narrative speech to bidirectional conversational speech.

1. Introduction

During daily language communication, connected natural language, such as discourse, is more commonly employed to convey high-level conceptual information and implicit social intentions compared to isolated linguistic units, such as a single word or phoneme. Recent neurophysiological evidence suggests that the brain parses connected natural language into the smallest linguistic units, such as phonemes or syllables, and progressively combines these units into nested structures—words, sentences, paragraphs, and discourse—forming a hierarchical organization across multiple timescales (Bornkessel-Schlesewsky et al., 2015; Chang et al., 2022; Hickok and Poeppel, 2004). Here, we refer to linguistic units as phonemes, words, sentences, etc., while linguistic hierarchy denotes the hierarchical organization of these linguistic units.

Moreover, our brain entrains its neural responses to such a linguistic hierarchy at timescales corresponding to each linguistic unit to achieve speech comprehension (Ding et al., 2016; Jin et al., 2018; Lerner et al., 2011; Xu et al., 2005). For example, by employing naturalistic stimuli such as movies or storylines, previous studies have shown that a listener's cortical activity can dynamically track the linguistic structures of stimuli across different timescales, ranging from phonemes and words to sentences and discourse (Ding et al., 2016; Keitel et al., 2018; Luo and Ding, 2020). Notably, such a linguistic hierarchy differs from syntactic structure in that it reflects the human brain's selective sensitivity to linguistic units of different timescales. Additionally, cortical activity demonstrates a spatially gradient architecture, with early auditory regions and the superior temporal cortex responding more to smaller units, such as phonemes and words, while the parietal and frontal cortices respond more to larger units, such as sentences and paragraphs (Lerner et al., 2011; Hasson et al., 2015; Yeshurun et al., 2017). However, previous studies have primarily investigated the linguistic hierarchy of unidirectional narrative speech and its underlying cortical

* Corresponding author at: State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, 100875, China.
E-mail address: luchunming@bnu.edu.cn (C. Lu).
[1] Contributed equally.

https://doi.org/10.1016/j.neuroimage.2025.121180
Received 10 September 2024; Received in revised form 24 March 2025; Accepted 28 March 2025
Available online 28 March 2025

architecture in the listener's brain (Hasson et al., 2015). Thus, little is known about the linguistic hierarchy of turn-taking-based bidirectional conversational speech and how such a hierarchy, if it exists, is co-represented in the brains of interlocutors.

There are crucial differences between unidirectional narrative speech, solely produced by an individual, and bidirectional conversational speech, jointly produced by two individuals. Previous linguistic theories have suggested that joint activities, such as conversations, lead to the emergence of shared representations; that is, two individuals not only understand but also anticipate and align with each other's mental states and intentions through dynamically shifting their roles (Hasson and Frith, 2016; Jiang et al., 2021; Pickering and Garrod, 2004). For example, Pickering and Garrod's (2004) theory suggests that interlocutors automatically align and form shared representations at all linguistic levels during conversation, from low-level phonemic representations to high-level semantic representations of the entire conversation as the interaction progresses (Pickering and Garrod, 2004; Stolk et al., 2016). Additionally, certain features unique to conversations, such as topic switching (corresponding to discourse in a narrative) and turn-taking (corresponding to paragraphs in a narrative), contribute to the formation of a dyadic-specific hierarchical structure (Levinson, 2016; Speer et al., 2024). Meanwhile, features universal to both individual speech and conversations, such as single sentences, may lead to similar characteristics between individual and dyadic levels of the linguistic hierarchy. To the best of our knowledge, however, little empirical neurocognitive evidence has been reported to suggest a parsing mechanism by which the brain segments naturalistic conversational speech into linguistic units of different timescales, thereby revealing a linguistic hierarchy at the dyadic level.

The present study aimed to fill the gap in the literature by investigating the potential linguistic hierarchy and its cortical representation during a naturalistic conversation task, combining fNIRS hyperscanning with a transformer-based PLM. Hyperscanning simultaneously measures the real-time brain activity of two or more individuals, enabling unprecedented exploration of neural mechanisms involved in real-time interpersonal interactions (Montague et al., 2002). Previously, Jiang et al. (2012) conducted the first hyperscanning study on naturalistic bidirectional conversation and found that face-to-face communication, compared to back-to-back communication, significantly increased INS in the left inferior frontal gyrus between interlocutors. Dai and colleagues (Dai et al., 2018) examined interpersonal interactions in a "cocktail party" scenario and discovered selectively enhanced INS at the temporoparietal junction between the listener and the attended speaker compared to the listener and the unattended speaker. Later, Salazar et al. (2021) and Liu et al. (2019) found that neural synchrony between communicators in the frontal lobe, right superior temporal sulcus, and temporoparietal junction reflects shared cognitive processing of semantics and syntax rather than simply shared physical stimuli. Thus, INS appears to be a reliable and valid indicator of the cortical representation of interpersonal interaction at the dyadic level. Although fNIRS can only measure the outer cortical layers of the human brain with limited temporal and spatial resolution, it is currently the only technique that provides high ecological validity alongside relatively acceptable spatial resolution compared to electroencephalogram (EEG) and functional magnetic resonance imaging (fMRI).

We focused on linguistic units such as turns and topics, which are larger in timescale than words and sentences and unique to bidirectional conversational speech. A "turn" was defined as a sequence of utterances from a single speaker until the other speaker assumed the leading role, regardless of brief interjections or fillers from the listener. We specifically tested the human brain's ability to parse turns and topics based on the boundary effect. Specifically, semantic similarity was expected to be higher between sentences in the same turn than across different turns, and higher between turns in the same topic than across different topics, thereby demonstrating a linguistic hierarchy. Three alternative hypotheses were proposed: First, according to the shared representation hypothesis, we predicted that turns and topics would be hierarchically represented in conversational speech at both the dyadic and individual levels. Furthermore, we expected a cortical architecture with a gradient pattern supporting the representation of the linguistic hierarchy in conversational speech at both the dyadic (i.e., INS) and individual levels. This architecture would likely involve primary auditory and motor regions being more closely associated with the representation of smaller linguistic units, while associative brain regions such as the parietal and frontal cortices would be more closely associated with the representation of larger linguistic units. Second, since turns and topics are linguistic units unique to conversations, we hypothesized that these levels would exist only in the form of shared representations. In other words, no corresponding linguistic hierarchy or related brain responses would be observed at the individual level for these linguistic units. Finally, the third hypothesis predicted that both the linguistic hierarchy and cortical architecture would only be identified at the individual level.

2. Methods

2.1. Participants

Prior to the experiment, G* Power 3.1 was used to estimate the required sample size. The analysis revealed that 26 dyads (52 individuals) were necessary to achieve a statistical power of 0.8 with $\eta_p^2 = 0.25$ and $\alpha = 0.05$ for examining the difference in semantic similarity between linguistic units. Finally, 70 healthy adults (36 females) were recruited through advertisements in Beijing. The mean age of participants was 22.13 years (*standard deviation [SD]* = 2.66), and the mean years of education was 16.02 (*SD* = 2.47). All participants were right-handed, with no neurological or psychiatric disorders based on self-report and had normal or corrected-to-normal vision. Participants were randomly paired into 35 same-gender dyads. After data quality inspection, 32 and 30 dyads were left with valid data for the Familiar and Unfamiliar conditions, respectively.

The study protocol was approved by the Institutional Review Board of Beijing Normal University. All participants were provided with written informed consent.

2.2. Tasks and procedures

During the task session, first, 20 candidate topics were selected from a previous study (Zhou et al., 2023). These topics were assessed to best fit into the cultural style of China according to an assessment of 67 Chinese adults (35 females, mean age = 24.4 ± 2.94), thus providing our participants with cues to stimulate a sufficient number of turns during the conversation. For each dyad, each participant was requested to report their perceived familiarity level with each of the 20 topics on a 10-point Likert scale (1 representing the lowest level, 10 representing the highest level). The scores of two participants in a dyad were averaged to obtain a dyadic-level score for topic familiarity. For each dyad, the topic perceived to have the highest level of familiarity was used in the Familiar condition, while the topic with the lowest level of familiarity was used in the Unfamiliar condition. It should be noted that only one topic was used for each dyad of each condition based on their report on topic familiarity, but different topics might be used for different dyads. We used the Kruskal-Wallis H test to examine whether the numbers of dyads involved in each topic were balanced.

The experiment was conducted in a quiet room. During the experiment, the two participants in a dyad sat across from each other at a table in a face-to-face manner (Fig. 1a). Initially, a 5-minute resting-state session with eyes closed served as a control session to remove spontaneous brain activity unrelated to the external stimuli. The task sessions immediately followed the resting-state session, during which the dyads freely conversed on a topic that was familiar to them for 5 min and 30 s.

Previous studies have indicated that the decrease in the familiarity of

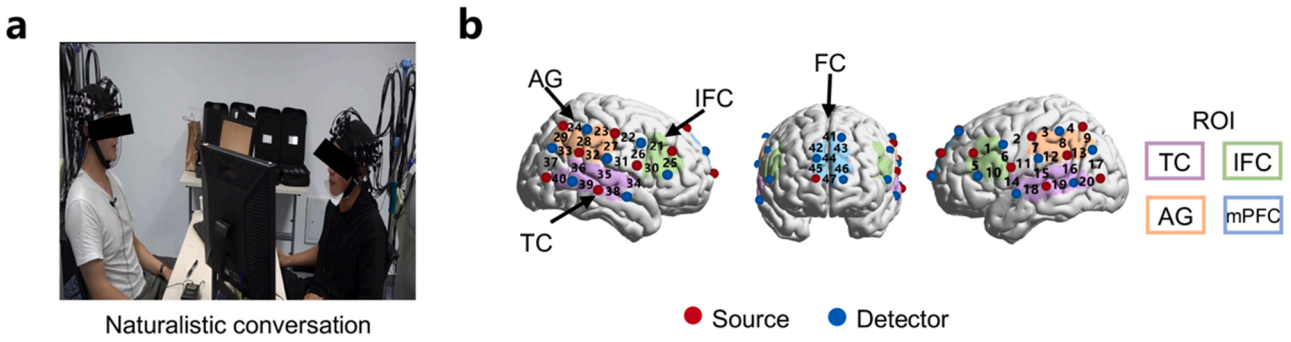


Fig. 1. The setup of the experiment. (a) Each dyad conversed on a topic for 5 min and 30 s. (b) The probe set was placed on the bilateral frontal, temporal, and parietal cortical layers. CH18, CH38, and CH47 were placed on T7, T8, and Fpz, respectively, according to the international 10–20 system. Colored areas denote regions of interest (ROI). TC: temporal cortex; IFC: inferior frontal cortex; AG: angular gyrus; mPFC: medial frontal cortex.

a topic can impact the fluency and quality of a conversation, but the processing of smaller linguistic units such as turns is not impacted (Rivers, 2018). Therefore, participants were additionally asked to freely converse on a less familiar topic (Unfamiliar condition). The purpose of this condition was to test the distinctiveness of linguistic units such as turns, i.e., whether the change in topic familiarity would impact the representation of turns, thereby further validating the linguistic hierarchy and cortical representation of turns. The order of the Familiar and Unfamiliar conditions was counterbalanced. Additionally, both participants in a dyad were hyperscanned using fNIRS to collect their hemodynamic signals in both conditions (Fig. 1a-b). The entire experimental procedure was video-recorded for subsequent behavioral analyses.

2.3. fNIRS data acquisition

Two LABNIRS systems (Shimadzu Corporation, Japan) were used to collect fNIRS data. For each system, three customized sets of 47 measurement channels (CHs) were used to collect data from each participant (Fig. 1b). Two sets of probes were used to cover the bilateral frontal, temporal, and parietal cortical layers, while the rest were used to cover the prefrontal cortex. The anatomical location of each channel was determined using the international 10–20 system. Specifically, CH18, CH38, and CH47 were placed on T7, T8, and Fpz, respectively. The positions of the probe sets were adjusted before the experiment to ensure consistency between the two individuals in a dyad and among dyads.

To confirm the anatomical locations of the optode probes, we obtained magnetic resonance imaging (MRI) data from 2 female and 2 male participants who wore plastic caps with the probes' true positions marked using Vitamin E balls. The high-resolution, T1weighted, magnetization-prepared, and rapid gradient-echo sequence was used (time repetition = 2530 ms; time echo = 3.30 ms; flip angle = 7°; slice thickness = 1.3 mm; in-plane resolution = 1.3 × 1.0 m^2; and number of inter-leaved sagittal slices = 128). Statistical Parametric Mapping 12 (Wellcome Department of Cognitive Neurology, London, UK) was used to normalize the MRI data to the standard Montreal Neurological Institute (MNI) coordinate space. The MNI coordinates of probes were generated according to the Automated Anatomical Labelling template using the NIRS_SPM toolbox (see Table S1). Based on this information, we were able to assess the consistency between the probes' true positions and the expected anatomical positions and to adjust the probes' true positions. This procedure was repeated several times until the true positions and the expected positions reached a high level of consistency, e.g., the difference in probabilities was < 10 % for 3 out of 4 scanned participants.

The optical density of near-infrared light (780, 805, and 830 nm) was measured at a sampling rate of 8.33 Hz. Then, changes in oxyhemoglobin (HbO), deoxyhemoglobin, and total hemoglobin concentration were calculated based on the modified Beer–Lambert law. Previous research has indicated that oxyhemoglobin is a sensitive marker for reflecting changes in local cerebral blood flow and offers a higher signal-to-noise ratio (Hoshi, 2007). Therefore, this study focuses solely on the HbO signal.

2.4. Behavioral data analysis

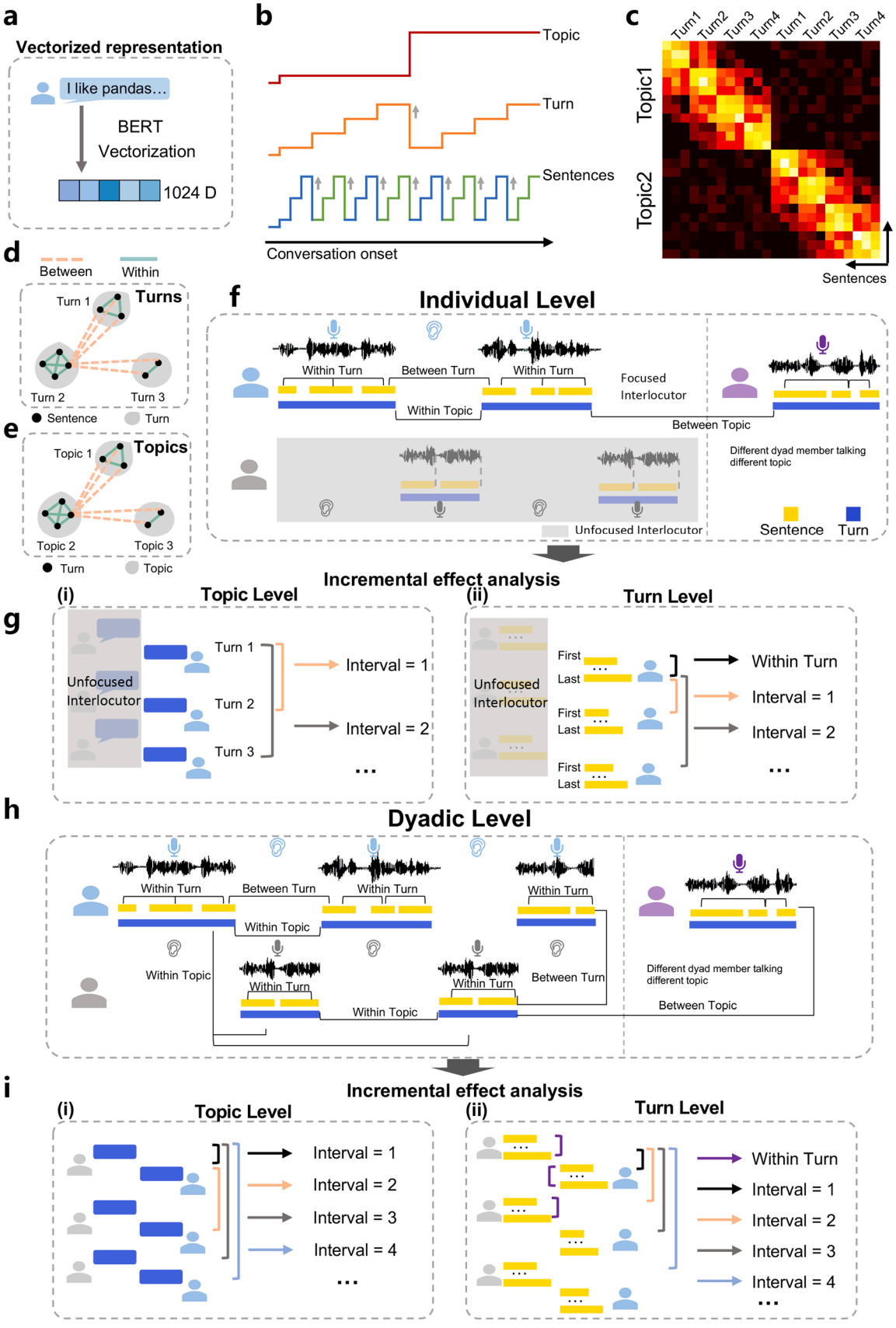
2.4.1. The linguistic characteristics of the conversational speech

For the conversational speech, five additional participants who were not involved in the conversation task were recruited to transcribe the recorded conversational speech into text using Nvivo (version 11) and Praat (version 6.3.01) software. For each video, these participants annotated the onset and offset of each sentence and turn for each interlocutor. To ensure consistency in transcription, we established strict guidelines requiring transcribers to follow a specific protocol (see the Supplementary Text 3). This approach guaranteed uniformity among transcribers throughout the project. Moreover, during the transcription process, various linguistic phenomena, such as empty pauses and filled pauses (e.g., "emm", "eeh"), disfluencies (e.g., hesitation, stuttering, interruption, mistakes, etc.), were carefully annotated. The proportion of these linguistic phenomena was 3.172 ± 2.013 %, and they were excluded from all analyses due to their semantic ambiguity for use with the PLM.

Based on the transcript, various indexes of the conversation were calculated, such as the average duration of sentences and turns, the average number of sentences and turns generated by each individual participant, and the average number of function and content words in sentences and turns. A linear mixed-effect model (LME) was used to test the difference in the number of content and function words across various topics, with word type (content vs. function words) and topic as fixed variables, and participant identity as a random variable.

2.4.2. Vectorized representation of the conversational speech using PLM

The Whole Word Masking (WWM) - RoBERTa model, downloaded from HuggingFace (Cui et al., 2021), was used to vectorize the transcribed text of conversational speech (Fig. 2a). As a derivation of Bidirectional Encoder Representations from Transformers (BERT), this model is trained on words rather than characters, making it suitable for Mandarin Chinese. This model features 24 layers of blocks, each with 16 attention heads, a hidden size of 1024, and approximately 340 million parameters in total. In the literature, there are two major approaches to vectorizing sentences: One uses the special "[CLS]" token to capture the overall feature of the input sequence, while the other averages the vectors of all words in a sentence to obtain a vectorized representation of the sentence. Recent evidence suggests that the method using the "[CLS]" token may slightly underperform other approaches in testing fMRI prediction capabilities across layers (Anderson et al., 2021). Thus, in this study the second approach was employed by averaging word vectors from the second-to-last hidden layer. Similar procedures were applied to the vector representation of turns by averaging word vectors



(caption on next page)

Fig. 2. The analytic pipeline of the linguistic hierarchy. (a) Vectorized representation of the conversational speech using PLM. (b) Hypothesized linguistic hierarchy in the speech. Sentences spoken by individual interlocutors (indicated by green and blue lines for different speakers) are chunked into turns and co-represented at the dyadic level (shown by the orange line). Turns are further segmented into topics (shown by the red line), which are also co-represented at the dyadic level, forming a linguistic hierarchy. (c) The illustrated semantic similarity matrix. Each grid represents cosine similarity between different sentences produced in conversation. Based on the hypothesized linguistic hierarchy, smaller units (i.e., sentences and turns) within the same larger units (i.e., turns and topics) should exhibit higher similarity than those across larger units, demonstrating a boundary effect. Additionally, the similarity between adjacent sentences or turns within the same turn or topic should be higher than that between non-adjacent ones, demonstrating a gradual decrease in similarity as the interval between them increases, showing the incremental context effect. (d-e) A demonstration of the relationship between linguistic units in the hierarchy. Gray circles represent different turns (d) or topics (e), with black dots within each circle denoting different sentences (d) or turns (e). Orange dashed lines signify "across", while green solid lines represent "within". (f) Calculating the semantic similarity at the individual level. The blue icon is focused as an example. (g) Incremental context effect at the individual level. Intervals between turns were coded as 1, 2, 3, and ≥3, respectively, and semantic similarity between each interval was calculated. (h) Calculating the semantic similarity at the dyadic level. In this case, both the blue and grey icon are focused. (i) Incremental context effect at the dyadic level. Intervals between turns were coded as 1, 2, 3, 4, and ≥4, respectively, and semantic similarity was calculated for at each interval.

across all sentences within a turn.

Identification of the linguistic hierarchy by testing the boundary effect. To confirm the linguistic hierarchy at the individual level, the speech produced by each individual interlocutor was analyzed (Fig. 2b-f). Previous studies have shown that when processing continuous stimuli, such as films or narratives, the brain segments them into discrete events with various temporal scales along the hierarchy, and each event is marked by a distinct neural population pattern (Baldassano et al., 2017; Lee and Chen, 2022). In the processing of speech, it is usually expected that the similarity between smaller linguistic units (such as sentences or turns) within larger units (such as turns or topics) is higher than that across larger units, showing a boundary effect (Fig. 2c). Therefore, in this study, we tested whether turns or topics in bidirectional conversational speech would be segmented, showing the boundary effect at both the semantic and brain response levels.

At the individual level, the speech produced by a specific interlocutor was extracted. As expected, this speech included multiple turns and sentences. At the turn level, based on these vectors, cosine similarity was computed between the last sentence of a given turn (e.g., $turn_i$) and first sentence of all other turns (e.g., $turn_{N-i}$) and then averaged to index the semantic similarity across turns (Fig. 2d) (Hoffman et al., 2018). Cosine similarity is a common method for calculating the similarity between high-dimensional semantic vectors, as it captures the relationship more precisely through the angle between the vectors rather than their magnitude (Haxby et al., 2014). A paired two-sample *t*-test was conducted on the averaged semantic similarity between within- and across-turns at the group level (i.e., in total 64 individuals). Then, for the topic level, similar procedures to the above were applied to the vector representation of turns by averaging word vectors across all sentences within a turn. Semantic similarity was calculated between a given turn of an interlocutor (e.g., $Individual_i$) and all turns of all other individuals (e.g., $Individual_{N-i}$) who had selected the same topic during their conversations. Similarities among those turns were further averaged to generate an index of the semantic similarity within topics. In a parallel manner, semantic similarity was determined between a given turn of an interlocutor (e.g., $Individual_i$) and all turns of all other individuals (e.g., $Individual_{N-i}$) who had chosen different topics. The similarity was also averaged across topics and a paired sample *t*-test was performed on these averaged semantic similarities between within- and across- topics at the group level (i.e., in total 64 individuals).

Next, the same analyses were conducted at the dyadic level, where the speech vectors produced by both interlocutors in a dyad were used (Fig. 2h). At the turn level, the semantic similarity of sentences within turns was calculated in the same manner as that of the individual level but was then averaged between the two interlocutors of a dyad, generating an index of the dyadic-level semantic similarity within turns. To calculate the semantic similarity across turns, the similarity was calculated between each sentence for individual A in a given turn (e.g., $turn_i$) and all sentences from other turns (e.g., $turn_{N-i}$) within the same interlocutor A or individual B. These similarity values were then averaged to derive the semantic similarity of across turns for interlocutor A. Finally, the values from both interlocutors (i.e., interlocutors A and B) were averaged to produce an index of dyadic-level sentence similarity for across turns. For topic level, the calculation of semantic similarity was the same as at the individual level, but the similarity was averaged between the two interlocutors of a dyad, generating indexes of the dyadic-level turn similarity both for within and across topics. Finally, paired sample *t*-tests were conducted to compare semantic similarities between within and across turns, as well as between within and across topics (Fig. 3d-e).

2.4.3. Validation of the linguistic hierarchy by testing the incremental context effect

A recent study (Schapiro et al., 2013) suggested that the similarity between linguistic units or events declines as the interval between units increases, showing an incremental context effect (Fig. 2c). Here we further predicted that, with the increasing number of turns within an interval, the semantic similarity between turns would decline.

To test this, we defined different lengths of intervals (i.e., 0, 1, 2, 3 and ≥ 3) to depict the increase in context length (Fig. 2f). For turns at the individual level (Fig. 2g-ii), when interval = 0 (i.e., within turns), semantic similarity was calculated between the last sentence and the first sentence of $turn_i$ within the same interlocutor. For interval = 1, semantic similarity was calculated between the last sentence of $turn_i$ and the first sentence of $turn_{i+1}$ within the same interlocutor. When interval = 2, 3, or ≥3, semantic similarity was calculated between the last sentence of $turn_i$ and the first sentence of $turn_{i+2}$, $turn_{i+3}$, $turn_{i+4}$, etc., within the same interlocutor. The similarity of turns was further averaged for each interval within an interlocutor.

At the dyadic level, the turns produced by both interlocutors were concatenated into a single speech stream and coded as $i, i+1, i+2, i+3, i+4, \ldots, N$ (Fig. 2h). Turns coded with odd numbers were produced by interlocutor A, while those coded with even numbers were produced by interlocutor B. At the turn level (Fig. 2i-ii), interval = 0 (i.e., within turn) referred to the semantic similarity between the last sentence and the first sentence of $turn_i$ by the same interlocutor A or B. Interval = 1 referred to the semantic similarity between the last sentence of $turn_i$ produced by interlocutor A and the first sentence of $turn_{i+1}$ produced by interlocutor B. Interval = 2 indicated the semantic similarity between the last sentence of $turn_i$ of interlocutor A and the first sentence of $turn_{i+2}$ also produced by interlocutor A. The same principle of calculation was applied to other intervals. The similarity was then further averaged for each interval.

A permutation test was employed to determine the statistical significance of semantic similarity for each interval. This test was conducted by randomly shuffling the order of each participant's turns and then recalculating the semantic similarity. This procedure was repeated 1000 times to generate a null distribution, and the p-value significance was obtained based on the position of the true value in the null distribution ($p < 0.05$). Next, the same procedure was employed to calculate the potential differences between the null distributions of different intervals. The position of the actual difference within the null distribution was used to compute the p-value. Pairwise comparisons were also conducted among different intervals (Fig. 3e-ii). The False Discovery Rate

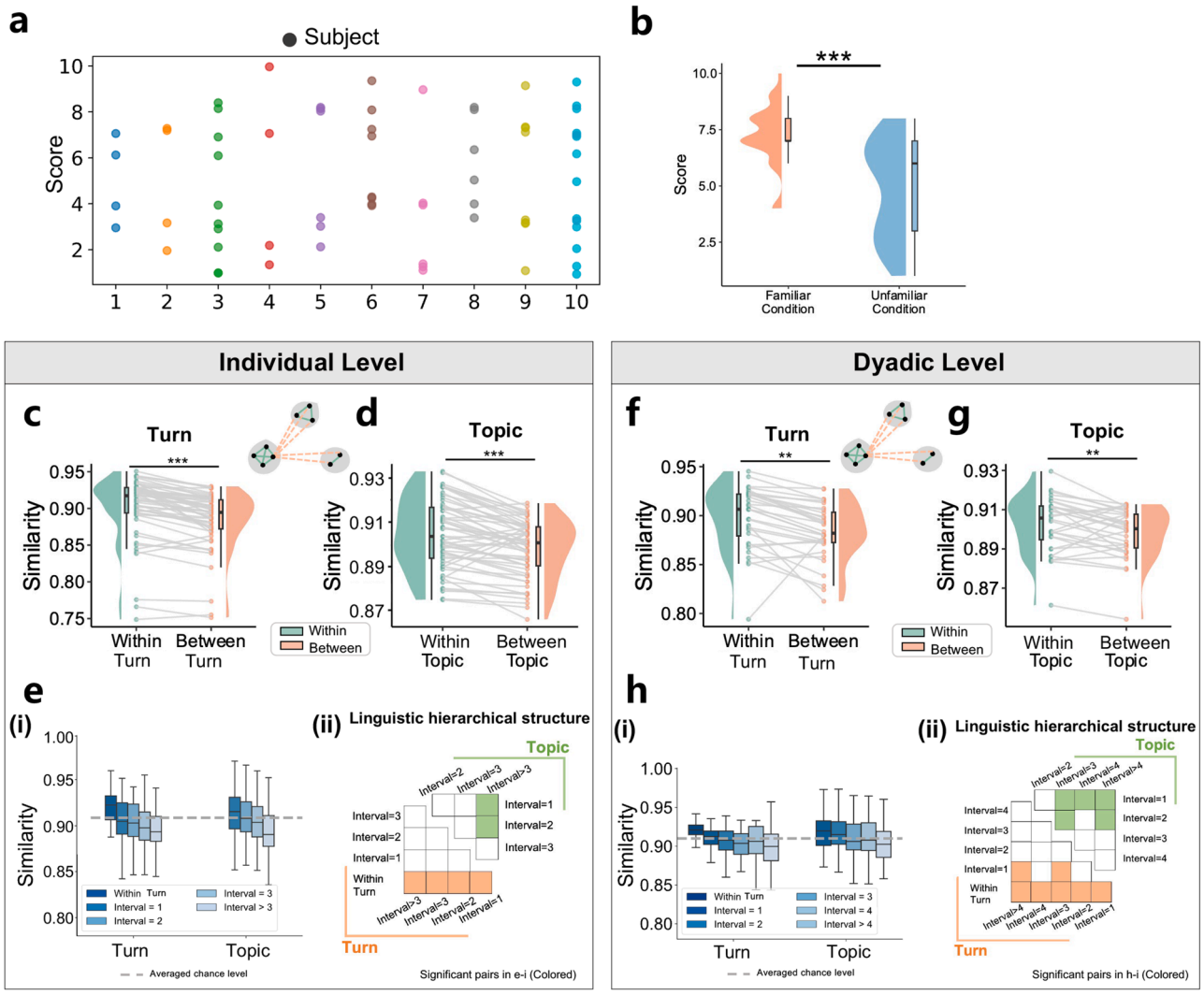


Fig. 3. (a) Participants' rating of familiarity on each of the 10 topics, with each dot representing a participant. (b) As expected, a significant difference was found between familiar and unfamiliar conditions in familiarity scores. Linguistic hierarchy of the conversational speech in the Familiar condition. (c-e) and (f-h) show results at the individual and dyadic levels, respectively. (c-d) and (f-g) show boundary effects of both turns and topics. Semantic similarity between sentences was significantly higher within turns (green) than across turns (orange). Similarly, the semantic similarity of turns was significantly higher within topics (green) than across topics (orange). In (e) and (h), semantic similarity between turns declines as the interval between them increases, showing an incremental context effect. The left panel shows the decline in semantic similarity as the interval increases at the turn or topic level. The gray dashed line represents the chance level. Notably, each interval has its own chance level, here we averaged across intervals for visualization purpose. The right panel indicates the statistical significance in pairwise comparisons between intervals. The results in both (e) and (h) are corrected using the FDR method ($q < 0.05$). **, $p < 0.01$; ***, $p < 0.001$.

(FDR) method was used to correct for multiple comparisons (Fig. 3e).

For the incremental context effect of topics, at the individual level, semantic similarity was calculated between turn_i and turn_{i+n} of the same interlocutor; however, there was no within turn, i.e., interval = 0. The similarity of sentences or turns was further averaged within an interlocutor for each interval. Then, at the dyadic level, interval = 1, 2, 3, 4, and ≥4 indicated the semantic similarity between turn_i and turn_{i+1}, turn_{i+2}, turn_{i+3}, etc. The similarity was then further averaged across dyads for each interval. The permutation tests as same as above were conducted.

2.5. fNIRS data analyses

2.5.1. Preprocessing

First, we conducted a quality check on the fNIRS data. In this study, sliding time windows with a length of 15 s were used to inspect artifacts. Data points lying beyond 3 standard deviations within the sliding window were identified as artifacts. If the artifact percentage of a channel exceeded 5 %, that channel was marked as bad. A participant was excluded if the proportion of bad channels exceeded 30 % of the total number of channels. No individuals or channels were excluded at this step.

Secondly, data from the first and last 15 s of each session were removed before preprocessing to ensure the fNIRS signals to reach a steady state. Then, the HbO data were detrended and corrected for motion artifacts using Discrete Wavelet Transformation method. Next, to remove global physiological noises such as changes in scalp blood flow, Principal Component Analysis (PCA) was applied to eliminate the first 80 % of signal variability. Additionally, a band-pass filter (0.01–0.5 Hz) was utilized to exclude high-frequency noise and low-frequency physiological noise.

2.5.2. Definition of ROIs

To increase statistical power in subsequent analyses, channels that passed the statistical tests of RSA were grouped into ROIs. According to the Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer

et al., 2002), channels were assigned to specific ROIs based on their MNI coordinates. A channel was categorized into an ROI if the probability of its MNI coordinate falling within the corresponding brain region exceeded 50 %. The preprocessed signals from all channels within each ROI were then averaged. Four ROIs were identified: the temporal cortex (TC), medial prefrontal cortex (mPFC), inferior frontal cortex (IFC), and angular gyrus (AG) (Fig. 1b).

2.5.3. RSA at the individual level

To investigate the cortical representations of the linguistic hierarchy in conversational speech, RSA was conducted (Kriegeskorte, 2008) to assess the relationship between linguistic similarity (i.e., turns or topics) and brain similarity at both the individual and dyadic levels (Fig. 4a). At the individual level, each interlocutor was analyzed either as a speaker when producing speech or as a listener when listening to the speech produced by their partner (Fig. 4c-d). For each interlocutor, a turn-level representational similarity matrix (RSM) was constructed by calculating the semantic similarity between the last sentence of each turn and the first sentence of all other turns. In this matrix, only the values at the lower triangle and the diagonal positions were used for the RSA. It should be noted that the values at the diagonal positions of the sentence matrix represent the semantic similarity between the first and last sentences of a turn, and thus they were included in the RSA, generating an individual-level semantic RSM. Next, an individual-level brain RSM was also built by calculating the similarity of preprocessed HbO signals corresponding to the sentences in the semantic RSM mentioned above.

For analyzing brain activity signals, the traditional approach involves using Pearson correlation to calculate the similarity between brain activities, a method widely used for measuring neural activity patterns during natural stimuli processing (Hasson et al., 2004). However, due to the varying lengths of individuals' turns, we employed the Dynamic Time Warping (DTW) method to calculate the distance between brain signals of varying turn lengths. DTW is a distance measurement method that can align signals of different lengths, which cannot be achieved using Pearson correlation, by finding an optimal match between them and then calculating the Euclidean distance. Specifically, the "dtw" function in Python (from the dtw-python package, version 1.3.1) was used (Giorgino, 2009). DTW measures the distance between two time series, always resulting in a value greater than 0, and allows for the calculation of similarity between two time series of unequal length by aligning sequences non-linearly, accommodating the dynamic nature of conversational speech (Silbert et al., 2014; Yamauchi et al., 2015). Additionally, we have multiplied the value obtained from DTW by −1, whereby converting the measure from distance to similarity, i.e., the larger the values, the greater the similarity.

To further validate the DTW's effectiveness, we conducted a random pair permutation test. In each permutation, we calculated the DTW similarity between adjacent sentences across turns (corresponding to an interval of 1). Then, we randomly re-paired the interlocutors within each of the dyads to form 32 fake dyads and recomputed the DTW similarity between interval of 1. The number of turns for each fake dyad was determined by the interlocutor with the fewer turns. This process was repeated 1000 times to generate null distributions for similarity. A significantly higher DTW similarity in true dyads compared to fake dyads (i.e., the null distribution) would indicate that DTW is a valid metric capturing neural similarities across sentences of varying lengths.

Similarly, the topic-level RSM was also built by calculating the cosine similarity between the turns of an interlocutor and all other turns of the same interlocutor. In this case, only the values in the lower triangle of the matrix were used for RSA. A brain-level RSM was also built by calculating the similarity of preprocessed HbO signals corresponding to the turns in the turn RSM using the DTW method. We also validated DTW at the topic level by performing a random pair permutation test, whereby the DTW similarity between adjacent turns was computed (interval = 1). This allowed us to assess whether the DTW similarity in true dyads was significantly different from that in fake dyads at the topic level, further confirming the method's suitability in capturing neural similarities across turns of varying lengths.

Finally, Spearman correlation was applied between the semantic RSM for turns and topics and the corresponding brain RSMs, obtaining a correlation coefficient for each interlocutor at each linguistic level and each fNIRS channel. To determine the statistical significance of these correlation coefficients, a matrix-shuffled permutation test was employed, which involved randomly shuffling the rows and columns of the brain RSM to generate a random matrix. A null Spearman correlation was then calculated between the original semantic RSM and the shuffled brain RSM. This procedure was repeated 1000 times to generate a null distribution, and the statistical significance for each channel was calculated using the one-tail equation: $p = \frac{1 + \text{number of null r values} \geq \text{empirical r}}{1 + 1000}$ (Song et al., 2021). The FDR method was used to correct for multiple comparisons across all fNIRS channels ($q < 0.05$, Fig. 4c-d).

2.5.4. RSA at the dyadic level

At the dyadic level, the RSM construction slightly differs from the single-brain perspective, as it incorporates neural coupling between two interlocutors (Fig. 4e).

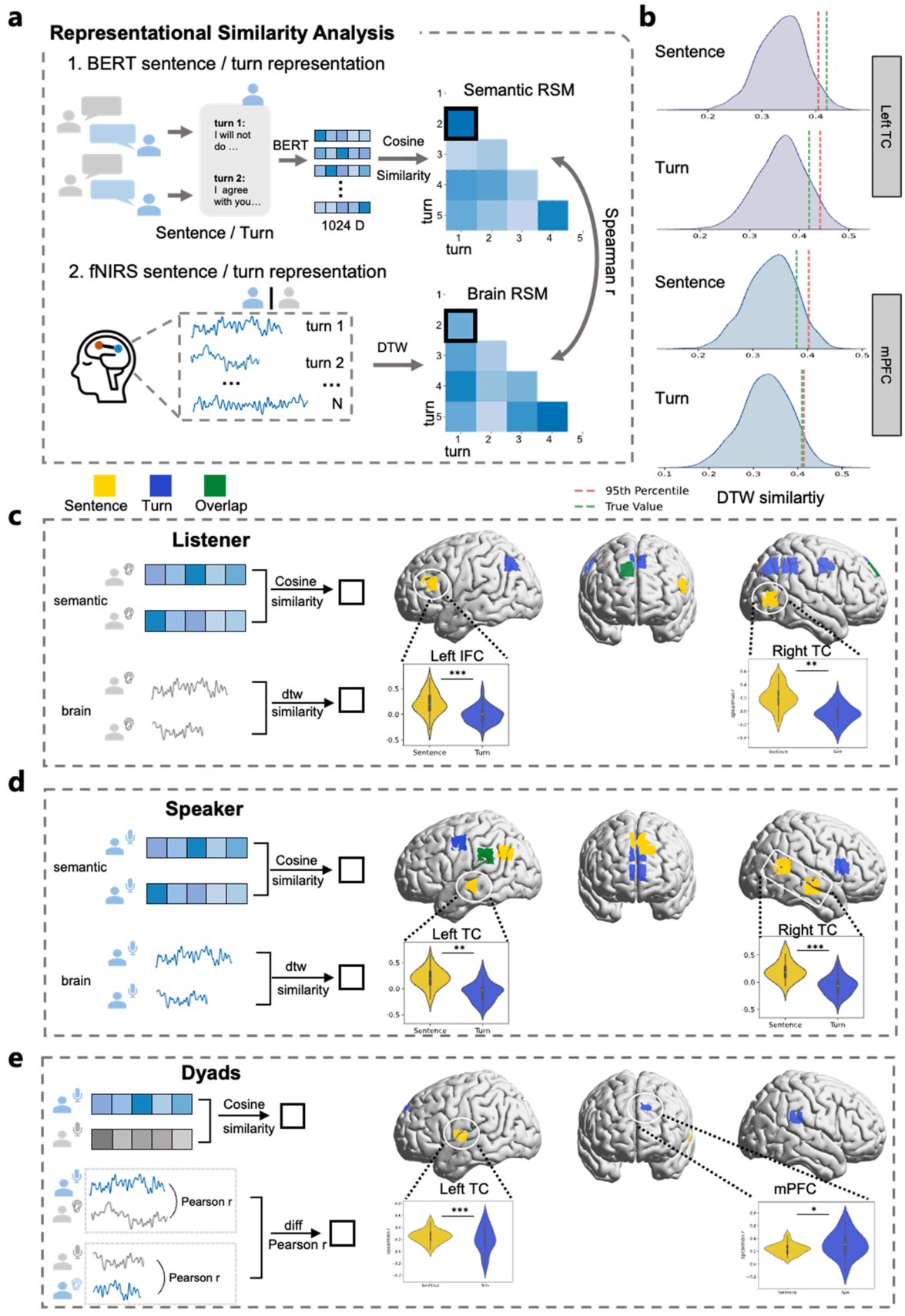
For turns, a semantic RSM was constructed by calculating the semantic similarity between the last sentence of each turn and the first sentence of all other turns. In this matrix, only the values in the lower triangle and diagonal positions were used for RSA. A brain-level RSM was also built by calculating the Pearson correlation of preprocessed HbO signals corresponding to the sentences in the sentence RSM. Differences in coupling between sentences were assessed using this correlation.

For topics, for each dyad, a turn-by-turn semantic similarity matrix was constructed by calculating the cosine similarity between the semantic vectors of each pair's turns. The matrix is organized in an alternating turn sequence (A1-B1-A2-B2, etc.), where 'A' and 'B' denote the two interlocutors and the numbers represent the *n*-th turn. We analyzed only the lower triangle of the matrix, excluding diagonal elements to avoid assessing similarity within the same individual's turn. Similarly, HbO signals are organized according to the same turn sequence. Since the time series had equal lengths in this step, we used Pearson correlation to calculate the similarity of brain activity instead of DTW as, for the time-aligned signals, DTW could introduce unnecessary path adjustments, potentially distorting the global relationship by overemphasizing local variations (Keogh et al., 2004). Specifically, we calculated the INS during each conversational turn, where one participant was speaking and the other listening, and used the difference in coupling between turns to construct the RSM, with rows and columns representing the differences in coupling between turns.

Finally, Spearman correlation was applied between the semantic RSM (for sentences or turns) and the corresponding brain RSM, obtaining a correlation coefficient for each dyad at each linguistic level and each fNIRS channel. The subsequent statistical analysis, including the application of a matrix shuffle permutation test to determine significance and the FDR method for multiple comparisons correction, follows the methodology described previously at the individual level ($q < 0.05$, Fig. 4e).

2.5.5. Reconstructing the linguistic hierarchy from the cortical representation

To further elucidate the cortical representation of turns and topics, we investigated the linguistic hierarchy as previously mentioned using brain activity. This analysis was only conducted on the ROI at the dyadic level because we did not find any cortical architecture at the individual level (Fig. 2h). The same procedure used in linguistic data analysis was employed for the HbO signal by calculating the DTW distance. Furthermore, to facilitate comparison, we converted the DTW distance into a normalized similarity. It should be noted that HbO signals from



(caption on next page)

Fig. 4. The cortical representations of the hierarchical structure of conversational speech. (a) Representational Similarity Analysis (RSA) pipeline. At the individual level, we first extracted the semantic vector of sentence/turn of each participant and constructed a semantic RSM by calculating the cosine similarity between sentences or turns. Then, we extracted brain activity for each sentence or turn, and calculated the similarity between brain activities through Dynamic Time Warping (DTW), and constructed a brain activity RSM. Finally, a spearman correlation was performed between the semantic RSM and the brain activity RSM for each interlocutor. Statistical significance was determined using a permutation test and corrected for multiple comparisons using the FDR method ($q < 0.05$). At the dyadic level, the main RSA pipeline was the same as at the individual level, but both semantic and brain RSMs were constructed based on the turn-taking sequence of speech of two interlocutors (shown as left panel in d). Brain activity similarity was indicated by the difference of INS, which was calculated by Pearson correlation. (b) Validation on the effectiveness of DTW through comparing DTW similarity between true dyads and fake dyads in the left TC and mPFC at both the turn and topic levels. Histograms represent the null distributions of DTW similarity derived from 1000 permutations of randomly paired fake dyads. The red vertical line indicates the DTW similarity of true dyads. The green vertical line marks the 95th percentile of the null distribution, serving as the significance threshold. (c-d) RSA results for the listener and speaker. Yellow indicates significant brain representation at the turn level, blue indicates significant brain representation at the topic level, and green represents brain regions involved in both levels (i.e., overlapping regions). The violin plots illustrate the differences in representational strength between the turn and topic levels. (e) RSA results for dyads. For (c-e), only the results survived the FDR correction are shown. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

speech produced by both interlocutors of a dyad were used. To avoid the problem of autocorrelation in the HbO signals, instead of comparing interlocutor A's brain activity in sentence 1 to his/her own activity in sentence 2, interlocutor A's brain activity in sentence 1 was correlated to interlocutor B's brain activity in sentence 2.

For turns, brain similarity of sentences within turns was calculated in the same manner as at the linguistic dyadic level, but we employed the DTW method to calculate brain similarity due to the varying length of sentences. Similarly, brain similarity across turns was further averaged between the two interlocutors in a dyad, generating an index of dyadic-level brain similarity for across turns. For topics, the calculation of brain similarity was consistent with that at the individual level but was further averaged between the two interlocutors in a dyad, generating indexes of dyadic-level turn similarity for within and across topics. Finally, paired *t*-test were conducted to compare brain similarities within and across turns, as well as within and across topics (Fig. 5b, e, h, k).

Finally, we validated the results of the incremental effects into a hierarchical structure matrix of the brain by comparing the significant differences between each interval condition. We then compared this brain-decoded hierarchical structure matrix to the previously identified linguistic hierarchical structure matrix (Fig. 3h-ii), using the Jaccard similarity method (Maitra, 2010) to evaluate the correlation between the two (Fig. 5d, g, j, m). To determine the statistical significance of the correlation, a permutation test was employed, which involved randomly shuffling the linguistic hierarchical structure matrix. This procedure was repeated 1000 times to generate a null distribution, and the significance of p-value was obtained based on the position of the true value in the null distribution ($p < 0.05$).

2.6. Distinctiveness of linguistic units with different timescales

To validate the results in the Familiar condition and examine the distinctiveness of turns compared to topics, the same procedures were repeated in the Unfamiliar condition. Previous evidence suggests that low topic familiarity might hinder the anticipatory processing of upcoming semantic information (Brothers et al., 2015; Park et al., 2023). Based on this perspective, we predicted that low familiarity of the topic might impact the cortical representation of topics in this study but might not impact that of turns. To test this assumption, first, we applied the same analytic procedures as above (see the Supplementary Text and Fig. S1) and replicated the linguistic hierarchy of the conversational speech in the Unfamiliar condition. Second, based on cortical architecture identified in the Familiar condition, we aimed to reconstruct the hierarchical structure of the conversational speech from brain activity in the Unfamiliar condition. Again, this analysis was only conducted at the dyadic level. Moreover, a cross-reconstruction test was performed at both the turn and topic levels in the same way as above.

3. Results

3.1. The linguistic characteristics of the conversational speech

On average, each topic involved 3 dyads ($SD = 1.011$) in both the Familiar and Unfamiliar conditions (Fig. 3a). The Kruskal-Wallis H test was conducted to examine whether there was a significant difference among topics in the number of dyads involved, but no significant differences were found either in the Familiar ($H(9) = 8.991$, $p = 0.432$) or Unfamiliar condition ($H(9) = 8.991$, $p = 0.431$). Also, no significant difference was found between the Familiar and Unfamiliar conditions in the number of dyads each topic involved ($H(9) = 0.324$, $p = 0.576$). Moreover, as expected, a paired sample *t*-test showed a significant difference between Familiar and Unfamiliar conditions regarding the familiarity scores ($t(31) = 5.050$, $p < 0.001$, Cohen's $d = 0.850$, Fig. 3b).

The conversational speech was transcribed into text, and then various features were extracted to characterize each interlocutor's speech. In the Familiar condition, the average duration of sentences produced by an interlocutor was 7.34 s ($SD = 2.282$), while the average duration of turns was 19.25 s ($SD = 10.342$). Additionally, on average, each interlocutor produced 9.21 turns (ranging from 4 to 22, $SD = 4.721$) and 24.42 sentences ($SD = 7.221$), with each sentence involving 5.90 function words ($SD = 3.621$) and 6.21 content words ($SD = 3.012$). Each turn comprised, on average, 23.42 function words and 25.10 content words ($SD = 15.020$ and 14.210, respectively). LME analysis revealed no significant differences in the number of content and function words across topics or word types. The results did not show any significant effects (main effect of word type: $F(1, 954) = 0.781$, $p = 0.435$, $\eta_p^2 = 0.012$; main effect of topic: $F(9, 954) = 1.079$, $p = 0.281$, $\eta_p^2 = 0.010$; interaction between word type and topic: $F(9, 954) = 0.306$, $p = 0.759$, $\eta_p^2 = 0.002$). The results of the Unfamiliar condition were reported in supplementary materials (SM).

3.2. The linguistic hierarchy of the conversational speech

First, a paired two-sample *t*-test was conducted on the averaged semantic similarity between within- and across-turns at the group level (i.e., in total 64 individuals). As expected, our results showed significantly higher semantic similarity between sentences within turns than across turns ($t(62) = 8.170$, $p < 0.001$, Cohen's $d = 1.101$, Fig. 3c). For topics of each interlocutor, consistent results with that of turn were obtained in that the semantic similarity within topics was significantly higher than that across topics ($t(62) = 5.831$, $p < 0.001$, Cohen's $d = 0.772$, Fig. 3d).

Next, the same analyses were conducted at the dyadic level. Here it should be noted that the speech vectors produced by both interlocutors in a dyad were used. The results showed that the semantic similarity was significantly higher within turns than across turns ($t(31) = 3.552$, $p < 0.01$, Cohen's $d = 0.634$, Fig. 3f). Similarly, for topics, the semantic similarity was significantly higher within topics than across topics ($t(31) = 2.816$, $p < 0.01$, Cohen's $d = 0.502$, Fig. 3 g). These findings confirmed the linguistic hierarchy presented in the speech both at the

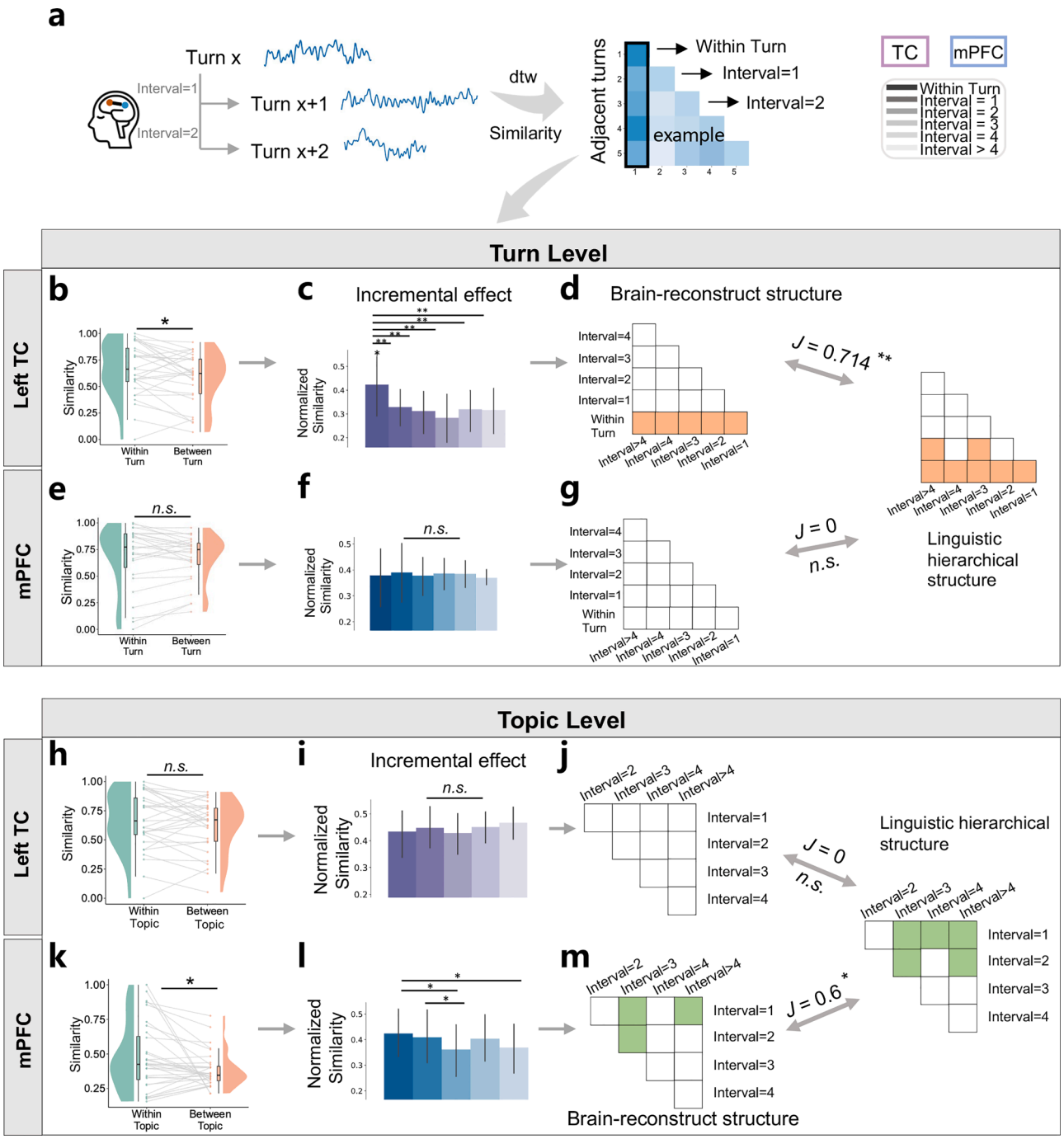


Fig. 5. Reconstruction of the linguistic hierarchy from brain activity in the Familiar condition. (a) The analytic pipeline. (b) and (h) The boundary effect of turns/topics reconstructed from the left TC. (c) and (i) The incremental effect reconstructed from the left TC. (d) and (j) The correlation was calculated between the left TC reconstructed linguistic hierarchy and the original linguistic hierarchy using the Jaccard method. (e-g) and (k-m) show the same results in the mPFC. *n.s.* represent non-significant results. *, $p < 0.05$; **, $p < 0.01$.

individual and dyadic levels.

3.3. The incremental context effect

The results showed that the semantic similarity was significantly higher when there was no interval (i.e., within turns) than when intervals ≥1 (i.e., the semantic similarity between the last sentence of the $turn_i$ and the first sentence of the $turn_{i+1}$, $turn_{i+2}$, $turn_{i+3}$, etc., within the same interlocutor) as well as the chance level (Fig. 3e-i, *ps* < 0.05, FDR corrected). There was a trend of gradual decrease in the semantic similarity though no significant differences were found between pairs of intervals (*ps* > 0.05, Fig. 3e-ii).

At the dyadic level, intervals between turns were coded as 1, 2, 3, 4 and ≥4, respectively. Odd intervals like 1, 3, and 5 indicate turns from different interlocutors, while even intervals like 2, 4, and 6 indicate turns from the same interlocutor. As expected, the semantic similarity when there was no interval (i.e., within turns) was significantly higher than when the interval ≥ 1 (see Materials and Methods, and Fig. 2i) as well as the chance level (Fig. 3h, $p < 0.05$, FDR corrected). Moreover, the semantic similarity when the interval was 1 (i.e., between the different interlocutors) was significantly higher than that when the interval = 3 (i.e., between the different interlocutors, $p = 0.007$),

suggesting a significant trend of gradual decrease in semantic similarity.

For the topic level, semantic vectors were averaged across sentences within each turn, and similarity was calculated between turns of the same interlocutor within the same topic. As expected, semantic similarity of turns declined as the interval length increased (Fig. 3e-i). Semantic similarity was significantly higher when interval = 1 than when interval > 3 ($p \leq 0.006$) as well as the chance level (Fig. 3e-ii, $p = 0.006$, FDR corrected). No other significant differences were found ($ps > 0.05$, FDR corrected, Fig. 3e-ii).

At the dyadic level, a similar pattern was found (Fig. 3h); that is, the semantic similarity when the interval = 1 was significantly higher than when the interval ≥ 3 ($p = 0.009$ for interval 1 vs. interval 3, $p = 0.039$ for interval 1 vs. interval 4) as well as the chance level ($ps < 0.05$), but none was found between interval = 1 (within the same interlocutors) and interval = 2 (between different interlocutors) ($p = 0.65$). These results suggested that the linguistic hierarchy of the bidirectional conversational speech is not only presented in the individual interlocutor but also shared by the dyadic interlocutors.

3.4. The cortical representations of the linguistic hierarchy of the conversational speech

At the individual level, the results from the listener indicated that the representation of smaller unit such as turns was significantly associated with the right temporal cortex (TC: CH40), left inferior frontal cortex (IFC: CH5), and medial prefrontal cortex (mPFC: CH42), while that of larger unit such as topics was associated with the right primary motor cortex (PMC: CH22), angular gyrus (AG: CH28, CH29) and mPFC (CH41, CH42) (Fig. 4c). Results from the speaker showed that the representation of turns was significantly associated with the bilateral TC (CH18, CH36 and CH38), left AG (CH12, CH13) and mPFC (CH41, CH43), while that of topics was associated with the left PMC (CH2), right IFC (CH30), left AG (CH12) and mPFC (CH44, CH47) (Fig. 4d).

We further examined the representational specificity of these brain regions in processing turns or topics. A permutation test on ROIs revealed that, in the listener, significantly higher cortical representation of turns compared to topics was found in the left IFC (mean difference: 0.172, $p < 0.001$) and right TC (mean difference: 0.165, $p = 0.003$, Fig. 4c). In the speaker, significantly higher cortical representation of turns than topics was found in the bilateral TC (left: mean difference: 0.16, $p = 0.002$; right: mean difference: 0.183, $p < 0.001$, Fig 4d). No significantly higher representation of topics than turns was found either in the speaker or the listener ($ps > 0.05$). These findings suggested that, at least at the individual level, there was no selective brain response to linguistic units with different time scales.

Next, to investigate the cortical representation of conversational speech at the dyadic level, RSA was conducted similarly to the individual level, with both semantic and brain RSMs constructed for the turn-taking sequences of two interlocutors, reflecting joint conversational activity (see Materials and Methods, Fig. 4e). The results showed that the representation of turns was significantly associated with the left TC (CH14), while that of topics was significantly associated with the right AG (CH32) and mPFC (CH43) (Fig. 4e).

Finally, an ROI-based contrast between turns and topics revealed at the dyadic level that the cortical representation for turns was significantly higher than that of topics in the left TC (mean difference: 0.071, $p < 0.001$); On the contrary, the cortical representation of topics was significantly higher than that of turns in the mPFC (mean difference: 0.062, $p = 0.048$). No significant difference was found in the right AG (mean difference: 0.020, $p > 0.05$). These findings further supported the first hypothesis, i.e., the linguistic hierarchy of the conversational speech was co-represented by a gradient cortical architecture from TC (turns) to mPFC (topics) at the dyadic level.

Additionally, to validate the effectiveness of DTW in capturing neural similarities between brain signals of varying lengths, we performed random pair permutation tests at both the turn and topic levels. Based on above findings, we selected the left TC and mPFC as the ROIs for the current validation analysis. At the turn level, the results demonstrated that, in the left TC, true dyads exhibited significantly higher DTW similarity compared to fake dyads ($p = 0.019$), but none was observed in the mPFC ($p = 0.146$, Fig. 4b bottom). At the topic level, true dyads showed a marginally significant increase in DTW similarity in the mPFC ($p = 0.059$), but no such pattern was found in the left TC ($p = 0.132$, Fig. 4b top). These findings supported the DTW as an effective metric in individual-level RSA analysis.

3.5. Reconstruction the linguistic hierarchy from the cortical representation

First, at the turn level, the results only showed a significant boundary effect in the left TC ($t(31) = 2.320$, $p = 0.030$, Cohen's $d = 0.425$, Fig. 5b). A significant incremental effect was also only found in the left TC, i.e., a trend of gradual decrease of the semantic similarity though no significant differences were found between pairs of intervals ($ps > 0.05$, Fig. 5c). To further examine the association of the incremental effect reconstructed from the brain with that calculated from the semantic embeddings of the conversational speech, we also calculated the correlation between the brain-reconstructed linguistic hierarchy (Fig. 5d) and the original linguistic hierarchy (Fig. 3e-ii) using the Jaccard similarity method (Maitra, 2010). The results showed a significant similarity compared to the chance level ($J = 0.714$, two-tailed permutation test, $p = 0.002$, Fig. 5d).

Second, at the topic level, a significant boundary effect was only found in the mPFC ($t(31) = 2.314$, $p = 0.020$, Cohen's $d = 0.413$, Fig. 5k). An incremental effect was also observed in the mPFC between interval=1 (the most adjacent turns between interlocutors) and interval ≥ 3 (between interlocutors, Fig. 5l). Notably, no significant difference was found between interval = 1 (between interlocutors) and interval = 2 (within interlocutors) ($p > 0.05$). We also found significant similarity between the linguistic hierarchy (Fig. 3e-ii) identified from the conversational speech and that reconstructed from the mPFC activity ($J = 0.600$, $p = 0.012$, Fig. 5m).

Finally, a cross-reconstruction analysis was conducted to test whether the left TC activity could reconstruct the boundary and incremental effects of topics, while mPFC activity could reconstruct those of turns. However, the results showed that in the mPFC, neither the boundary effect ($t(31) = 0.352$, $p = 0.731$, Cohen's $d = 0.063$, Fig. 5e) nor the incremental effect ($ps > 0.05$, FDR corrected, Fig. 5f) reached significance. There was not a significant correlation between the reconstructed and original linguistic hierarchy, either ($J = 0$, $p > 0.05$, Fig. 5g). Similarly, in the TC, there was no significant boundary effect ($t(31) = 1.352$, $p = 0.190$, Cohen's $d = 0.246$, Fig. 5h), nor was there an incremental effect ($ps > 0.05$, FDR corrected, Fig. 5i). No significant correlation was found, either ($J = 0$, $p > 0.05$, Fig. 5j). These findings further supported the close associations between the gradient pattern of the cortical representations from the TC to mPFC and the linguistic hierarchy from turns to topics in the naturalistic conversational speech.

3.6. Distinctiveness of linguistic units with different timescales

First, the results replicated the linguistic hierarchy of the conversational speech in the Unfamiliar condition. Second, based on cortical architecture identified in the Familiar condition, we aimed to reconstruct the hierarchical structure of the conversational speech from brain activity in the Unfamiliar condition. Again, this analysis was only conducted at the dyadic level. The results showed significantly higher similarity in brain activity of the left TC within turns than across turns ($t(58) = 2.684$, $p = 0.013$, Cohen's $d = 0.352$, Fig. 6a). An incremental effect also reached significance (Fig. 6b). Moreover, a significant correlation was found between the original linguistic hierarchy (Fig. S1f) and the brain-reconstructed linguistic hierarchy (Permutation two-tailed test, $J = 0.830$, $p = 0.008$, Fig. 6c). As expected, no significant

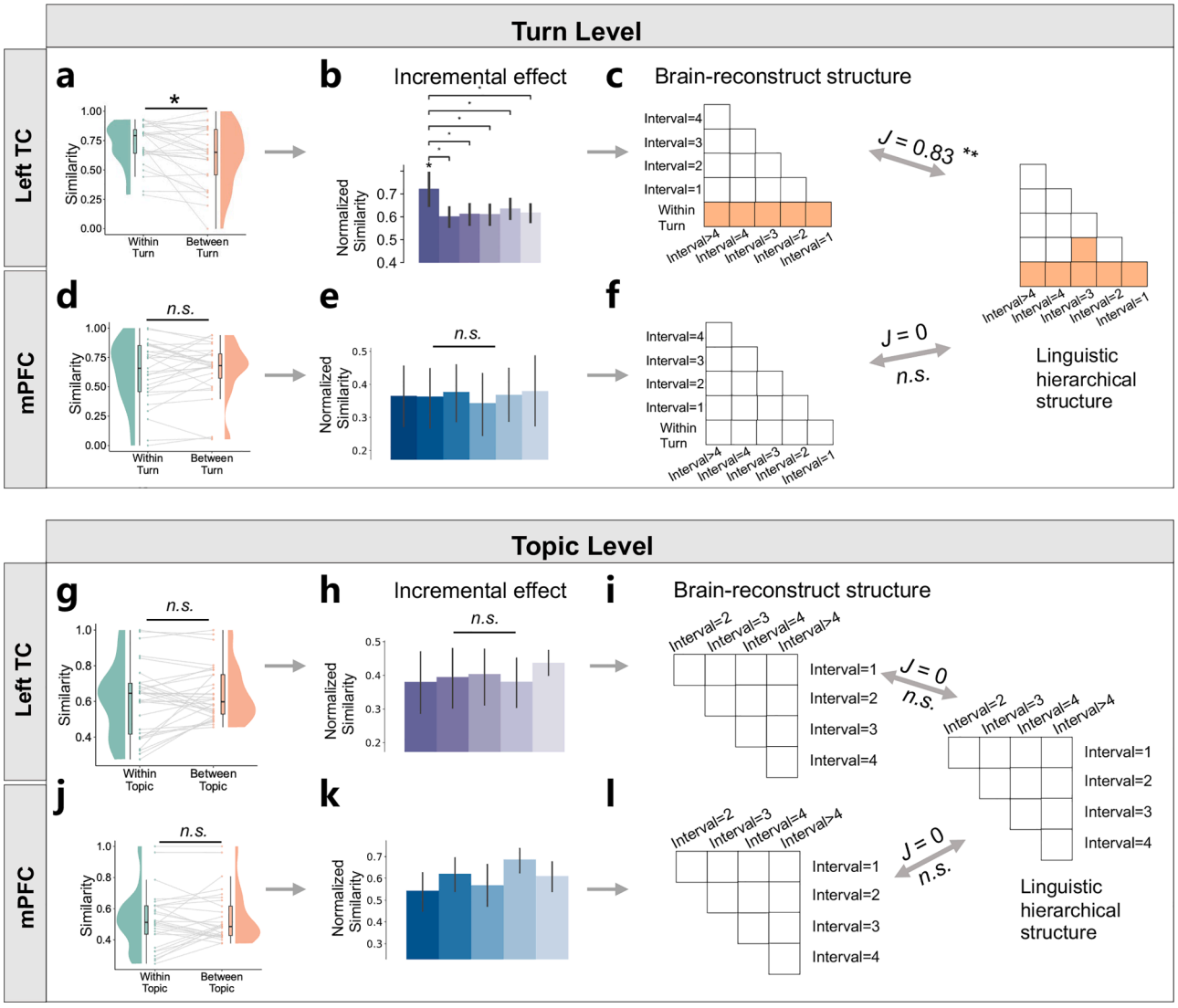


Fig. 6. Reconstruction of the hierarchical structure from brain activity in the Unfamiliar condition. (a) and (g) The boundary effect of the left TC activity at turn or topic level. (b) and (h) The incremental effect in the left TC. (c) and (i) Jaccard similarity in the left TC. (d-f) and (j-l) shows the same results in the left mPFC. *n.s.* represents not significant. *, $p < 0.05$; **, $p < 0.01$.

results were found at the topic level in the mPFC between within and across topics ($t(58) = -0.756$, $p = 0.451$, Cohen's $d = -0.090$, Fig. 6j), and neither an incremental effect ($ps > 0.05$, Fig. 6k) nor a significant correlation was found ($J = 0$, $p > 0.05$, Fig. 6l).

Moreover, a cross-reconstruction test was performed at both the turn and topic levels in the same way as above. Consistent with the Familiar condition, at the turn level, the mPFC was unable to reconstruct the difference between within and across turns ($t(29) = -0.462$, $p = 0.632$, Cohen's $d = -0.081$, Fig. 6d), and neither an incremental effect ($ps > 0.05$, Fig. 6e) nor a significant correlation was found ($J = 0$, $p > 0.05$, Fig. 6f). At the topic level, the left TC could not reconstruct the difference between within and across topics ($t(29) = -0.492$, $p = 0.621$, Cohen's $d = -0.090$, Fig. 6g), and neither an incremental effect ($ps > 0.05$, Fig. 6h) nor a significant correlation was found ($J = 0$, $p > 0.05$, Fig. 6i). Together, these findings suggested that the topic familiarity only modulated the higher levels of the linguistic hierarchy and validated the representational distinctiveness between turns and topics.

4. Discussion

In this study, the hierarchical linguistic structure of the natural language was empirically tested, for the first time, in the bidirectional conversational speech. While it has been demonstrated that our brain can parse the natural language into smaller elements during speech comprehension, most previous studies focus primarily on narrative comprehension, only involving a unidirectional transfer of information between individuals (Salazar et al., 2021). In this situation, speech tends to be well-formed (Chafe and Danielewicz, 1987), allowing for a clear depiction of the linguistic hierarchy. On the contrary, conversational speech is characterized by its dynamic, interactive, and often non-linear nature, with limited evidence of the linguistic hierarchy (Levinson, 2016). In a conversation, interlocutors continuously update their representations of each other's intention through reciprocal exchanges of linguistic information and adjust their behaviors (Levinson and Torreira, 2015; Pickering and Garrod, 2004), which presents unique challenges in discerning the linguistic hierarchy in the conversational speech. In our study, we provided original evidence, through employing the PLM to characterize the features of turns and topics, for the linguistic hierarchical principle in the bidirectional conversational speech, addressing a significant gap in the literature.

Second, we revealed a gradient pattern of the cortical architecture supporting the representation of the linguistic hierarchy in the conversational speech at the dyadic level, being consistent with the second hypothesis as well. Previous studies on the brain representation of

narrative language have majorly used the feature of the temporal receptive window (i.e., the TRW) (Caucheteux et al., 2023) with technologies such as fMRI (Deniz et al., 2021), EEG and magnetoencephalography (MEG) (Ding et al., 2016; Keitel et al., 2018). These studies have elucidated a gradient cortical architecture where smaller linguistic units carrying less information are represented and processed in primary cortical areas (such as auditory and motor areas), whereas larger linguistic units carrying more information are processed in higher-order associative cortical layers (such as the angular gyrus, frontal lobes, and the default mode network). For instance, researchers employed fMRI to record participants' brain responses while listening to stories, revealing that different levels of the cortical architecture are responsible for processing linguistic units across various temporal scales (Lerner et al., 2011). However, no studies have yet explored the linguistic hierarchy of the conversational speech and the complex pattern of brain responses to the bidirectional flow of information in the conversation. The present results, for the first time, revealed such a cortical architecture at the dyadic level, suggesting that a spatially gradient cortical architecture underlies the dynamic updates and reciprocity occurring between interlocutors during a bidirectional conversation.

Additionally, in this cortical architecture, the left TC was found to be more associated with the representation of turns, while the mPFC was more associated with the representation of topics, showing a gradient pattern of cortical representation. The left superior temporal cortex (STC) has been widely recognized in previous studies as a primary structure involved in auditory speech processing compared to non-speech or unintelligible speech (Binder et al., 2000), which clearly distinguishes its functional role from that of the primary auditory cortex (Hamilton et al., 2021). The STC is particularly sensitive to the extraction of semantic (Devauchelle et al., 2009) and syntactic features (Friederici et al., 2000) from speech and the integration of lexical-semantic and syntactic information (Awad et al., 2007). On the other hand, the mPFC plays a crucial role in social cognition, particularly in understanding and inferring the mental states of others (Van Overwalle and Baetens, 2009). Previous studies have found that bilateral dorsolateral and inferior prefrontal areas show increased activation as sentences became less causally related (Ferstl et al., 2007), suggesting a key role of the mPFC in integrating information to form a coherent narrative structure. Therefore, unlike narrative comprehension, the higher level of the cortical architecture in the mPFC might be distinct in dynamically updating and reciprocally interacting between interlocutors during conversation to form a coherent conversational structure.

Third, the above cortical architecture was only identified at the dyadic level rather than the individual level. Previous theory has indicated that individuals engaged in joint activities like a conversation construct a shared cognitive and linguistic framework to facilitate communication and understanding (Tomasello et al., 2005). This framework, grounded in shared intentions and mutual knowledge, is essential for successful interaction. Recent theories also suggest that interactions involve more than mirroring, incorporating complementary behaviors and dynamically coupled interactions that further complexify social exchanges (Hasson and Frith, 2016). During a conversation, participants not only share intentions but also dynamically build and adjust their mental representations to achieve mutual understanding, continuously coordinating and reaching consensus on content (Clark and Brennan, 1991). Although these theories do not explicitly address the interplay between shared and individual representations, they imply that the formation of group-level intentions and representations may weaken individual-level representations. Supporting this notion, a recent hyperscanning study found that during group interactions, increased inter-brain synchronization in the dorsolateral prefrontal cortex (DLPFC) among group members was accompanied by decreased neural activation in the DLPFC at the individual level (Yang et al., 2020). This perspective is reflected in our findings, showing a unique cortical architecture for the linguistic hierarchy only at the dyadic level. This result also aligns with the interactive alignment model, which posits that conversational participants gradually align at multiple levels, including situational models, semantic, syntactic, lexical, prosodic, and phonetic representations, forming similar conceptual and linguistic representations (Pickering and Garrod, 2004). It is also consistent with the hierarchical model for interpersonal verbal communication (Jiang et al., 2021), which suggests that neural synchrony between interlocutors specifically underpins the shared representation of speech, from basic visual-auditory integration to mutual understanding and the representation of social concepts and relationships.

Finally, there are limitations in this study. First, although fNIRS is the most appropriate technique to test the cortical representation of naturalistic conversational speech, it may still lead to a possibility of false negative results due to its limited coverage of brain regions, especially those in deeper structures than the outer cortex. Although this limitation didn't affect our conclusion about the linguistic hierarchy and the gradient cortical architecture, it should be further elaborated in future studies. Second, we did not look at smaller linguistic units such as sentences and words. Future studies are needed to involve more levels of linguistic units to capture the complete linguistic hierarchy and the cortical architecture.

In conclusion, this study provided original evidence for the linguistic hierarchy and the gradient cortical architecture supporting the representation of the linguistic hierarchy in the conversational speech. This finding suggests that the linguistic hierarchy, as outlined in previous studies, is a general principle of human natural language, no matter whether in the unidirectional narrative speech or the bidirectional conversational speech. Moreover, this effect was only found at the dyadic level between interlocutors. Together, these findings filled a gap in the literature by extending the linguistic hierarchy of natural language and its gradient cortical representation from unidirectional narrative comprehension to bidirectional naturalistic conversation.

Code availability

All analyses were performed using Python 3.11.5 with standard functions and toolboxes. All codes used are available upon request.

Data availability

The data that support the findings of this study are available from the corresponding author upon request.

CRediT authorship contribution statement

Ruhuiya Aili: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Siyuan Zhou:** Writing – review & editing, Methodology, Investigation, Data curation, Conceptualization. **Xinran Xu:** Investigation. **Xiangyu He:** Investigation. **Chunming Lu:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare no competing financial interests.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (62293550, 62293551).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2025.121180.

Data availability

Data will be made available on request.

References

Anderson, A.J., Kiela, D., Binder, J.R., Fernandino, L., Humphries, C.J., Conant, L.L., Raizada, R.D.S., Grimm, S., Lalor, E.C., 2021. Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. J. Neurosci. 41, 4100–4119. https://doi.org/10.1523/JNEUROSCI.1152-20.2021.

Awad, M., Warren, J.E., Scott, S.K., Turkheimer, F.E., Wise, R.J.S., 2007. A common system for the comprehension and production of narrative speech. J. Neurosci. 27, 11455–11464. https://doi.org/10.1523/JNEUROSCI.5257-06.2007.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J.W., Hasson, U., Norman, K.A., 2017. Discovering event structure in continuous narrative perception and memory. Neuron 95, 709–721. https://doi.org/10.1016/j.neuron.2017.06.041 e5.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S.F., Springer, J.A., Kaufman, J. N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex 10, 512–528. https://doi.org/10.1093/cercor/10.5.512.

Bornkessel-Schlesewsky, I., Schlesewsky, M., Small, S.L., Rauschecker, J.P., 2015. Neurobiological roots of language in primate audition: common computational properties. Trends Cogn. Sci. 19, 142–150. https://doi.org/10.1016/j.tics.2014.12.008.

Brothers, T., Swaab, T.Y., Traxler, M.J., 2015. Effects of prediction and contextual support on lexical processing: prediction takes precedence. Cognition 136, 135–149. https://doi.org/10.1016/j.cognition.2014.10.017.

Caucheteux, C., Gramfort, A., King, J.-R., 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. Nat. Hum. Behav. https://doi.org/10.1038/s41562-022-01516-2.

Chafe, W., Danielewicz, J., 1987. Properties of Spoken and Written language, in: Comprehending Oral and Written Language. Academic Press, San Diego, CA, US, pp. 83–113.

Chang, C.H.C., Nastase, S.A., Hasson, U., 2022. Information flow across the cortical timescale hierarchy during narrative construction. Proc. Natl. Acad. Sci. 119, e2209307119. https://doi.org/10.1073/pnas.2209307119.

Clark, H.H., Brennan, S.E., 1991. Grounding in communication. Perspectives On Socially Shared Cognition. American Psychological Association, Washington, DC, US, pp. 127–149. https://doi.org/10.1037/10096-006.

Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., 2021. Pre-training with whole word masking for Chinese BERT. IEEE ACM Trans. Audio Speech Lang. Process. 29, 3504–3514. https://doi.org/10.1109/TASLP.2021.3124365.

Dai, B., Chen, C., Long, Y., Zheng, L., Zhao, H., Bai, X., Liu, W., Zhang, Y., Liu, L., Guo, T., Ding, G., Lu, C., 2018. Neural mechanisms for selectively tuning in to the target speaker in a naturalistic noisy situation. Nat. Commun. 9, 2405. https://doi.org/10.1038/s41467-018-04819-z.

Deniz, F., Tseng, C., Wehbe, L., Gallant, J.L., 2021. Semantic representations during language comprehension are affected by context. https://doi.org/10.1101/2021.12.15.472839.

Devauchelle, A.-D., Oppenheim, C., Rizzi, L., Dehaene, S., Pallier, C., 2009. Sentence syntax and content in the human temporal lobe: an fMRI adaptation study in auditory and visual modalities. J. Cogn. Neurosci. 21, 1000–1012. https://doi.org/10.1162/jocn.2009.21070.

Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D., 2016. Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci. 19, 158–164. https://doi.org/10.1038/nn.4186.

Ferstl, E.C., Neumann, J., Bogler, C., von Cramon, D.Y., 2007. The extended language network: a meta-analysis of neuroimaging studies on text comprehension. Hum. Brain Mapp. 29, 581–593. https://doi.org/10.1002/hbm.20422.

Friederici, A.D., Meyer, M., von Cramon, D.Y., 2000. Auditory language comprehension: an event-related fMRI study on the processing of syntactic and lexical information. Brain Lang. 74, 289–300. https://doi.org/10.1006/brln.2000.2313.

Giorgino, T., 2009. Computing and visualizing dynamic time warping alignments in R: the dtw package. J. Stat. Softw. 31, 1–24. https://doi.org/10.18637/jss.v031.i07.

Hamilton, L.S., Oganian, Y., Hall, J., Chang, E.F., 2021. Parallel and distributed encoding of speech across human auditory cortex. Cell 184, 4626–4639. https://doi.org/10.1016/j.cell.2021.07.019 e13.

Hasson, U., Chen, J., Honey, C.J., 2015. Hierarchical process memory: memory as an integral component of information processing. Trends Cogn. Sci. 19, 304–313. https://doi.org/10.1016/j.tics.2015.04.006.

Hasson, U., Frith, C.D., 2016. Mirroring and beyond: coupled dynamics as a generalized framework for modelling social interactions. Philos. Trans. R. Soc. B Biol. Sci. 371, 20150366. https://doi.org/10.1098/rstb.2015.0366.

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

Haxby, J.V., Connolly, A.C., Guntupalli, J.S., 2014. Decoding neural representational spaces using multivariate pattern analysis. Annu. Rev. Neurosci. 37, 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99. https://doi.org/10.1016/j.cognition.2003.10.011.

Hoffman, P., Loginova, E., Russell, A., 2018. Poor coherence in older people's speech is explained by impaired semantic and executive processes. eLife. https://doi.org/10.7554/eLife.38907.

Hoshi, Y., 2007. Functional near-infrared spectroscopy: current status and future prospects. J. Biomed. Opt. 12, 062106. https://doi.org/10.1117/1.2804911.

Jiang, J., Dai, B., Peng, D., Zhu, C., Liu, L., Lu, C., 2012. Neural synchronization during face-to-face communication. J. Neurosci. 32, 16064–16069. https://doi.org/10.1523/JNEUROSCI.2926-12.2012.

Jiang, J., Zheng, L., Lu, C., 2021. A hierarchical model for interpersonal verbal communication. Soc. Cogn. Affect. Neurosci. 16, 246–255. https://doi.org/10.1093/scan/nsaa151.

Jin, P., Zou, J., Zhou, T., Ding, N., 2018. Eye activity tracks task-relevant structures during speech and auditory sequence perception. Nat. Commun. 9, 5374. https://doi.org/10.1038/s41467-018-07773-y.

Keitel, A., Gross, J., Kayser, C., 2018. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. PLOS Biol. 16, e2004473. https://doi.org/10.1371/journal.pbio.2004473.

Keogh, E., Lonardi, S., Ratanamahatana, C.A., 2004. Towards parameter-free data mining. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Seattle WA USA, pp. 206–215. https://doi.org/10.1145/1014052.1014077. Presented at the KDD04: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Kriegeskorte, N., 2008. Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci. https://doi.org/10.3389/neuro.06.004.2008.

Lee, H., Chen, J., 2022. A generalized cortical activity pattern at internally generated mental context boundaries during unguided narrative recall. eLife 11, e73693. https://doi.org/10.7554/eLife.73693.

Lerner, Y., Honey, C.J., Silbert, L.J., Hasson, U., 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci. 31, 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011.

Levinson, S.C., 2016. Turn-taking in Human communication – origins and implications for language processing. Trends Cogn. Sci. 20, 6–14. https://doi.org/10.1016/j.tics.2015.10.010.

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

Liu, W., Branigan, H.P., Zheng, L., Long, Y., Bai, X., Li, K., Zhao, H., Zhou, S., Pickering, M.J., Lu, C., 2019. Shared neural representations of syntax during online dyadic communication. NeuroImage 198, 63–72. https://doi.org/10.1016/j.neuroimage.2019.05.035.

Luo, C., Ding, N., 2020. Cortical encoding of acoustic and linguistic rhythms in spoken narratives. eLife 9, e60433. https://doi.org/10.7554/eLife.60433.

Maitra, R., 2010. A re-defined and generalized percent-overlap-of-activation measure for studies of fMRI reproducibility and its use in identifying outlier activation maps. NeuroImage 50, 124–135. https://doi.org/10.1016/j.neuroimage.2009.11.070.

Montague, P.R., Berns, G.S., Cohen, J.D., McClure, S.M., Pagnoni, G., Dhamala, M., Wiest, M.C., Karpov, I., King, R.D., Apple, N., Fisher, R.E., 2002. Hyperscanning: simultaneous fMRI during linked social interactions. NeuroImage 16, 1159–1164. https://doi.org/10.1006/nimg.2002.1150.

Park, J.J., Baek, S.-C., Suh, M.-W., Choi, J., Kim, S.J., Lim, Y., 2023. The effect of topic familiarity and volatility of auditory scene on selective auditory attention. Hear. Res. 433, 108770. https://doi.org/10.1016/j.heares.2023.108770.

Pickering, M.J., Garrod, S., 2004. Toward a mechanistic psychology of dialogue. Behav. Brain Sci. 27, 169–190. https://doi.org/10.1017/S0140525X04000056.

Rivers, W.M., 2018. Teaching Foreign Language Skills: Second Edition. University of Chicago Press.

Salazar, M., Shaw, D.J., Gajdoš, M., Mareček, R., Czekóová, K., Mikl, M., Brázdil, M., 2021. You took the words right out of my mouth: dual-fMRI reveals intra- and inter-personal neural processes supporting verbal interaction. NeuroImage 228, 117697. https://doi.org/10.1016/j.neuroimage.2020.117697.

Schapiro, A.C., Rogers, T.T., Cordova, N.I., Turk-Browne, N.B., Botvinick, M.M., 2013. Neural representations of events arise from temporal community structure. Nat. Neurosci. 16, 486–492. https://doi.org/10.1038/nn.3331.

Silbert, L.J., Honey, C.J., Simony, E., Poeppel, D., Hasson, U., 2014. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. Proc. Natl. Acad. Sci. U.S.A. 111, E4687–E4696. https://doi.org/10.1073/pnas.1323812111.

Song, H., Finn, E.S., Rosenberg, M.D., 2021. Neural signatures of attentional engagement during narratives and its consequences for event memory. Proc. Natl. Acad. Sci. 118, e2021905118. https://doi.org/10.1073/pnas.2021905118.

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

Stolk, A., Verhagen, L., Toni, I., 2016. Conceptual alignment: how brains achieve mutual understanding. Trends Cogn. Sci. 20, 180–191. https://doi.org/10.1016/j.tics.2015.11.007.

Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H., 2005. Understanding and sharing intentions: the origins of cultural cognition. Behav. Brain Sci. 28, 675–691. https://doi.org/10.1017/S0140525X05000129.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15, 273–289. https://doi.org/10.1006/nimg.2001.0978.

Van Overwalle, F., Baetens, K., 2009. Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. NeuroImage 48, 564–584. https://doi.org/10.1016/j.neuroimage.2009.06.009.

Xu, J., Kemeny, S., Park, G., Frattali, C., Braun, A., 2005. Language in context: emergent features of word, sentence, and narrative comprehension. NeuroImage 25, 1002–1015. https://doi.org/10.1016/j.neuroimage.2004.12.013.

Yamauchi, T., Xiao, K., Bowman, C., Mueen, A., 2015. Dynamic time warping: a single dry electrode EEG study in a self-paced learning task. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). Presented at the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 56–62. https://doi.org/10.1109/ACII.2015.7344551.

Yang, J., Zhang, H., Ni, J., De Dreu, C.K.W., Ma, Y., 2020. Within-group synchronization in the prefrontal cortex associates with intergroup conflict. Nat. Neurosci. 23, 754–760. https://doi.org/10.1038/s41593-020-0630-x.

Yeshurun, Y., Nguyen, M., Hasson, U., 2017. Amplification of local changes along the timescale processing hierarchy. Proc. Natl. Acad. Sci 114, 9475–9480. https://doi.org/10.1073/pnas.1701652114.

Zhou, S., Xu, X., He, X., Zhou, F., Zhai, Y., Chen, J., Long, Y., Zheng, L., Lu, C., 2023. Biasing the neurocognitive processing of videos with the presence of a real cultural other. Cereb. Cortex 33, 1090–1103. https://doi.org/10.1093/cercor/bhac122.